CSE 151: Machine learning **Winter 2019**

Homework 8

Submission instructions:

- Please type up your solutions.
- If a problem asks for a numerical answer, you need only provide this answer. There is no need to show your work, unless you would like to.
- Upload the PDF file for your homework to `gradescope` by 6pm on Tuesday March 12.

Part A: Clustering and informative projections

1. *Suboptimality of Lloyd's algorithm.* 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? Give the locations of the centers as well as the k-means cost.

   (b) Suppose we call Lloyd's k-means algorithm on this data, with $k = 3$ and with initialization $\mu_1 = -10, \mu_2 = -8, \mu_3 = 0$. What is the final set of cluster centers obtained by the algorithm? What is the k-means cost of this set of centers?

2. *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?
   - $x \mapsto (u_1 \cdot x, u_2 \cdot x)$
   - $x \mapsto (u_1 \cdot x)u_1 + (u_2 \cdot x)u_2$
   - $x \mapsto U^T x$
   - $x \mapsto UU^T x$

3. A certain random variable $X \in \mathbb{R}^3$ has mean and covariance as follows:

   $$\mathbb{E}X = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \quad \text{cov}(X) = \begin{pmatrix} 5 & -3 & 0 \\ -3 & 5 & 0 \\ 0 & 0 & 4 \end{pmatrix}$$

(a) Consider the direction $u = (1, 1, 1)/\sqrt{3}$. What are the mean and variance of $X \cdot u$?

(b) The eigenvectors of $\text{cov}(X)$ can be found in the following list; which ones are they?

$$\begin{pmatrix}1\\0\\0\end{pmatrix}, \quad \begin{pmatrix}0\\1\\0\end{pmatrix}, \quad \begin{pmatrix}0\\0\\1\end{pmatrix}, \quad \frac{1}{\sqrt{2}}\begin{pmatrix}1\\1\\0\end{pmatrix}, \quad \frac{1}{\sqrt{2}}\begin{pmatrix}0\\1\\1\end{pmatrix}, \quad \frac{1}{\sqrt{2}}\begin{pmatrix}1\\-1\\0\end{pmatrix}$$

(c) Find the eigenvalues corresponding to each of the eigenvectors in part (b). Make it clear which eigenvalue belongs to which eigenvector.

(d) Suppose we used principal component analysis (PCA) to project points X into *two* dimensions. Which directions would it project onto?

(e) Continuing from part (d), what would be the resulting two-dimensional projection of the point $x = (4, 0, 2)$?

(f) Continuing from part (e), suppose that starting from the 2-d projection, we tried to reconstruct the original x. What would the three-dimensional reconstruction be, exactly?

4. M is a 2×2 real-valued symmetric matrix with eigenvalues $\lambda_1 = 2, \lambda_2 = -1$ and corresponding eigenvectors

$$u_1 = \frac{1}{\sqrt{5}}\begin{pmatrix}2\\1\end{pmatrix}, \quad u_2 = \frac{1}{\sqrt{5}}\begin{pmatrix}-1\\2\end{pmatrix}.$$

(a) What is M?

(b) What are the eigenvalues of the matrix $M + 2I$?

(c) What are the eigenvalues of the matrix $M^2 = MM$?

Part B: Programming problems

For this homework, we will work with 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.

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 50×85 matrix of real values, in `predicate-matrix-continuous.txt`.

Load the real-valued array, and also the animal names.

1. 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
```

Turn in the resulting dendrogram.

2. We would also 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: you might need a different size, e.g. `rcParams['figure.figsize'] = 10, 10`. In order to annotate each point with the name of the animal, use `pyplot.annotate`.

   Turn in the resulting plot. Does this *embedding* seem sensible to you?