STOR 565 Homework

Show all work. Note: all logarithms are natural logarithms.

1. 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 $F_x(t) = 1$?

2. 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 $y > 0$.

3. 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 $n^{-1} \sum_{i=1}^n \langle \mathbf{x}_i, \mathbf{v} \rangle^2 = \mathbf{v}^t \mathbf{S} \mathbf{v}$.

4. Let $x = (x_1, \ldots, x_d)^t$ be a vector in $\mathbb{R}^d$.

a. Show that $||x|| \leq |x_1| + \cdots + |x_d|$. 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 $|x_1| + \cdots + |x_d| \leq ||x|| \, d^{1/2}$. Find an example where the bound holds with equality.

5. Let X, X' be independent random variables with the same distribution. In this case we say that X' is an independent copy of X. Show that $\text{Var}(X) = \frac{1}{2}\mathbb{E}(X - X')^2$

6. 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

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

7. Let $x = x_1, \ldots, x_n$ be a univariate sample, and let $\tilde{x} = \tilde{x}_1, \ldots, \tilde{x}_n$ be the standardized version of x with $\tilde{x}_i = (x_i - m(x))/s(x)$. Show that $m(\tilde{x}) = 0$ and $s(\tilde{x}) = 1$.

8. 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 $+1$.

9. (Norms of outer products) Let $u \in \mathbb{R}^k$ and $v \in \mathbb{R}^l$ 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 $||\mathbf{v}||$.