# **Homework Six**, for Tue 3/1 CSE 250B

**Your homework must be typeset, and the PDF file should be uploaded to Gradescope by midnight on the due date.**

1. *Experiments with clustering.* For this problem, we'll be using the *animals with attributes* data set. Go to

   `http://attributes.kyb.tuebingen.mpg.de`

   and, under "Downloads", choose the "base package" (the very first file in the list). Unzip it and look over the various text files.

   (a) This is a small data set that has information about 50 animals. The animals are listed in `classes.txt`. For each animal, the information consists of values for 85 features: does the animal have a tail, is it slow, does it have tusks, etc. The details of the features are in `predicates.txt`. The full data consists of a $50 \times 85$ matrix of real values, in `predicate-matrix-continuous.txt`.

   (b) Load the real-valued array, and also the animal names. Run $k$-means on the data and ask for $k = 10$ clusters. For each cluster, list the animals in it. Does the clustering make sense?

   Python notes: you can find an implementation of $k$-means in `sklearn.cluster`.

   (c) Now hierarchically cluster this data, using average linkage (ideally, Ward's method). Plot the resulting dendrogram. Does the hierarchical clustering seem sensible to you?

   Python notes: Use `scipy.cluster.hierarchy.linkage`. In the `dendrogram` method, set the `orientation` parameter to `'right'` and label each leaf with the corresponding animal name. You will need to make the plot larger by prefacing your code with

   ```
   from pylab import rcParams
   rcParams['figure.figsize'] = 5, 10
   ```

   (or try a different size if this doesn't seem quite right).

   To turn in:

   - The list of $k$-means clusters
   - The dendrogram

2. *Placement of the cluster center.* Let's return to the topic of $k$-means clustering. For a cluster of points $C \subset \mathbb{R}^p$, the optimal placement of the center $\mu$ is the point for which

   $$\sum_{x \in C} \|x - \mu\|^2$$

   is minimized. Here $\|\cdot\|$ denotes $\ell_2$ distance.

   (a) Show (using calculus or otherwise) that this optimal center $\mu$ is simply the mean of the points $C$.

   (b) Show (using a small example in one dimension) that this is no longer true if $\ell_1$ distance is used. Can you characterize the optimal center location in the $(\mathbb{R}^1, \ell_1)$ case?

3. *A bad case for $k$-means.* Consider the following data set consisting of five points in $\mathbb{R}^1$:

   $$-10, -8, 0, 8, 10.$$

   We would like to cluster these points into $k = 3$ groups.

   (a) What is the optimal $k$-means solution? (Just give the centers.)

   (b) Suppose we call Lloyd's $k$-means algorithm on this data, with $k = 3$. Exhibit a particular initialization of the centers (to three distinct data points) under which the final answer is suboptimal.

4. *An experiment with PCA.* The *animals with attributes* data, described above, represents each of 50 animals as a vector in $\mathbb{R}^{85}$.

   We would like to visualize these animals in 2-d. Show how to do this with a PCA projection from $\mathbb{R}^{85}$ to $\mathbb{R}^2$. Show the position of each animal, and label them with their names. (Python notes: remember how to enlarge the figure.) Does this *embedding* seem sensible to you?

5. *Projections.* Let $u_1, u_2 \in \mathbb{R}^p$ be two vectors with $\|u_1\| = \|u_2\| = 1$ and $u_1 \cdot u_2 = 0$. Define $U$ to be the matrix whose columns are $u_1$ and $u_2$.

   (a) What are the dimensions of each of the following?

   - $U$
   - $U^T$
   - $UU^T$
   - $u_1 u_1^T$

   (b) What are the differences, if any, between the following four projections?

   - $x \mapsto (u_1 \cdot x, u_2 \cdot x)$
   - $x \mapsto (u_1 \cdot x)u_1 + (u_2 \cdot x)u_2$
   - $x \mapsto U^T x$
   - $x \mapsto UU^T x$