STOR 565 Homework

1. Some general questions about clustering.

   a. Describe the difference between a complete clustering scheme/algorithm, and an incomplete one.

   b. How is clustering different from classification?

2. Some general questions about binary trees.

   a. Draw a binary tree with 3 nodes. How many leaves does it have? How many internal nodes does it have?

   b. Draw a binary tree with 5 nodes. How many leaves does it have? How many internal nodes does it have?

   c. Draw essentially different binary trees with 7 nodes. Do they have the same number of internal nodes? Do they have the same number of leaves?

   d. Formulate a conjecture about the relationship between the number of internal nodes and the number of leaves in a rooted binary tree.

   e. (Optional) Prove your conjecture using induction.

3. *Measuring the variability of a set of vectors.* Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ be a sample containing n observations of dimension p. 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 mean

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

   If the general case seems difficult, consider first the case when $p = 1$. You may also want to follow the outline in Exercise 4.1 on p.87 of Boyd and Vanderberghe.

Now let $\mathbf{X} = [\mathbf{x}_1^t, \ldots, \mathbf{x}_n^t]^t$ be the $n \times p$ data matrix associated with the sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$. 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 real numbers.

b. Express the $p \times p$ sample covariance matrix $\hat{\Sigma} = \{s(\mathbf{x}_{\cdot j}, \mathbf{x}_{\cdot k}) : 1 \leq j, k \leq p\}$ in terms of $\mathbf{X}$ and $\overline{\mathbf{x}}$. Note: we have *not* assumed that $\overline{\mathbf{x}} = 0$.

c. Carefully describe V_1 and V_2 in plain English.

d. Provide an equivalent expression for V_1 in terms of the diagonal entries of $\hat{\Sigma}$.

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

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

g. Show that

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

h. Using the identity from part (g), and some additional calculations, show that

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

4. 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 $\nabla f(\mathbf{x}) = 2\mathbf{A}\mathbf{x}$.

b. Show that $\nabla^2 f(\mathbf{x}) = 2\mathbf{A}$.

5. Let $w \in \mathbb{R}^d$ be a vector and $b \in \mathbb{R}$ a constant. Show that the sets $C = \{x : w^t x - b \geq 0\}$ and $D = \{x : w^t x = b\}$ in $\mathbb{R}^d$ are convex.

6. Show that if $C_1, \ldots, C_n \subseteq \mathbb{R}^d$ are convex then so is their intersection $\cap_{i=1}^n C_i$.