Math 156 Term Project 4

Deadline: Friday May 29th 11:59pm

Problem 1

Let $\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^N \subset \mathbb{R}^D$ be a dataset of N points. Recall that the sample mean is given by

$$\bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}^{(i)},$$

and the sample covariance matrix is given by

$$\mathbf{S} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}^{(i)}\mathbf{x}^{(i)^T} - \bar{\mathbf{x}}\bar{\mathbf{x}}^T \in \mathbb{R}^{D\times D}.$$

Take the orthogonal eigendecomposition of the sample covariance matrix $\mathbf{S} = \mathbf{U}\mathbf{L}\mathbf{U}^T$, where $\mathbf{U} = [\mathbf{u}_1, \cdots, \mathbf{u}_D]$ consisting of the orthonormal eigenbasis, arranged in the order so that their associated eigenvalues $\mathbf{L} = \text{diag}(\lambda_1, \ldots, \lambda_D)$ are in descending order, i.e. $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_D \geq 0$. These eigenvectors are called the *principal components of the dataset*.

(a) Show that every data point $\mathbf{x}^{(i)} \in \mathcal{D}$ can be written in the form

$$\mathbf{x}^{(i)} = \bar{\mathbf{x}} + \sum_{j=1}^{D} \mathbf{u}_j^T(\mathbf{x}^{(i)} - \bar{\mathbf{x}})\mathbf{u}_j. \qquad \text{(10 points)}$$

Further more, show that this identity can be expressed in matrix form,

$$\mathbf{x}^{(i)} = \bar{\mathbf{x}} + \mathbf{U}\mathbf{U}^T(\mathbf{x}^{(i)} - \bar{\mathbf{x}}). \qquad \text{(5 points)}$$

(b) Let $V \subset \mathbb{R}^d$ be an arbitrary M-dimensional subspace spanned by the orthonormal vectors $\{\mathbf{v}_1, \cdots, \mathbf{v}_M\}$. The orthogonal projection π_V of a vector $\mathbf{x} \in \mathbb{R}^D$ onto V is given by

$$\pi_V(\mathbf{x}) = \sum_{j=1}^{M} \mathbf{v}_j^T\mathbf{x}\mathbf{v}_j. \qquad .$$

Let $\mathbf{V} = [\mathbf{v}_1, \cdots, \mathbf{v}_M] \in \mathbb{R}^{D\times M}$, show that $\pi_V(\mathbf{x}) = \mathbf{V}\mathbf{V}^T\mathbf{x}$ (10 points).

(c) Principal Component Analysis (PCA) reduces the dimensionality of the data by projecting the data onto the subspace spanned by the first M principal components ($M \leq D$). In particular, the first M eigenvectors $\mathbf{u}_1, \cdots, \mathbf{u}_M$ of the sample covariance matrix $\mathbf{S}$ spans the *principal subspace* $\mathcal{P}_M$. PCA projects the data points onto this M-dimensional subspace $\mathcal{P}_M$ by utilizing the orthogonal projection discussed in part (b).

$$\tilde{\mathbf{x}}^{(i)} = \bar{\mathbf{x}} + \pi_{\mathcal{P}_M}(\mathbf{x}^{(i)} - \bar{\mathbf{x}}).$$

The mean squared error of the PCA projection onto the M-dimensional principal subspace $\mathcal{P}_M$ is

$$E_M = \frac{1}{N}\sum_{i=1}^{N} \|\tilde{\mathbf{x}}^{(i)} - \mathbf{x}^{(i)}\|^2.$$

Show that

$$E_M = \sum_{j=M+1}^{D} \lambda_j. \qquad (25 \text{ points})$$

Hence by looking at the eigenvalues of the sample covariance matrix $\mathbf{S}$ we can decide how to choose the cutoff value M for the dimension.
Hint: There's a chance that you might like the following block matrix form

$$\mathbf{U} = [\mathbf{u}_1, \cdots, \mathbf{u}_D] = [\ \mathbf{U}_1 \mid \mathbf{U}_2\ ]$$

where $\mathbf{U}_1 = [\mathbf{u}_1 \cdots, \mathbf{u}_M] \in \mathbb{R}^{D\times M}$ consists of the first M columns of $\mathbf{U}$, and $\mathbf{U}_2 = [\mathbf{u}_{M+1}, \cdots, \mathbf{u}_D] \in \mathbb{R}^{D\times(D-M)}$ consists of the other columns of $\mathbf{U}$.

Problem 2

In this problem, you will implement PCA and use it for high-dimensional data visualization. There are a few components in this problem that you can make it with your own style. First, as usual, you can use any programming language and any external, openly available library to finish our tasks. Secondly, you can select the data set that you're using. For example, you can use the Iris Data Set from UCI Machine Learning Repository. Perform PCA on the data set with effective dimension $M = 2$. Plot the result as Figure 12.8 in the textbook (p. 569) while coloring data points from different class in different color/shape. (50 points)

Problem 3 (Bonus)

Generate your own data set $\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^{N}$ with $N = 600$ data points and $\mathbf{x}^{(i)} \in \mathbb{R}^D, D = 56$ (56 attributes). The data set should consist of three blobs,

$$\mathbf{x}^{(i)} \sim \mathcal{N}(\mathbf{0}, 0.25\mathbf{I}) \text{ for } i = 1, \cdots, 200,$$
$$\mathbf{x}^{(i)} \sim \mathcal{N}(\mathbf{u}, 0.25\mathbf{I}) \text{ for } i = 201, \cdots, 400,$$
$$\mathbf{x}^{(i)} \sim \mathcal{N}(-\mathbf{u}, 0.25\mathbf{I}) \text{ for } i = 401, \cdots, 600,$$

where $\mathbf{u} \in \mathbb{R}^D$ is a random unit vector. (I'm okay with you using uniform distribution in the cube $[0, 1]^D$ and normalize. The process of generating an unit vector in a true random manner is not important to our task.) Perform the same task of high-dimensional data visualization as in Problem 2. Explain your result. (up to 20 points)