STOR 565 Homework

A. Calculus and Elementary Inequalities

1. By graphing the functions $f(x) = 1+x$ and $g(x) = e^x$, argue informally that $1+x \leq e^x$ for every number x, and find one value of x where equality holds. Deduce from this inequality that $\log y \leq y - 1$ for every $y > 0$.

2. Let $x = x_1, \ldots, x_n$ be a univariate sample of n numbers. It is a standard, and important, fact that the quantity $h(a) = \sum(x_i - a)^2$ is minimized when (and only when) a is the sample mean $m(x) = n^{-1} \sum_{i=1}^n x_i$. Here we show this in two different ways.

   a. Take a derivative of h to find the number a that minimizes or maximizes the function h, and then take another derivative to show that the number you found minimizes the function.

   b. Consider the expression for h. Add and subtract $m(x)$ inside the parentheses, expand the square, and take the sum of these terms. Note that one of the sums is zero, and one of the terms does not depend on a. Use this to show that the sample mean minimizes $h(a)$.

   c. Use what you've shown above to find the following

$$\operatorname*{argmin}_{a \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^n (x_i - a)^2 \quad \text{and} \quad \min_{a \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^n (x_i - a)^2$$

3. (Inequalities from Calculus) Use calculus to establish the following inequalities.

   a. $(1 + u/3)^3 \geq 1 + u$ for every $u \geq 0$

   b. $x + x^{-1} \geq 2$ for $x \geq 1$

4. Inequalities for $\log(1 + x)$ and $\log(1 - x)$ from Taylor's theorem.

   a. Expand the function $h(v) = \log v$ in a third order Taylor series around the point $v = 1$. (Thus you will be expressing $h(1 + x)$ in terms of x, $h(1)$, $h'(1)$, $h''(1)$, and $h'''(u)$ for some u between 1 and $1 + x$. Note that x may be negative.)

   b. By examining the final term in the series, show that $\log(1 + x) \geq x - x^2/2$ for $x \geq 0$.

c. By examining the final term in the series, show that $\log(1-x) \leq -x - x^2/2$ for $0 \leq x < 1$.

5. Let $h(u) = (1+u)\log(1+u) - u$. (This function appears in Bennett's exponential inequality for sums of independent, bounded random variables.)

a. By considering the first few terms of the Taylor expansion of $h(\cdot)$ around zero, show that for every $u \geq 0$

$$h(u) \geq \frac{u^2}{2+2u}$$

b. (Optional) Use calculus to establish the stronger bound that for every $u \geq 0$

$$h(u) \geq \frac{u^2}{2+2u/3}$$

6. Show that $xy \leq 3x^2 + y^2/3$ for $x, y \geq 0$.

7. Show that $|e^a - e^b| \leq e^b\, e^{|a-b|}\, |a-b|$.

8. Let $a_1, \ldots, a_n$ be real numbers. Show that $n^{-1}\sum_{k=1}^n |a_k| \leq (n^{-1}\sum_{k=1}^n a_k^2)^{1/2}$.

9. Let $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ be numbers in the interval $[-1, 1]$. Establish the inequality

$$|a_1 \cdots a_n - b_1 \cdots b_n| \;\leq\; \sum_{i=1}^n |a_i - b_i|$$

Hint: Use induction and the fact that $a_1a_2 - b_1b_2 = (a_1 - b_1)a_2 + b_1(a_2 - b_2)$.

10. Show that for each number $u \in \mathbb{R}$ we have

$$\min(u, 1-u) \;=\; u\,\mathbb{I}(u < 1/2) + (1-u)\,\mathbb{I}(u \geq 1/2)$$

Hint: Consider separately the cases $u < 1/2$ and $u \geq 1/2$.

11. Find the gradient and Hessian of the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$f(x) = x_1^2 x_2 + 3x_1 - 5x_2 + 1$$

12. Let $f : \mathbb{R}^d \to \mathbb{R}$ be defined by $f(\mathbf{x}) = \mathbf{x}^t\mathbf{A}\mathbf{x}$ where $\mathbf{A} \in \mathbb{R}^{d\times d}$ is symmetric.

a. Show that the gradient of f is given by $\nabla f(\mathbf{x}) = 2\mathbf{A}\mathbf{x}$.

b. Show that the Hessian of f is given by $\nabla^2 f(\mathbf{x}) = 2\mathbf{A}$.

13. Use calculus to show that for $u \in (0, 1)$

$$\frac{u^2}{2(1-u)} \geq -\log(1-u) - u$$

B. Linear Algebra and Matrices

1. Let $\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}^t \mathbf{v} = \sum_{i=1}^d u_i \, v_i$ be the usual inner product in $\mathbb{R}^d$. Recall that the norm of a vector $\mathbf{u} \in \mathbb{R}^d$ is defined by $||\mathbf{u}|| = \langle \mathbf{u}, \mathbf{u} \rangle^{1/2}$. Use this definition, and the definition of vector sums and scalar multiplication to establish the following.

a. Show that $\langle \mathbf{u}, \mathbf{v} \rangle = \langle \mathbf{v}, \mathbf{u} \rangle$

b. Show that $\langle a\mathbf{u}, b\mathbf{v} \rangle = ab \, \langle \mathbf{u}, \mathbf{v} \rangle$

c. Show that $\langle \mathbf{u} + \mathbf{w}, \mathbf{v} \rangle = \langle \mathbf{u}, \mathbf{v} \rangle + \langle \mathbf{w}, \mathbf{v} \rangle$

d. Show that $||\mathbf{u}|| = 0$ if and only if $\mathbf{u} = 0$.

e. Use the definition of the norm to show that $||\mathbf{u} + \mathbf{v}||^2 = ||\mathbf{u}||^2 + 2\langle \mathbf{u}, \mathbf{v} \rangle + ||\mathbf{v}||^2$.

f. Use this equation and the Cauchy Schwarz inequality to establish the triangle inequality for the vector norm, namely $||\mathbf{u} + \mathbf{v}|| \leq ||\mathbf{u}|| + ||\mathbf{v}||$.

g. The standard Euclidean distance between two vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$ is defined by $d(\mathbf{u}, \mathbf{v}) = ||\mathbf{u} - \mathbf{v}||$. Establish that $d(\mathbf{u}, \mathbf{v}) \leq d(\mathbf{u}, \mathbf{w}) + d(\mathbf{w}, \mathbf{v})$ for any vectors $\mathbf{u}, \mathbf{v}, z \in \mathbb{R}^d$. Draw a picture illustrating this result.

2. Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ be the data matrix associated with n samples $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ such that $\sum_{i=1}^n \mathbf{x}_i = 0$. Answer the following. You may use arguments from class, but clearly explain your work.

a. Define the sample covariance matrix $\mathbf{S}$ in terms of $\mathbf{X}$. What are the dimensions of $\mathbf{S}$?

b. Show that $\mathbf{S} = n^{-1} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^t$

c. Show that $\mathbf{S}$ is symmetric and non-negative definite

d. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ be the eigenvalues of $\mathbf{S}$. Show that $\sum_{k=1}^p \lambda_k = n^{-1} ||\mathbf{X}||^2$

e. Show that if $p > n$ then $\text{rank}(\mathbf{S}) < p$ and $\mathbf{S}$ is not invertible. Hint: recall that $\text{rank}(\mathbf{S}) = \text{rank}(\mathbf{X}^t \, \mathbf{X}) = \text{rank}(\mathbf{X}) \leq \min(n, p)$.

f. For any vector $\mathbf{v} \in \mathbb{R}^p$ we have $n^{-1} \sum_{i=1}^n \langle \mathbf{x}_i, \mathbf{v} \rangle^2 = \mathbf{v}^t \mathbf{S} \mathbf{v}$.

3. Let $\mathbf{u} = (u_1, \ldots, u_d)^t$ be a vector in $\mathbb{R}^d$.

a. Show that $||\mathbf{u}|| \leq |u_1| + \cdots + |u_d|$. Hint: use the fact that for $a, b \geq 0$ one has $a \leq b$ if and only if $a^2 \leq b^2$. Give an examples with $d = 2$ where the bound holds with equality, and where one has strict inequality.

b. Use the Cauchy-Schwarz inequality to get the upper bound $|u_1| + \cdots + |u_d| \leq ||\mathbf{u}|| \, d^{1/2}$. Find an example where the bound holds with equality.

4. Let $a_1, \ldots, a_n$ be positive numbers. Use the Cauchy-Schwartz inequality for inner products to show that $n^2 \leq (\sum_{k=1}^n a_k)(\sum_{k=1}^n a_k^{-1})$. Hint: Begin with the identity $1 = a_k^{1/2} \, a_k^{-1/2}$ which holds for $k = 1, \ldots, n$.

5. (Norms of outer products) Let $\mathbf{u} \in \mathbb{R}^k$ and $\mathbf{v} \in \mathbb{R}^l$ be vectors. Find an expression relating the Frobenius norm of the outer product $||\mathbf{u}\mathbf{v}^t||$ to the Euclidean norms of the vectors $||\mathbf{u}||$ and $||\mathbf{v}||$.

6. Show that if $\mathbf{v}_1, \mathbf{v}_2$ are eigenvectors of a symmetric matrix $\mathbf{A}$ with different eigenvalues, then $\mathbf{v}_1, \mathbf{v}_2$ are orthogonal. Hint: Begin by taking transposes to show that $\mathbf{v}_1^t \mathbf{A} \mathbf{v}_2$ and $\mathbf{v}_2^t \mathbf{A} \mathbf{v}_1$ are equal; then use the definition of an eigenvector and simplify.

7. Recall that the Frobenius norm of a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ is given by $||A|| = \sqrt{\sum_{i=1}^m \sum_{j=1}^n a_{ij}^2}$, the square root of the sum of the squares of the entries of the matrix. Establish the following properties of the Frobenius norm for matrices.

(a) $||\mathbf{A}|| = 0$ if and only if $\mathbf{A} = 0$

(b) $||b\mathbf{A}|| = |b| \, ||\mathbf{A}||$

(c) $||\mathbf{A}||^2 = \sum_{i=1}^m ||a_{i\cdot}||^2 = \sum_{j=1}^n ||a_{\cdot j}||^2$. Here $a_{i\cdot}$ denotes the ith row of A, and $a_{\cdot j}$ denotes the jth column of A.

(d) $||\mathbf{A}\mathbf{B}|| \leq ||\mathbf{A}|| \, ||\mathbf{B}||$. Hint: Use Cauchy-Schwarz.

8. Recall that the trace of an $n \times n$ matrix $\mathbf{A} = \{a_{ij}\}$ is the sum of its diagonal elements, that is $\mathrm{tr}(\mathbf{A}) = \sum_{i=1}^n a_{ii}$.

a. Show that $\mathrm{tr}(\mathbf{A}) = \mathrm{tr}(\mathbf{A}^t)$.

b. Let $\mathbf{A}$ be an $n \times p$ matrix, and let $\mathbf{B}$ be a $p \times n$ matrix. Note that if $n \neq p$ then $\mathbf{A}\mathbf{B}$ and $\mathbf{B}\mathbf{A}$ are square matrices with different dimensions. Nevertheless, use the fact that $(\mathbf{A}\,\mathbf{B})_{ii} = \sum_{j=1}^p a_{ij} \, b_{ji}$ to establish the important identity $\mathrm{tr}(\mathbf{A}\,\mathbf{B}) = \mathrm{tr}(\mathbf{B}\,\mathbf{A})$.

c. By applying the identity above multiple times, show that if $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ are square matrices of the same dimension then

$$\mathrm{tr}(\mathbf{A}\,\mathbf{B}\,\mathbf{C}) \;=\; \mathrm{tr}(\mathbf{B}\,\mathbf{C}\,\mathbf{A}) \;=\; \mathrm{tr}(\mathbf{C}\,\mathbf{A}\,\mathbf{B})$$

Show that if $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ are symmetric then, in addition, we have $\mathrm{tr}(\mathbf{A}\,\mathbf{B}\,\mathbf{C}) = \mathrm{tr}(\mathbf{A}\,\mathbf{C}\,\mathbf{B})$. Note that this equality is *not* true in general.

d. Suppose that $\mathbf{B} = \{b_{ij}\}$ is an $m \times n$ matrix. By considering $(\mathbf{B}^t\,\mathbf{B})_{ii}$, show that

$$\mathrm{tr}(\mathbf{B}^t\,\mathbf{B}) \;=\; \sum_{i=1}^{m}\sum_{j=1}^{n} b_{ij}^2$$

which is the square of the Frobenius norm $||\mathbf{B}||^2$ of $\mathbf{B}$.

9. Suppose that $\mathbf{v}_1, \ldots, \mathbf{v}_k$ are orthogonal vectors in $\mathbb{R}^n$. Show that $||\sum_{i=1}^k \mathbf{v}_i||^2 = \sum_{i=1}^k ||\mathbf{v}_i||^2$. Interpret this in terms of the Pythagorean formula relating the length of the hypotenuse of a right triangle to the lengths of the other edges.

10. Let $\mathbf{A}$ and $\mathbf{B}$ be invertible $n \times n$ matrices. Argue that $(\mathbf{A}\,\mathbf{B})^{-1} = \mathbf{B}^{-1}\,\mathbf{A}^{-1}$.

11. Let $\mathbf{A}$ be an $n \times n$ matrix. Show that if $\mathbf{A}$ has rank n then $\mathbf{A}\mathbf{x} = 0$ if and only if $\mathbf{x} = 0$. Hint: If $\mathbf{A}$ has rank n then its columns are linearly independent.

12. Let $A \in \mathbb{R}^{d\times d}$ be symmetric. The spectral theorem tells us that there is an orthonormal basis $v_1, \ldots, v_d$ for $\mathbb{R}^d$ such that each v_i is an eigenvector of A.

a. Show that the $d \times d$ matrix $\Gamma = [v_1, \ldots, v_d]$ is orthogonal, that is $\Gamma^t\Gamma = I_d$. Note that this implies $\Gamma\Gamma^t = I_d$, though you do not need to show this.

b. Let $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ be the $d \times d$ diagonal matrix with D_{ii} equal to the ith eigenvalue of A and all other entries equal to zero. Show that $A\Gamma = \Gamma D$.

c. Conclude from the expression above that A can be written in the form $A = \Gamma D \Gamma^t$

13. Recall that any symmetric matrix $A \in \mathbb{R}^{d\times d}$ can be written in the form $A = \Gamma D \Gamma^t$, where $\Gamma \in \mathbb{R}^{d\times d}$ is an orthogonal matrix and $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ is a diagonal matrix with D_{ii} equal to the ith eigenvalue of A and all other entries equal to zero. Suppose that A is non-negative definite, so that each $\lambda_i \geq 0$. Define $A^{1/2} = \Gamma D^{1/2} \Gamma^t$ where $D^{1/2} = \mathrm{diag}(\lambda_1^{1/2}, \ldots, \lambda_d^{1/2})$. Show that $A^{1/2}$ is symmetric and satisfies $A^{1/2}A^{1/2} = A$.

14. Let $A, B \in \mathbb{R}^{m \times n}$ be a matrices.

a. Show that $A = B$ iff $Ax = Bx$ for all $x \in \mathbb{R}^n$.

b. Let $v_1, \ldots, v_n$ be a basis for $\mathbb{R}^n$. Show that if $Av_i = Bv_i$ for $1 \leq i \leq n$ then $Ax = Bx$ for all $x \in \mathbb{R}^n$.

15. (Non-negative definite matrices) Recall that a symmetric matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$ is non-negative definite (written $\mathbf{A} \geq 0$) if $\mathbf{u}^t \mathbf{A} \mathbf{u} \geq 0$ for every vector $\mathbf{u} \in \mathbb{R}^d$, and is positive definite (written $\mathbf{A} > 0$) if $\mathbf{u}^t \mathbf{A} \mathbf{u} > 0$ for every non-zero vector $\mathbf{u} \in \mathbb{R}^d$.

a. Show that if a matrix $\mathbf{A} \geq 0$ then its diagonal entries are non-negative. Hint: Let $\mathbf{u}$ be a standard basis vector having one component equal to 1 and all other components equal to 0.

b. Show that if $\mathbf{A} \geq 0$ then all its eigenvalues are non-negative.

c. It is tempting to think that if $\mathbf{A} \geq 0$ then all its entries are non-negative, but this is not the case. Consider the matrix

$$\mathbf{A} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

Show that $\mathbf{A}$ is non-negative definite, but not positive definite. What is the rank of $\mathbf{A}$?

d. Modify the $(1,1)$ entry of $\mathbf{A}$ to produce a positive definite matrix $\mathbf{B}$. What is the rank of $\mathbf{B}$?

16. Let $\mathbf{U}\,\mathbf{D}\,\mathbf{V}^t$ be the singular value decomposition of an $m \times n$ matrix $\mathbf{A}$ with rank r, where $\mathbf{U}$, $\mathbf{D}$, and $\mathbf{V}$ are as given in class. Describe the matrices $\mathbf{U}$, $\mathbf{D}$, and $\mathbf{V}$ in detail (you may repeat what is in the notes), and establish the identity $A = \sum_{i=1}^{r} \sigma_i(\mathbf{A}) \mathbf{u}_i \mathbf{v}_i^t$. Here $\mathbf{u}_i$ and $\mathbf{v}_i$ are the ith columns of $\mathbf{U}$ and $\mathbf{V}$, respectively.

17. *More on the SVD.* Let $\mathbf{A}$ be an $m \times n$ matrix with real valued entries.

a. Show that a number λ is a non-zero eigenvalue of $\mathbf{A}\,\mathbf{A}^t$ if and only if it is a non-zero eigenvalue of $\mathbf{A}^t\,\mathbf{A}$.

b. Let $\mathbf{A} = \sum_{j=1}^{r} \sigma_j(\mathbf{A})\, u_j\, v_j^t$ be the SVD expansion of $\mathbf{A}$. Show that for each $1 \le d \le r$ we have

$$\left\| \sum_{j=1}^{d} \sigma_j(\mathbf{A})\, u_j\, v_j^t \right\|_F = \sum_{j=1}^{d} \sigma_j(\mathbf{A})^2$$

Hint: It may be helpful to use the identity $\|\mathbf{B}\|_F = \mathrm{tr}(\mathbf{B}^t\mathbf{B})$

18. Define hyperplanes $H_i = \{x : x^t u_i = c_i\}$ for $1 \le i \le m$ where $u_1, \ldots, u_m \in \mathbb{R}^n$ are linearly independent, and $c_1, \ldots, c_m \in \mathbb{R}$. What can you say about the intersection $\cap_{i=1}^m H_i$? Hint: Consider the linear equation $\mathbf{U}x = c$ where $\mathbf{U}$ has rows $u_1^t, \ldots, u_m^t$ and $c = (c_1, \ldots, c_m)^t$.

19. Let $u, v \in \mathbb{R}^d$. Show that the set $H = \{x : ||x - u|| = ||x - v||\}$ of points equidistant from u and v is a hyperplane. In particular, find a direction vector w and offset b such that $H = \{x : x^t w = b\}$. (Hint: square each norm in the definition of H and simplify.) Show that H is orthogonal to the line connecting u and v, and note that u and v are equally far from H. Thus, H is the perpendicular bisector of line connecting u and v.

20. Let $\mathbf{u}_1 = (-1, 2, 0)^t$ and $\mathbf{u}_2 = (2, 4, 3)^t$. Find the projections of $\mathbf{u}_1$ and $\mathbf{u}_2$ onto $\mathbf{v}$ where:

a. $\mathbf{v} = (0, 1, 0)^t$

b. $\mathbf{v} = (1, 1, 1)^t$

c. $\mathbf{v} = (1, 0, -1)^t$

21. Let $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^d$ be orthonormal vectors with span $V = \{\alpha\mathbf{v}_1 + \beta\mathbf{v}_2 : \alpha, \beta \in \mathbb{R}\}$. For $\mathbf{u} \in \mathbb{R}^d$ define the projection of $\mathbf{u}$ onto V to be the vector $\mathbf{v} \in V$ that is closest to $\mathbf{u}$,

$$\mathrm{proj}_V(\mathbf{u}) = \operatorname*{argmin}_{\mathbf{v} \in V} ||\mathbf{u} - \mathbf{v}||.$$

Show that $\mathrm{proj}_V(\mathbf{u}) = \langle \mathbf{u}, \mathbf{v}_1 \rangle\, \mathbf{v}_1 + \langle \mathbf{u}, \mathbf{v}_2 \rangle\, \mathbf{v}_2$. Hint: Adapt the argument used in class for the projection onto a one-dimensional subspace.

22. *Measuring the variability of a set of vectors.* Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ be a sample of n p-dimensional vectors. We can measure the extent to which a vector $\mathbf{u} \in \mathbb{R}^p$ acts as representative for the sample through the sum of squares

$$S(\mathbf{u}) := \sum_{i=1}^{n} ||\mathbf{x}_i - \mathbf{u}||^2.$$

a. Show that $S(\mathbf{u})$ is minimized when $\mathbf{u}$ is equal to the centroid

$$\overline{\mathbf{x}} \;=\; \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i.$$

If the general case seems difficult, consider first the case when $p = 1$.

Consider the two variance-type quantities

$$V_1 \;=\; \frac{1}{n}\sum_{i=1}^{n}||\mathbf{x}_i - \overline{\mathbf{x}}||^2 \quad \text{and} \quad V_2 \;=\; \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}||\mathbf{x}_i - \mathbf{x}_j||^2.$$

Note that V_1 and V_2 are non-negative.

b. Carefully describe V_1 and V_2 in plain English.

c. Give necessary and sufficient conditions under which $V_1 = 0$.

d. Give necessary and sufficient conditions under which $V_2 = 0$.

e. Show that

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{x}_i^t\,\mathbf{x}_j \;=\; (\sum_{i=1}^{n}\mathbf{x}_i)^t\,(\sum_{j=1}^{n}\mathbf{x}_j) \;=\; n^2\,||\overline{\mathbf{x}}||^2$$

f. Using the identity from part e., and some additional calculations, show that

$$V_1 \;=\; V_2 \;=\; \frac{1}{n}\sum_{i=1}^{n}||\mathbf{x}_i||^2 - ||\overline{\mathbf{x}}||^2$$

23. Let $\mathbf{x}_1 = (-1, 2, 0)$ and $\mathbf{x}_2 = (2, 4, 3)$. Find the projections of $\mathbf{x}_1$ and $\mathbf{x}_2$ onto $\mathbf{u}_0$ where:

a. $\mathbf{u}_0$ = (0,1,0)

b. $\mathbf{u}_0$ = (1,0,-1)

24. Consider the 3×3 matrix $\mathbf{A}$ below:

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 3 \\ 0 & 1 & 2 \\ 1 & 0 & 1 \end{pmatrix}$$

a. What is rank($\mathbf{A}$)?

b. What is det($\mathbf{A}$)?

c. Calculate the eigenvalues and corresponding eigenvectors of $\mathbf{A}$. (Note: in this example the number of non-zero eigenvalues is *less* than the rank of the matrix.)

C. Probability

1. Let X, Y be random variables and let $a, b > 0$. Define events

$$A = \{|X| \geq a\} \quad B = \{|Y| \geq b\} \quad C = \{|X + Y| \geq a + b\}$$

a. Argue that if $a > |X|$ and $b > |Y|$ then $a + b > |X + Y|$.

b. Conclude that $A^c \cap B^c \subseteq C^c$.

c. Show using Boolean algebra that $C \subseteq A \cup B$.

d. Conclude using the properties of probability that $\mathbb{P}(C) \leq \mathbb{P}(A) + \mathbb{P}(B)$.

e. Reason similarly to show that $\mathbb{P}(|XY| \geq a) \leq \mathbb{P}(|X| \geq a/b) + \mathbb{P}(|Y| \geq b)$.

2. Let P be a probability measure on a set $\mathcal{X}$. Recall that if A and B are subsets of $\mathcal{X}$ and $P(B) > 0$, then the conditional probability of A given B is defined by

$$P(A \,|\, B) \;=\; \frac{P(A \cap B)}{P(B)}$$

Show the following.

a. If A and B are disjoint then $P(A \cup B \,|\, C) = P(A \,|\, C) + P(B \,|\, C)$

b. $P(A^c \,|\, B) = 1 - P(A \,|\, B)$

c. If $A \subseteq B$ then $P(A \,|\, C) \leq P(B \,|\, C)$

3. Let $\mathcal{X}$ be a set and let A, B be subsets of $\mathcal{X}$. Recall that the indicator function of A is defined by

$$\mathbb{I}_A(x) \;=\; \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \in A^c \end{cases}$$

a. Show that $\mathbb{I}_{A^c} = 1 - \mathbb{I}_A$.

b. Show that $\mathbb{I}_A - \mathbb{I}_B = \mathbb{I}_{B^c} - \mathbb{I}_{A^c}$.

c. Show that $\mathbb{I}_{A \cap B} = \mathbb{I}_A \, \mathbb{I}_B$.

d. Let $u, v \in \{0, 1\}$. Show that $\mathbb{I}(u \neq v) = |\mathbb{I}(u = 1) - \mathbb{I}(v = 1)|$. Hint: Consider separately the cases $\mathbb{I}(u \neq v) = 0$ and $\mathbb{I}(u \neq v) = 1$.

4. Let A, B, C be events in a random experiment with probability measure P. Carefully show the following.

a. $\max(P(A), P(B)) \leq P(A \cup B)$.

b. $P(A) \leq P(A \cup B) + P(B^c)$.

c. $P(A \,|\, B \cap C) \geq P(A \cap B \,|\, C)$.

d. $P(A \,|\, B \cap C) \geq P(A \,|\, C)\, P(B \,|\, A \cap C)$.

5. Let (X, Y) be a discrete random pair with joint probability mass function $p(x, y)$. Recall from the lecture notes that we may define $\mathbb{E}(Y|X) = \varphi(X)$ where $\varphi(x) = \sum_y y\, p(y|x)$. Establish the following.

a. If $Y \geq 0$ then $\mathbb{E}(Y|X) \geq 0$

b. $\mathbb{E}(aY + b|X) = a\, \mathbb{E}(Y|X) + b$

c. $\mathbb{E}\{\, \mathbb{E}(Y|X)\, \} = \mathbb{E}Y$

6. Let $X_1, \ldots X_n$ be independent and identically distributed random variables. Calculate $\mathbb{E}[X_1|X_1 + \ldots + X_n = x]$. (Hint: Consider $\mathbb{E}[S_n|X_1 + \ldots + X_n = x]$ where $S_n = X_1 + \ldots + X_n$ and use symmetry.)

7. Let X, Y be non-negative random variables with joint density function $f(x, y) = y^{-1}\, e^{-x/y}\, e^{-y}$ for $x, y \geq 0$.

a. Find the marginal density $f(y)$ of Y

b. Find the conditional density $f(x \,|\, y)$ of X given $Y = y$

c. Find $\mathbb{E}[X \,|\, Y = y]$

d. Find $\mathbb{E}[X \,|\, Y]$

8. Let X be a discrete random variable taking values in a finite (or countably infinite) set $\mathcal{X}$, and having probability mass function $p(x) = \mathbb{P}(X = x)$. Let $h : \mathcal{X} \to [a, b]$ be any function.

a. Write down the sum for $\mathbb{E}h(X)$.

b. Show that $\mathbb{E}h(X) = a$ if $p(x) > 0$ only when $h(x) = a$.

c. Establish the reverse implication: if $\mathbb{E}h(X) = a$ then $p(x) > 0$ only when $h(x) = a$. Hint: Assume to the contrary that $p(x') > 0$ for some $x' \in \mathcal{X}$ with $h(x') \neq a$. As h takes values in $[a, b]$, we have $h(x') > a$. Use this to show $\mathbb{E}h(X) > a$.

d. Following the arguments above, show that $\mathbb{E}h(X) = b$ if and only if $p(x) > 0$ implies $h(x) = b$.

9. Let $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ be a jointly distributed pair. Assume that $\mathcal{X}$ and $\mathcal{Y}$ are finite. Recall that X and Y are independent if $\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\,\mathbb{P}(Y = y)$.

a. Show that if X and Y are independent then $\mathbb{P}(X = x \,|\, Y = y)$ does not depend on y.

b. Let $y \in \mathcal{Y}$ be fixed. Show that if $\mathbb{P}(Y = y \,|\, X = x)$ does not depend on x then it is equal to $\mathbb{P}(Y = y)$.

c. Suppose that for each $y \in \mathcal{Y}$ the conditional probability $\mathbb{P}(Y = y \,|\, X = x)$ does not depend on x. Show that X and Y are independent.

D. Order

1. Let $\{a_1, \ldots, a_n\}$ and $\{b_1, \ldots, b_n\}$ be two sequences of numbers. Rigorously establish the following inequalities.

a. $\min\{a_i\} + \min\{b_i\} \;\leq\; \min\{a_i + b_i\} \;\leq\; \min\{a_i\} + \max\{b_i\}$

b. $-\min\{a_i\} = \max\{-a_i\}$ and $-\max\{a_i\} = \min\{-a_i\}$

c. $\max\{a_i\} - \max\{b_i\} \;\leq\; \max\{|a_i - b_i|\}$

Use part (b) to find a chain of inequalities like that in part (a) for maxima

2. In each case below find $\min_{x\in\mathcal{X}} f(x)$, $\operatorname{argmin}_{x\in\mathcal{X}} f(x)$, $\max_{x\in\mathcal{X}} f(x)$, and $\operatorname{argmax}_{x\in\mathcal{X}} f(x)$. Indicate when the min or the max do not exist. It may help to sketch the functions.

a. $f(x) = \sin x$ with $\mathcal{X} = [0, 2\pi]$ and $\mathcal{X} = [0, \pi]$

b. $f(x) = x^2$ with $\mathcal{X} = [-2, 2]$, $\mathcal{X} = (-2, 2]$, $\mathcal{X} = (-2, 2)$

c. $f(x) = \min(x, 1)$ with $\mathcal{X} = [0, 2]$ and $\mathcal{X} = (-2, 2]$

3. Let $U_1, \ldots, U_m$ be random variables. Find an inequality relating $\mathbb{E}(\min_{1\leq j\leq m} U_j)$ and $\min_{1\leq j\leq m} \mathbb{E}U_j$. Hint: Begin by noting that $\min_{1\leq j\leq m} U_j \leq U_k$ for each k.

4. (Saddle points and minimax) Let $\mathcal{X}$ and $\mathcal{Y}$ be sets and let $f : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}$ be any function.

a. Show that, with no further assumptions,

$$\sup_{y\in\mathcal{Y}} \inf_{x\in\mathcal{X}} f(x, y) \;\leq\; \inf_{x\in\mathcal{X}} \sup_{y\in\mathcal{Y}} f(x, y) \tag{1}$$

This simple fact plays an important role in optimization, where it implies the weak duality property of the Lagrange dual problem, and in game theory, where it has connections with Nash equilibria. A pair $(\tilde{x}, \tilde{y}) \in \mathcal{X} \times \mathcal{Y}$ is called a *saddle point* for f if

$$f(\tilde{x}, y) \;\leq\; f(\tilde{x}, \tilde{y}) \;\leq\; f(x, \tilde{y}) \quad \text{for every } x \in \mathcal{X} \text{ and } y \in \mathcal{Y}$$

b. Show that if $(\tilde{x}, \tilde{y})$ is a saddle point for f then

$$f(\tilde{x}, \tilde{y}) \;=\; \inf_{x\in\mathcal{X}} f(x, \tilde{y}) \quad \text{and} \quad f(\tilde{x}, \tilde{y}) \;=\; \sup_{y\in\mathcal{Y}} f(\tilde{x}, y)$$

To see how these inequalities explain the use of the terminology "saddle point", assume that f is nice and smooth, and sketch what it will look like in a neighborhood around the point $(\tilde{x}, \tilde{y})$.

c. Show that the existence of a saddle point implies equality in inequality (1) above.

d. Evaluate both sides of (1) when $\mathcal{X} = [0, 1]$, $\mathcal{Y} = [-1, 1]$, and $f(x, y) = x^2 y$.

5. Let $a_1, \ldots, a_n$ be real numbers, and let $b_1, \ldots, b_n$ be positive. Show that

$$\min_{1 \le i \le n} \frac{a_i}{b_i} \;\le\; \frac{a_1 + \cdots + a_n}{b_1 + \cdots + b_n} \;\le\; \max_{1 \le i \le n} \frac{a_i}{b_i}$$

E. Convexity

1. Show that the following subsets of $\mathbb{R}^d$ are convex.

   a. The emptyset

   b. The hyperplane $H = \{x : x^t u = b\}$

   c. The halfspace $H_+ = \{x : x^t u > b\}$

   d. The ball $B(x_0, r) = \{x : ||x - x_0|| \leq r\}$

2. Let $C_1, \ldots, C_n \subseteq \mathbb{R}^d$ be convex. Show that the intersection $\cap_{i=1}^n C_i$ is convex.

3. Recall that the convex hull of a set $A \subseteq \mathbb{R}^d$, denoted $\text{conv}(A)$, is the intersection of all convex sets C containing A. Show that $\text{conv}(A)$ is equal to the set of all convex combinations $\sum_{i=1}^k \alpha_i x_i$, where $k \geq 1$ is finite, $x_1, \ldots, x_k \in A$, and the coefficients α_i are non-negative and sum to one.

4. (Set sums and scaler products) Given sets $A, B \subseteq \mathbb{R}^d$ and a constant $\alpha \in \mathbb{R}$ define the set sum and set scaler product as follows:

$$A + B \;=\; \{x + y : x \in A \text{ and } y \in B\} \qquad \alpha A \;=\; \{\alpha x : x \in A\}$$

   a. (Optional) Show that if A is open then $A + B$ is open regardless of whether B is open.

   b. Show that if A and B are convex, then so is $A + B$.

   c. If A is convex is $A + B$ necessarily convex?

   d. Show by example that, in general, $2A \neq A + A$.

   d. Show that if A is convex then $\alpha A + \beta A = (\alpha + \beta)A$ for all $\alpha, \beta \geq 0$.

5. Identify the extreme points (if any) of the following convex sets.

   a. The hyperplane $H = \{x : x^t u = b\}$

   b. The halfspace $H_+ = \{x : x^t u > b\}$

   c. The closed ball $\overline{B}(x_0, r) = \{x : ||x - x_0|| \leq r\}$

6. Let $f : C \to \mathbb{R}$ be a strictly convex function defined on a convex set $C \subseteq \mathbb{R}^n$. Show that $\text{argmax}_{x \in C} f(x)$ is contained in the set of extreme points of C.

7. (Operations on convex functions that produce new convex functions) Let $C \subseteq \mathbb{R}^d$ be a convex set and let $f_1, \ldots, f_n : C \to \mathbb{R}$ be convex functions. Use the definition of convexity to establish the following.

a. If $a_1, \ldots, a_n$ are non-negative then $g(x) = \sum_{i=1}^n a_i f_i(x)$ is convex on C.

b. The function $g(x) = \max_{1 \le i \le n} f_i(x)$ is convex on C.

c. If $h : \mathbb{R} \to \mathbb{R}$ is convex and increasing then $g(x) = h(f(x))$ is convex on C. (Recall that h is increasing if $u \le v$ implies $h(u) \le h(v)$).

8. Define what it means for a function to be strictly convex. Define the notion of a global minima. Show that the global minima of a strictly convex function is necessarily unique.

9. Let $h_\alpha : \mathbb{R} \to [0, \infty)$ be defined by $h_\alpha(x) = |x|^\alpha$ where $\alpha > 0$ is fixed. Sketch $h_\alpha(x)$ for $\alpha = 1/2, 1, 2$. For which values of α is $h_\alpha(x)$ convex? Justify your answer.

10. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function. For $\gamma \in \mathbb{R}$ the γ-level set of f is defined to be the set of points x where $f(x)$ is less than or equal to γ. Formally,

$$L_\gamma(f) \;=\; \{x : f(x) \le \gamma\}$$

a. Draw some level sets for the convex functions $f(x) = x^2$ and $f(x) = e^{-x}$. Note that $L_\gamma(f)$ may be empty.

b. Show that for each γ the level set $L_\gamma(f)$ is convex. Hint: If $L_\gamma(f)$ is empty then it is trivially convex. Otherwise, use the definition of a convex set.

11. Let $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ be positive constants.

a. Use Jensen's inequality to establish the Arithmetic-Geometric mean inequality

$$\frac{1}{n} \sum_{i=1}^n a_i \;\ge\; \left(\prod_{i=1}^n a_i \right)^{1/n}.$$

b. Establish the inequality

$$(\Pi_{k=1}^n a_k)^{1/n} \;+\; (\Pi_{k=1}^n b_k)^{1/n} \;\le\; (\Pi_{k=1}^n (a_k + b_k))^{1/n}$$

Hint: First divide the LHS by the RHS.

12. Use the second derivative condition to establish whether the following functions are convex or concave. In each case, sketch the function.

a. The function $f(x) = e^x$ on $(-\infty, \infty)$.

b. The function $f(x) = \sqrt{x}$ on $(0, \infty)$.

c. The function $f(x) = 1/x$ on $(0, \infty)$.

d. The function $f(x) = \log x$ on $(0, \infty)$.

Now let $X > 0$ be a positive random variable. Write out the conclusion of Jensen's inequality for each of the functions above.

13. Define the function $f(x) = x \log x$ for $x \in (0, \infty)$

a. Sketch the function $f(x)$ and show that it is convex.

b. Find the minimum and argmin of $f(x)$.

b. Let $X > 0$ be a random variable. What can you say about the relationship between $\mathbb{E}(X \log X)$ and $\mathbb{E}X \log \mathbb{E}X$?

14. Let $f_1, \ldots, f_k : \mathbb{R}^p \to \mathbb{R}$ be convex functions.

a. Show that for each number t the set $L_t = \{x : \sum_{j=1}^k f_j(x) \leq t\}$ is convex. Hint: Use results from the previous homework.

b. Show that for each t the sets $\{\beta \in \mathbb{R}^p : \sum_{j=1}^p \beta_j^2 \leq t\}$ and $\{\beta \in \mathbb{R}^p : \sum_{j=1}^p |\beta_j| \leq t\}$ are convex.

15. Show that the Lagrange dual function, defined by

$$\tilde{L}(\lambda) \;=\; \min_{w,b} L(w, b, \lambda)$$

is concave. Hint: Argue that the dual function is the minimum of linear (hence concave) functions, and is therefore concave. The SVM dual problem is given by the program

$$\max \tilde{L}(\lambda) \quad \text{s.t.} \quad \sum_{i=1}^n \lambda_i \, y_i = 0 \;\; \text{and} \;\; \lambda_1, \ldots, \lambda_n \geq 0$$

Carefully define the constraint set for λ in this problem and argue that this set is convex. (Note that there are $n+1$ constraints.) Thus the dual problem seeks to maximize a concave function over a convex set.

16. Show that the following functions $f, g, h : [0, 1] \to \mathbb{R}$ used to define impurity measures for growing trees are concave.

a. $m(p) = \min(p, 1 - p)$

b. $g(p) = p(1 - p)$

c. $h(p) = -p \log p - (1 - p) \log(1 - p)$, with the convention that $0 \log 0 = 0$

Which of these functions is strictly concave?

F. Statistics

1. Let X, X' be independent random variables with the same distribution. Show that $\mathrm{Var}(X) = \frac{1}{2}\mathbb{E}(X - X')^2$

2. In this problem we find an upper bound on the variance of a random variable with values in a finite interval. Let X be a random variable taking values in the finite interval $[0, c]$. You may assume that X is discrete, though this is not necessary for this problem.

a. Show that $\mathbb{E}X \leq c$ and $\mathbb{E}X^2 \leq c\,\mathbb{E}X$.

b. Recall that $\mathrm{Var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2$. Use the inequalities above to show that

$$\mathrm{Var}(X) \;\leq\; c^2[u(1-u)] \quad \text{where} \quad u \;=\; \frac{\mathbb{E}X}{c} \in [0,1].$$

c. Use this inequality and simple calculus to show that $\mathrm{Var}(X) \leq c^2/4$ if $X \in [0, c]$.

d. Use this result to show that if X is a random variable taking values in an interval $[a, b]$ with $-\infty < a < b < \infty$ then $\mathrm{Var}(X) \leq (b-a)^2/4$

e. It turns out that the general bound cannot be improved. To see this, show that the variance of the random variable $X \in [a, b]$ with $\mathbb{P}(X = a) = \mathbb{P}(X = b) = 1/2$ is equal to the bound you found above.

3. The empirical cumulative distribution function (CDF) of a sample $x = x_1, \ldots, x_m$ is defined by

$$F_x(t) \;=\; m^{-1} \sum_{i=1}^{m} \mathbb{I}(x_i \leq t)$$

The sum in the definition counts the number of data points that are less than or equal to t, so $F_x(t)$ is the fraction of data points that are less then or equal to t. Suppose that x has four points: -3, -1, -1, and 5.

a. Find the following values of the empirical CDF by using the formula above: $F_x(-4)$, $F_x(0)$, $F_x(-1)$, $F_x(6)$

b. Sketch the empirical CDF for this data set as a function of t.

c. For what values of t is $F_x(t) = 0$?

d. For what values of t is $F_x(t) = 1$?

4. Let $r(x, y)$ be the sample correlation of a bivariate data set $(x, y) = (x_1, y_1), \ldots, (x_n, y_n)$.

   a. Let $ax + b$ denote the data set $ax_1 + b, \ldots, ax_n + b$ and define $cy + d$ similarly. Show that $r(ax + b, cy + d) = r(x, y)$ if $a, c > 0$.

   b. Use the Cauchy-Schwarz inequality to show that $r(x, y)$ is always between -1 and $+1$.

5. Show that if $f(x)$ is bounded and $X \sim \text{Poiss}(\lambda)$ then $\mathbb{E}[\lambda f(X + 1)] = \mathbb{E}[X f(X)]$. Here Poiss($\lambda$) denotes the usual Poisson distribution with pmf $p(k) = e^{-\lambda}\lambda^k/k!$ for $k \geq 0$.

6. Let X be a standard normal random variable and let $Y = X^2$.

   a. Use the cdf method to find the density of Y.

   b. Show that one of the events $\{Y \leq 1\}$ and $\{X \leq 1\}$ is contained in the other.

   c. Show that X, Y are *not* independent.

   d. What is $\text{Cov}(X, Y)$?

G. Normal and Multinormal Distributions

1. Let X have a $\mathcal{N}(\mu, \sigma^2)$ distribution. Show that $\mathbb{E}X = \mu$.

2. Chi-squared distribution. A random variable X has a chi-squared distribution with $k \geq 1$ degrees of freedom, written $X \sim \chi^2_k$, if X has the same distribution as $Z_1^2 + \cdots + Z_k^2$ where $Z_1, \ldots, Z_k$ are iid $\sim \mathcal{N}(0, 1)$.

   a. Find $\mathbb{E}X$ and $\mathrm{Var}(X)$ when $X \sim \chi^2_k$. You may use the fact that $\mathbb{E}Z^4 = 3$ if $Z \sim \mathcal{N}(0, 1)$.

   b. If $X \sim \chi^2_k$ and $Y \sim \chi^2_l$ are independent, what is the distribution of $X + Y$?

3. Let $\Gamma(x)$ be the standard Gamma function, defined for $x > 0$. Show that if $Z \sim \mathcal{N}(0, 1)$ then for each $p \geq 1$

$$\mathbb{E}|Z|^p \;=\; \frac{2^{p/2}}{\sqrt{\pi}}\Gamma((1+p)/2)$$

Deduce from this fact and Stirling's approximation that $||Z||_p := (\mathbb{E}|Z|^p)^{1/p} = O(p^{1/2})$.

4. Show that if $Y \sim \mathcal{N}(0, \sigma^2)$ and $c > 0$ then $\mathbb{E}\{|Y| I(|Y| > c)\} \leq \sigma \exp\{-c^2/2\sigma^2\}$

5. Let U_1, U_2 be uncorrelated random variables with mean zero and variance one. Define $U = (U_1, U_2)^t$. Let $X = (X_1, X_2)^t$ be a random vector with components

$$X_1 \;=\; a\,U_1 + b\,U_2 \;\text{ and }\; X_2 \;=\; c\,U_1 + d\,U_2$$

   a. Find $\mathbb{E}[U]$.

   b. What is $\mathrm{Var}(U)$?

   c. Find $\mathbb{E}X$.

   d. Find the matrix $\mathrm{Var}(X)$ by directly calculating each entry using the definitions of X_1 and X_2.

   e. Find $\mathbf{A}$ such that $X = \mathbf{A}U$.

   f. Find $\mathrm{Var}(X)$ using the formula for $\mathrm{Var}(\mathbf{A}U)$.

   g. In terms of a, b, c and d, when is $\mathbf{A}$ invertible?

6. Let $X \in \mathbb{R}^k$ be a random vector and $A \in \mathbb{R}^{r \times k}$. Use the definition of expected value, variance, and linear algebra to establish the following.

a. $\mathbb{E}(AX) = A\,\mathbb{E}X$

b. $\mathrm{Var}(X)$ is symmetric and non-negative definite

c. $\mathrm{Var}(X)_{ij} = \mathrm{Cov}(X_i, X_j)$

d. $\mathrm{Var}(AX) = A\,\mathrm{Var}(X)A^t$

7. Let $X \sim \mathcal{N}_k(\mu, \Sigma)$ and let $Y = AX + b$ where $A \in \mathbb{R}^{l\times k}$ and $b \in \mathbb{R}^l$.

a. Find $\mathbb{E}Y$ and $\mathrm{Var}(Y)$.

b. Argue carefully that Y is multinormal and find its distribution.

c. Fix $v \in \mathbb{R}^l$. Find the distribution of $U = \langle v, Y\rangle$.

8. Let $X \sim \mathcal{N}_d(\mu, \Sigma)$ and let $Y = \Sigma^{1/2}Z + \mu$ where $Z \sim \mathcal{N}_d(0, I)$.

(a) Show that $\mathbb{E}Y = \mathbb{E}X$ and that $\mathrm{Var}(Y) = \mathrm{Var}(X)$.

(b) Fix $v \in \mathbb{R}^d$. Find the distributions of the random variable v^tX.

9. Show that if $X \sim \mathcal{N}_d(\mu, \Sigma)$ and $U = X^TAX$ then $\mathbb{E}U = \mathrm{tr}(A\Sigma) + \mu^TA\mu$. (It may be helpful to use the fact that $\mathrm{tr}(UV) = \mathrm{tr}(VU)$.)

10. (Bivariate normal distribution). Let $X = (X_1, X_2)^t \sim \mathcal{N}_2$ with

$$\mathbb{E}X_1 = \mu_1,\ \mathbb{E}X_2 = \mu_2,\ \mathrm{Var}(X_1) = \sigma_1^2,\ \mathrm{Var}(X_2) = \sigma_2^2,\ \mathrm{Corr}(X_1, X_2) = \rho \in [-1, 1]$$

a. Find $\mu = \mathbb{E}X$ and $\Sigma = \mathrm{Var}(X)$ in terms of the quantities above.

b. Find the determinant of Σ and conclude that Σ is invertible if and only if $\rho \in (-1, 1)$.

c. Find Σ^{-1} when $\rho \in (-1, 1)$.

d. Write down the density $f(x)$ of X in the case $\rho \in (-1, 1)$.

11. Let U and V be independent $\mathcal{N}(0, 1)$ random variables. Define $Y = V$ and let

$$X = \begin{cases} U & \text{if } UV \geq 0 \\ -U & \text{if } UV < 0 \end{cases}$$

a. Let $A \subseteq [0, \infty)$ be a Borel set. Show that $\mathbb{P}(X \in A) = \mathbb{P}(U \in A)$. Hint: Begin with the decomposition $\mathbb{P}(X \in A) = \mathbb{P}(X \in A, UV \geq 0) + \mathbb{P}(X \in A, UV < 0)$.

b. Carry out a similar analysis for sets $A \subseteq (-\infty, 0)$. Use this and the previous step to show that X has a $\mathcal{N}(0,1)$ distribution.

c. Show that $XY = |UV| \geq 0$ and that $\text{Corr}(X,Y) = 2/\pi < 1$. Conclude from these facts that X and Y are not jointly normal.

d. Show that X^2 and Y^2 are independent.

12. Let $\mathbf{A}$ be a $k \times p$ random matrix with independent entries $A_{ij} \sim \mathcal{N}(0,1)$, and let $x \in \mathbb{R}^p$ be a fixed vector.

a. Show that the random variables $U_i := (\mathbf{A}x)_i$ are independent with mean zero and variance $||x||^2$. Conclude that $\mathbb{E}||\mathbf{A}x||^2 = k\,||x||^2$.

b. Let $Z = (Z_1, \ldots, Z_k)^t$ be a random vector with $Z_i = U_i/||x||$. Show that $Z \sim \mathcal{N}_k(0, I)$, where I is the $k \times k$ identity matrix.

H. Concentration Type Inequalities

1. State and prove Markov's probability inequality and Chebyshev's probability inequality.

2. Let $X \geq 0$ be a random variable with $\mathbb{E}X = 10$ and $\mathbb{E}X^2 = 120$.

   a. Find an upper bound on $\mathbb{P}(X \geq 14)$ using Markov's inequality.

   b. Let $0 < c < 10$. Find an upper bound on $\mathbb{P}(X \geq c)$ using Markov's inequality. Note that the bound is greater than one, and therefore uninformative. Argue informally that this is not a shortcoming of Markov's inequality, that is, $\mathbb{P}(X \geq c)$ may be equal to one.

   c. Find an upper bound on $\mathbb{P}(X \geq 14)$ involving $\mathbb{E}X^2$.

   d. Find an upper bound on $\mathbb{P}(X \geq 14)$ using Chebyshev's inequality. How does this bound compare to those above?

3. Let X be a random variable with $\mathrm{Var}(X) = 3$. Use Chebyshev's inequality to find upper bounds on $\mathbb{P}(|X - \mathbb{E}X| > 1)$ and $\mathbb{P}(|X - \mathbb{E}X| > 2)$. Comment on the potential usefulness of these bounds.

4. Recall that the moment generating function of a random variable X is defined by $M_X(s) = \mathbb{E}e^{sX}$ for all s such that the expectation is finite. Find the moment generating function (MGF) of the following distributions.

   a. Poisson(λ)

   b. $\mathcal{N}(0, 1)$

5. (The weak law of large numbers). Let $U_1, U_2, \ldots, U$ be iid random variables with finite variance, and let $X = n^{-1} \sum_{i=1}^{n} U_i$ be the average of $U_1, \ldots, U_n$.

   a. Find $\mathbb{E}X$ in terms of $\mathbb{E}U$.

   b. Find $\mathrm{Var}(X)$ in terms of $\mathrm{Var}(U)$.

   c. Use these calculations and Chebyshev's inequality to establish that

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} U_i \; - \; \mathbb{E}U \right| \geq t \right) \; \leq \; \frac{\mathrm{Var}(U)}{nt^2}$$

d. What can you conclude from the inequality above when t is fixed and n tends to infinity?

6. Let X and Y be independent random variables with moment generating functions $M_X(s)$ and $M_Y(s)$, respectively. Show that $S = X + Y$ has moment generating function $M_S(s) = M_X(s)\, M_Y(s)$.

7. (Hoeffding's MGF Bound) Let X be a discrete random variable with probability mass function $p(\cdot)$. Assume that $a \leq X \leq b$ and that $\mathbb{E}X = 0$. Let $M_X(s) = \mathbb{E}e^{sX}$ be the moment generating function of X and define $\varphi(s) := \log M_X(s)$.

a. Show that
$$\varphi'(s) = \frac{\mathbb{E}[Xe^{sX}]}{\mathbb{E}e^{sX}} \quad \text{and} \quad \varphi''(s) = \frac{\mathbb{E}[X^2e^{sX}]}{\mathbb{E}e^{sX}} - (\varphi'(s))^2$$

b. Verify that $\varphi(0) = \varphi'(0) = 0$

Now fix $t > 0$ and let U be a new random variable having the "exponentially tilted" probability mass function
$$q(x) \;=\; \frac{p(x)e^{tx}}{\mathbb{E}e^{tX}}$$

c. Verify that $q(\cdot)$ is a probability mass function, that is, $q(x) \geq 0$ and $\sum_x q(x) = 1$.

d. Argue that $a \leq U \leq b$. This follows from the fact that U has the same possible values as X, only with different probabilities.

e. Show that $\mathbb{E}(U) = \varphi'(t)$ and that $\mathrm{Var}(U) = \varphi''(t)$.

f. Using the variance bound for bounded random variables, conclude from (c) and (d) that $\varphi''(t) \leq (b-a)^2/4$.

g. Use the second order Taylor series expansion of φ around $s = 0$ and what you've shown above to establish that $\varphi(s) \leq s^2(b-a)^2/8$ for $s > 0$.

h. Exponentiating the inequality in (g) gives Hoeffding's MGF bound.

8. Let $X_1, \ldots, X_n$ be iid Uniform$(-\theta, \theta)$ random variables.

a. Use Chebyshev's inequality to find a bound on $\mathbb{P}(\sum_{i=1}^n X_i \geq t)$ for $t \geq 0$.

b. Use Hoeffding's inequality to find a bound on $\mathbb{P}(\sum_{i=1}^n X_i \geq t)$ for $t \geq 0$.

9. Let X be a non-negative random variable such that $\mathbb{E}X^2$ is finite. Show that for each $0 < \lambda < 1$ we have the inequality

$$\mathbb{P}(X \geq \lambda \, \mathbb{E}X) \;\geq\; (1-\lambda)^2 \, \frac{(\mathbb{E}X)^2}{\mathbb{E}X^2}$$

Hint: Use the Cauchy-Schwartz inequality and the identity $X = X\,\mathbb{I}(X \geq c) + X\,\mathbb{I}(X < c)$.

10. Let $X_1, \ldots, X_n$ be independent Bernoulli random variables with $\mathbb{E}X_i = p_i$. Let $S = X_1 + \cdots + X_n$ and let $\mu = \mathbb{E}S = \sum_{i=1}^n p_i$. Use Chernoff's bound and a MGF computation to show that for all $t > \mu$

$$\mathbb{P}(S > t) \leq \exp\{t - \mu - t\log(t/\mu)\}$$

How does this bound compare to Hoeffding's inequality?

11. Let $X \sim \chi^2_k$ have a chi-squared distribution with k degrees of freedom.

(a) Show that if Z is standard normal and $s < 2$ then $\mathbb{E}\exp\{sZ^2\} = (1-2s)^{-1/2}$.

(b) Show that the MGF of X is equal to $\varphi_X(s) = (1-2s)^{-k/2}$.

(c) Use the Chernoff bound to establish that for $0 \leq \epsilon \leq 1$,

$$P(X \leq (1-\epsilon)k) \leq \exp\left\{-\frac{k}{4}(\epsilon^2 - \epsilon^3)\right\}$$

12. *Independent Copies.* Let X, X' be independent random variables with the same distribution. In this case we say that X' is an independent copy of X.

(a) Show that $\mathrm{Var}(X) = \frac{1}{2}\mathbb{E}(X - X')^2$

(b) Argue formally or informally that $\mathbb{E}(X' \,|\, X) = \mathbb{E}X$

(c) Using the result of part (b) and Jensen's inequality for conditional expectations, show that $\mathbb{E}|X - \mathbb{E}X| \leq \mathbb{E}|X - X'|$. This is a key step in establishing a number of important bounds in empirical process theory.

13. Let $X_1, \ldots, X_n \in \mathcal{X}$ be i.i.d. and let $\mathcal{G}$ be a family of function $g : \mathcal{X} \to [-c, c]$. Define

$$f(x_1^n) \;=\; \sup_{g \in \mathcal{G}} \left| n^{-1} \sum_{i=1}^n g(x_i) \;-\; \mathbb{E}g(X) \right|$$

Find the difference coefficients $c_1, \ldots, c_n$ of f, and use these to establish concentration bounds for the random variable $f(X_1^n)$.

14. Let $X_1, \ldots, X_n \in \mathbb{R}^d$ be independent random vectors such that $\mathbb{E}X_i = 0$ and $||X_i|| \leq c_i/2$ with probability one, where $||u|| = (u^t u)^{1/2}$ is the ordinary Euclidean norm. Let $\alpha = (1/4)\sum_{i=1}^n c_i^2$.

(a) Show that $\mathbb{E}||\sum_{i=1}^n X_i|| \leq \sqrt{\alpha}$.

(b) Use the bounded difference inequality and the inequality in part (a) to show that for all $t \geq \sqrt{\alpha}$

$$P\left(||\sum_{i=1}^n X_i|| > t\right) \;\leq\; \exp\left\{-\frac{(t-\sqrt{\alpha})^2}{2\alpha}\right\}$$

15. Let X be a random variable satisfying the concentration type inequality $\mathbb{P}(|X| > t) \leq a\, e^{-b\,t^2}$ for all $t \geq 0$, where $a \geq 1$ and $b \geq 0$. Show that

$$\mathbb{E}|X| \leq \sqrt{\frac{1+\log a}{b}}.$$

Hint: Note that for $s \geq 0$ we have $\mathbb{E}X^2 \leq s + \int_s^\infty \mathbb{P}(X^2 \geq t)\, dt$. Use Cauchy-Schwartz.

16. Let $X_1, \ldots, X_n$ be iid $\sim$ Bern(p). Note that $|X_i - p| \leq \max(p, 1-p)$.

(a) Use Bernstein's inequality to get an upper bound on $\mathbb{P}(n^{-1}\sum_{i=1}^n X_i - p \geq t)$ for $t \geq 0$.

(b) Argue that one can restrict attention to $t \in [0, 1-p]$. Using this fact and the bound in part (a) show that if $p \geq 1/2$ then for all $t \geq 0$

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i - p \geq t\right) \;\leq\; \exp\left\{\frac{-3nt^2}{8p(1-p)}\right\}$$

(c) Compare the bound in part (b) to a naive inequality based on the central limit theorem and tail bounds for the standard normal distribution.

17. Let $X_1, \ldots, X_n$ be random variables with moment generating functions $M_{X_i}(s) \leq M(s)$ for each $s \geq 0$.

(a) Using the argument in class for Gaussian random variables, show that

$$\mathbb{E}\max(X_1, \ldots, X_n) \;\leq\; \inf_{s:s>0} \frac{\log n + \log M(s)}{s}.$$

Suppose now that $U_1, \ldots, U_n$ are Gamma(α, β) random variables.

(b) Show that the moment generating function of U_i is $M(s) = (1 - s\,\beta)^{-\alpha}$.

(c) Using the bound from part (a) and an appropriate choice of s, which can be found by inspection, show that

$$\mathbb{E}\max(U_1, \ldots, U_n) \;\leq\; \frac{2\,\beta\,\log n}{1 - n^{-1/\alpha}}.$$

18. Let $U_1, \ldots, U_n$ be independent Uniform$(0, \theta)$ random variables. Find $\mathbb{E}\left[\max_{1 \leq j \leq n} U_j\right]$.

I. Classification

1. Let (X, Y) be a jointly distributed pair with $X \in \mathcal{X}$ and $Y \in \{0, 1\}$. Suppose that $\mathcal{X}$ is finite and that (X, Y) has joint probability mass function $p(x, y)$.

a. Express the prior probabilities $\pi_0 = \mathbb{P}(Y = 0)$ and $\pi_1 = \mathbb{P}(Y = 1)$ in terms of $p(x, y)$.

b. Express the class conditional probability mass function $p_0(x) = \mathbb{P}(X = x \,|\, Y = 0)$ in terms of $p(x, y)$ and the prior probabilities.

c. Show that the marginal pmf of X can be written as $p(x) = \pi_0\, p_0(x) + \pi_1\, p_1(x)$ where $p_1(x) = \mathbb{P}(X = x \,|\, Y = 1)$.

e. Use Bayes rule to show that $\eta(x) := P(Y = 1 \,|\, X = x) = \pi_1 p_1(x)/p(x)$

2. Consider a classification problem in which the predictor X is uniformly distributed on the unit interval $[0, 1]$ and the response $Y \in \{0, 1\}$ as usual. For $x \in [0, 1]$ let $\eta(x) = \mathbb{P}(Y = 1 \,|\, X = x)$. Specify the Bayes rule ϕ^* and the Bayes risk R^* in each of the following cases.

a. $\eta(x) = 1/3$ for all x

b. $\eta(x) = x$

c. $\eta(x) \in \{0, 1\}$ for all x

In each of the cases above, find the prior probability $\pi_1 = \mathbb{P}(Y = 1)$, or indicate why this is not possible without more information.

3. Let $(X, Y) \in \mathbb{R}^2 \times \{0, 1\}$ be a random predictor-response pair. Suppose that the predictor X is a pair (X_1, X_2) where $X_1, X_2 \in [0, 1]$ are independent, X_1 is uniform on $[0, 1]$, and X_2 has density $g(x_2) = 3x_2^2$ for $0 \leq x_2 \leq 1$. Suppose that $\eta(x_1, x_2) = (x_1 + x_2)/2$.

a. Find the Bayes rule ϕ^* for this problem and identify its decision boundary.

b. Find the unconditional density of X

c. Find the Bayes risk associated with (X, Y)

d. Find the prior probability that $Y = +1$.

e. Find the class-conditional density of X given $Y = 1$.

4. Suppose that you are given access to a database consisting of many email messages that have been labeled as spam or normal. You decide to construct a simple classification rule,

the only feature being whether or not the word "meeting" appears somewhere in the email. Using relative frequencies to estimate probabilities you find the following:

$$\hat{P}(\text{spam}) = .3 \quad \hat{P}(\text{`meeting' present} \,|\, \text{spam}) = .01 \quad \hat{P}(\text{`meeting' present} \,|\, \text{normal}) = .04$$

Using this information, calculate a simple classification rule for spam detection. What can you say about the error rate of your rule on the database?

5. Argue as carefully as you can that if the Bayes risk R^* for a pair (X, Y) is equal to $1/2$ then Y is independent of X.

6. Consider the labeled data set $(-2, 1), (-1, 1), (0, 0), (1, 1), (2, 0) \in \mathbb{R} \times \{0, 1\}$.

   a. Sketch the 1-nearest neighbor rule for this dataset by drawing a line and indicating which points are assigned to zero and which are assigned to one.

   b. Sketch the 3-nearest neighbor rule for this dataset by drawing a line and indicating which points are assigned to zero and which are assigned to one.

7. Let $(X, Y) \in \mathbb{R} \times \{0, 1\}$ be a random predictor-response pair. Suppose that Y has prior probabilities $\pi_1 = \mathbb{P}(Y = 1)$ and $\pi_0 = \mathbb{P}(Y = 0)$, and that X is continuous with marginal density f and class conditional densities f_0 and f_1. Let $\eta(x) = \mathbb{P}(Y = 1 \,|\, X = x)$ as usual.

   a. Show that the Bayes rule ϕ^* can be written in the form

$$\phi^*(x) \;=\; \begin{cases} 1 & \text{if } \log \frac{\eta(x)}{1-\eta(x)} \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

   b. Find a simple expression for the Bayes rule $\phi^*(x)$ in terms of $\pi_1 f_1(x)$ and $\pi_0 f_0(x)$.

Suppose that f_1 is $\mathcal{N}(\mu_1, \sigma^2)$ and that f_0 is $\mathcal{N}(\mu_0, \sigma^2)$ where $\mu_1 > \mu_0$.

   c. Using the results above, find an expression for the Bayes rule $\phi^*(x)$ in terms of the parameters π_0, π_1, μ_0, μ_1, and σ^2.

   d. What is the form of the rule in part (b) when $\pi_1 = 1/2$? Explain why this makes intuitive sense.

   e. Suppose for simplicity that $\mu_1 = u$ and $\mu_0 = -u$ for some $u > 0$. What form does the Bayes rule take when u increases (tends to infinity), and in particular, how does the rule depend on π_1 versus π_0? A informal but clear answer is fine.

8. Consider the setting of linear discriminant analysis in which the class-conditional densities f_0 and f_1 have the multivariate normal form $f_k = \mathcal{N}(\mu_k, \Sigma_k)$.

a. Using the expression for the multivariate normal density, show that the discriminant functions $\delta_k(x) = \log(\pi_k\, f_k(x))$ have the form

$$\delta_k(x) = -\frac{1}{2} x^t \Sigma_k^{-1} x + \langle x, \Sigma_k^{-1}\mu_k\rangle - \frac{1}{2}\left[\log(2\pi)^d \pi_k^{-2} \det(\Sigma_k) + \mu_k^t \Sigma_k^{-1}\mu_k\right]$$

b. Show that when $\Sigma_0 = \Sigma_1 = \Sigma$ the decision boundary $B = \{x : \delta_1(x) = \delta_0(x)\}$ has the form

$$B \;=\; \{x : x^t\, \Sigma^{-1}(\mu_1 - \mu_0) + (c_0 - c_1) = 0\}$$

where c_0, c_1 are real valued constants, and argue that this set is a hyperplane.

9. Let (X, Y) be a jointly distributed pair with $X \in \mathbb{R}^d$ and $Y \in \{0, 1\}$. Suppose that we have added a zeroth component to the vector X that is always equal to 1, so that the augmented vector $X \in \mathbb{R}^{d+1}$. The logistic regression method for binary classification is based on the assumption that

$$\log \frac{\mathbb{P}(Y = 1 \,|\, X = x)}{\mathbb{P}(Y = 0 \,|\, X = x)} \;=\; \log \frac{\eta(x)}{1 - \eta(x)} \;=\; \langle \beta, x\rangle \tag{2}$$

for some vector $\beta \in \mathbb{R}^{d+1}$ of coefficients. In words, equation (2) says that the conditional log-odds ratio of $Y = 1$ vs. $Y = 0$ is linear in the feature vector x.

a. Show, by inverting the relation (2), that

$$\eta(x) = \eta(x : \beta) \;=\; \frac{e^{\langle \beta, x\rangle}}{1 + e^{\langle \beta, x\rangle}} \;=\; \frac{1}{1 + e^{-\langle \beta, x\rangle}}$$

Here we write $\eta(x : \beta)$ to remind ourselves that η depends on β.

b. Equation (2) is sometimes written in the form $\mathrm{logit}(\eta(x)) = \langle \beta, x\rangle$, where $\mathrm{logit}(u) = \log[u/(1-u)]$ for $0 < u < 1$ is the logistic (or logit) function. Sketch the logistic function.

Given a data set $D_n = (x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^{d+1} \times \{0, 1\}$ logistic regression estimates the coefficient vector β in (2) by maximizing the conditional log likelihood function

$$\ell(\beta) \;=\; \log \prod_{i=1}^{n} \mathbb{P}_\beta(Y = y_i \,|\, X = x_i)$$

where $\mathbb{P}_\beta(Y = 1 \,|\, X = x) = \eta(x : \beta)$ and $\mathbb{P}_\beta(Y = 0 \,|\, X = x) = 1 - \eta(x : \beta)$.

c. Use the expression for $\eta(x : \beta)$ in (a) to show that the conditional log likelihood function can be written in the form

$$\ell(\beta) \;=\; \sum_{i=1}^{n} \Big[y_i \langle \beta, x_i \rangle - \log(1 + e^{\langle \beta, x_i \rangle}) \Big]$$

d. Show that $\nabla \ell(\beta) = \sum_{i=1}^{n} x_i \, [y_i - \eta(x_i : \beta)]$. Hint: Evaluate the partial derivative $\partial \ell(\beta) / \partial \beta_j$ for a fixed index j between 1 and d.

10. Let $D_n = (x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \{0, 1\}$ be a data set for classification and let $\gamma = \{A_1, \ldots, A_K\}$ be a partition of $\mathcal{X}$. Define the histogram classification rule $\hat{\phi}_\gamma$ based on γ. Show that $\hat{\phi}_\gamma$ minimizes the training error $R_n(\phi)$ over all classification rules ϕ that are constant on the cells of γ, meaning $\phi(u) = \phi(v)$ if u, v are in the same cell of γ.

11. Recall that the Bayes Rule ϕ^* for a jointly distributed pair (X, Y) with response $Y \in \{0, 1\}$ is defined by

$$\phi^*(x) = \operatorname*{argmax}_{k=0,1} \mathbb{P}(Y = k \,|\, X = x)$$

a. How would you modify this definition in the case where the response takes values in the finite set $\{0, 1, \ldots, K\}$, that is, each feature vector x is associated with one of K possible outcomes?

b. Show that in the binary case $Y \in \{0, 1\}$ the Bayes Rule has the form

$$\phi^*(x) \;=\; \begin{cases} 1 & \text{if } \eta(x) \geq 1/2 \\ 0 & \text{otherwise} \end{cases}$$

12. Let $D_n = (X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{X} \times \{0, 1\}$ be iid observations for a classification problem. Recall that the empirical risk of a fixed classification rule $\phi : \mathcal{X} \to \{0, 1\}$ is defined by

$$\hat{R}_n(\phi) \;=\; \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(\phi(X_i) \neq Y_i)$$

and that the risk of ϕ is $R(\phi) = \mathbb{P}(\phi(X) \neq Y)$.

a. Show that $\mathbb{E}[\hat{R}_n(\phi)] = R(\phi)$

b. Show that $\mathrm{Var}(\hat{R}_n(\phi)) = n^{-1} \, R(\phi)(1 - R(\phi)) \leq 1/(4n)$

c. Argue carefully that $n\hat{R}_n(\phi)$ has a $\mathrm{Bin}(n, R(\phi))$ distribution

d. Use Chebyshev's inequality to show that for $t \geq 0$

$$\mathbb{P}(|\hat{R}_n(\phi) - R(\phi)| \geq t) \;\leq\; \frac{R(\phi)(1 - R(\phi))}{n\,t^2} \;\leq\; \frac{1}{4\,n\,t^2}$$

e. Use Hoeffding's inequality to show that for $t \geq 0$

$$\mathbb{P}(|\hat{R}_n(\phi) - R(\phi)| \geq t) \;\leq\; 2\exp\{-2nt^2\}$$

13. Consider a classification problem in which you have access to a test set containing $m = 120$ iid observations $(X_i, Y_i) \in \mathcal{X} \times \{0, 1\}$. You would like to use the test set to assess the risk of a given rule ϕ using the empirical risk $\hat{R}_m(\phi)$. Chebyshev's inequality and Hoeffding's inequality provide bounds on $\mathbb{P}(|\hat{R}_m(\phi) - R(\phi)| \geq t)$ for $t \geq 0$. Compute and compare these probability bounds, with $m = 120$, at the following values of t: 1/20, 1/11, 1/9, and 1/5.

14. Consider a classification problem in which you would like to assess the risk of a given rule ϕ using its empirical risk $\hat{R}_m(\phi)$ on a test data set D_m. In particular, you wish to determine the size n of the test set necessary to conclude that

$$\mathbb{P}(|\hat{R}_n(\phi) - R(\phi)| \geq \delta) \leq \epsilon$$

Use Chebyshev's and Hoeffding's inequalities to find suitable values for n as a function of δ and ϵ. How do the resulting quantities depend on δ and ϵ? Generally speaking, which inequality permits you to use a smaller test set?

15. Let D_n and D_m be independent training and test sets, respectively. Suppose that the rule $\hat{\phi}_n(x) = \phi_n(x : D_n)$ is derived from the training set.

a. Define the test set error $\hat{R}_m(\hat{\phi}_n)$.

b. Show that $\mathbb{E}[\hat{R}_m(\hat{\phi}_n) \,|\, D_n] = R(\hat{\phi}_n)$

c. What is $\mathbb{E}\hat{R}_m(\hat{\phi}_n)$? Compare this to your answer above.

16. Let $D_n = (x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \{0, 1\}$ be a data set for classification. For each region $A \subseteq \mathcal{X}$ let $|A|$ denote the number of points x_i in A and let $p(A) = |A|^{-1} \sum_{x_i \in A} y_i$ be the fraction of points $x_i \in A$ labeled 1. Suppose that the region A can be expressed as the disjoint union $A = A_1 \cup A_2$ of two other regions.

a. Using the definition, show that

$$p(A) \;=\; \frac{|A_1|}{|A|}p(A_1) + \frac{|A_2|}{|A|}p(A_2)$$

b. Show that $|A| = |A_1| + |A_2|$. Conclude from this and part (a) that for any concave function $f : [0,1] \to \mathbb{R}$

$$f(p(A)) - \frac{|A_1|}{|A|}f(p(A_1)) - \frac{|A_2|}{|A|}f(p(A_2)) \;\geq\; 0$$

This establishes that the impurity differences defined in the lecture for the misclassification, Gini, and entropy impurity measures are non-negative.

c. Let $m(p) = \min(p, 1-p)$. Show that $|A|m(p(A))$ is the number of misclassifications if every point in A is assigned to the majority class.

d. Consider two partitions γ_1 and γ_2 of $\mathcal{X}$ that are identical except that a cell A of γ_1 is split into two cells A_1 and A_2 in γ_2. What can you say about the training error of the corresponding histogram classification rules (based on majority voting in cells)?

17. Let $D_n = (x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^p \times \{\pm 1\}$ be sequence of labeled pairs. Show that the constraint set

$$C := \{w, b : y_i(x_i^t w - b) \geq 1 \text{ for } i = 1, \ldots, n\}$$

appearing in the primal SVM optimization problem is convex. To make things a bit more formal, treat the elements of C as vectors $v = (w_1, \ldots, w_p, b)^t \in \mathbb{R}^{p+1}$. Hint: Show that C is the intersection of n sets, one for each i, and then show that each of these sets is convex.

18. Write down the primal problem, with optimal value $p*$, and argue using the previous question and results from a previous homework that the primal problem is a convex program. Now consider the Lagrangian $L : \mathbb{R}^p \times \mathbb{R} \times \mathbb{R}^n_+$, which is defined by

$$L(w, b, \lambda) \;:=\; \frac{1}{2}||w||^2 \;-\; \sum_{i=1}^{n} \lambda_i \left\{ y_i(w^t x_i - b) - 1 \right\}$$

Establish that

$$\max_{\lambda \geq 0} L(w, b, \lambda) \;=\; \begin{cases} ||w||^2 & \text{if } y_i(x_i^t w - b) \geq 1 \text{ for } i = 1, \ldots, n \\ +\infty & \text{otherwise} \end{cases}$$

To see why this is true, note that if one of the constraints $y_i(x_i^t w - b) \geq 1$ is *not* satisfied, then one can increase the corresponding λ_i to make the Lagrangian arbitrarily large. Using the last display above, argue informally that the primal problem can be written as

$$p^* \;=\; \min_{w,b} \max_{\lambda \geq 0} L(w, b, \lambda)$$

J. Regression

1. Let $(x, y) \in \mathbb{R}^p \times \mathbb{R}$ be a fixed predictor-response pair, and define a function $f : \mathbb{R}^p \to \mathbb{R}$ by $f(\beta) = (y - x^t\beta)^2$.

   a. Show that f is convex.

   b. Now let $D_n = (x_1, y_1), \ldots, (x_n, y_n)$ be n predictor-response pairs. What can you say about the convexity of the sum of squares $g(\beta) = \sum_{i=1}^n (y_i - x_i^t\beta)^2$?

   c. Fix $\lambda \geq 0$ and define the penalized performance criterion

$$h_\alpha(\beta) \;=\; \sum_{i=1}^n (y_i - x_i^t\beta)^2 \;+\; \lambda \sum_{j=1}^p |\beta_j|^\alpha$$

   Argue that h_α is convex if $\alpha \geq 1$. Hint: Recall that a sum of convex functions is convex.

2. Consider a data set with design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ and response vector $\mathbf{y} \in \mathbb{R}^n$. Fix $\lambda > 0$ and define the penalized loss $\hat{R}_{n,\lambda}(\beta) = ||\mathbf{y} - \mathbf{X}\beta||^2 + \lambda\,||\beta||^2$. Following the calculus based arguments for OLS, show that $\hat{R}_{n,\lambda}(\beta)$ has unique minimizer $\hat{\beta}_\lambda = (\mathbf{X}^t\mathbf{X} + \lambda I_p)^{-1}\mathbf{X}^t y$.

3. Let $(X, Y) \in \mathbb{R}^p \times \mathbb{R}$ be a jointly distributed pair following the signal plus noise model $Y = f(X) + \varepsilon$ where ε is independent of X, $\mathbb{E}\varepsilon = 0$, and $\mathrm{Var}(\varepsilon) = \sigma^2$.

   a. Find simple expressions for $\mathbb{E}Y$ and $\mathrm{Var}(Y)$.

   b. Argue that $\mathbb{E}(Y|X) = f(X)$. Thus f is the regression function of Y based on X.

   c. Show that $\varphi = f$ minimizes the risk $R(\varphi) = \mathbb{E}(\varphi(X) - Y)^2$ over prediction rules $\varphi : \mathbb{R}^p \to \mathbb{R}$. What is the minimum value of $R(\varphi)$?

4. Let $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{X} \times \mathbb{R}$ be iid observations from the signal plus noise model $Y = f(X) + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

   a. Define the empirical risk $\hat{R}_n(\varphi)$ of a rule $\varphi : \mathbb{R}^p \to \mathbb{R}$.

   b. Assuming that $\mathrm{Var}(\varphi(X)) < \infty$, find the expectation and variance of $\hat{R}_n(\varphi)$. You may use the fact that $\mathbb{E}\varepsilon^3 = 0$ and $\mathbb{E}\varepsilon^4 = 3\sigma^4$ under our normality assumption.

c. What does Chebyshev's inequality tell you in this setting? What sort of assumptions could you make to control the size of the upper bound?

d. Can you apply Hoeffding's inequality in this case? If so, what is the bound?

5. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^{p+1}$ be fixed vectors with initial component equal to one 1. Suppose that we observe responses $y_1, \ldots, y_n \in \mathbb{R}$ generated from the linear model $y_i = \beta^t \mathbf{x}_i + \varepsilon_i$, where $\beta \in \mathbb{R}^{p+1}$ is an unknown coefficient vector and $\varepsilon_1, \ldots, \varepsilon_n$ are iid $\sim \mathcal{N}(0, \sigma^2)$.

a. Argue that $y_1, \ldots, y_n$ are independent and that $y_i \sim \mathcal{N}(\mathbf{x}_i^t \beta, \sigma^2)$.

b. Find the joint likelihood $L(\beta)$ of $y_1, \ldots, y_n$.

c. Find the log likelihood $\ell(\beta)$ of $y_1, \ldots, y_n$ and show that maximizing $\ell(\beta)$ over β is equivalent to minimizing the empirical risk $\hat{R}_n(\beta) = n^{-1} \sum_{i=1}^n (y_i - \mathbf{x}_i^t \beta)^2$ over β.

d. Define the response vector $\mathbf{y}$ and design matrix $\mathbf{X}$ associated with the data above, giving the dimensions of each. Show carefully that $\hat{R}_n(\beta) = n^{-1} ||y - X\beta||^2$.

6. Let $\mathbf{y}$ and $\mathbf{X}$ be the response vector and design matrix, respectively, associated with observations $(\mathbf{x}_i, y_i)$ of the previous problem. Recall from class that the OLS coefficient $\hat{\beta} = (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \mathbf{y}$

a. Show that $\mathbf{y} = \mathbf{X}\beta + \varepsilon$ with $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 I)$. Conclude that $\mathbf{y} \sim \mathcal{N}_n(\mathbf{X}\beta, \sigma^2 I)$.

b. Show that $\hat{\beta} = \beta + (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \varepsilon$.

c. Find $\mathbb{E}\hat{\beta}$ and $\mathrm{Var}(\hat{\beta})$.

d. Argue that $\hat{\beta} \sim \mathcal{N}_{p+1}(\beta, \sigma^2 (\mathbf{X}^t \mathbf{X})^{-1})$, and conclude that $\hat{\beta}_j \sim \mathcal{N}(\beta_j, \sigma^2 (\mathbf{X}^t \mathbf{X})^{-1}_{jj})$.

e. Use the distribution of $\hat{\beta}_j$ to find a 95% confidence interval for β_j.

7. Let $\mathbf{y}$ and $\mathbf{X}$ be the response vector and design matrix, respectively, associated with observations $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n) \in \mathbb{R}^p \times \mathbb{R}$.

a. Show that $\mathbf{X}^t \mathbf{X} + \lambda I_p$ is symmetric and positive definite if $\lambda > 0$. Conclude that $\mathbf{X}^t \mathbf{X} + \lambda I_p$ is invertible if $\lambda > 0$.

b. Find a simple relationship between the eigenvalues of $\mathbf{X}^t \mathbf{X} + \lambda I_p$ and those of $\mathbf{X}^t \mathbf{X}$.

8. Let $\hat{\beta}_\lambda$ be the minimizer of $\hat{R}_{n,\lambda}(\beta) = ||\mathbf{y} - \mathbf{X}\beta||^2 + \lambda ||\beta||^2$.

a. Show that $\hat{\beta}_0$ is the usual OLS estimator (when the rank of X is equal to p).

b. Show that $||\mathbf{y} - \mathbf{X}\hat{\beta}_\lambda||^2 \leq ||\mathbf{y} - \mathbf{X}\beta||^2$ for every β such that $||\beta|| \leq ||\hat{\beta}_\lambda||$. Hint: Assume the stated inequality fails to hold and show that this implies that $\hat{\beta}_\lambda$ is not the minimizer of $\hat{R}_{n,\lambda}(\beta)$.