MATH 373: Introduction to Machine Learning
Homework 2

Computational Problems

For each of these problems, feel free to use R and/or Python.

1. Using code like that demonstrated in class, download the .png file containing an image of a house posted to Redfin in the Data folder on the course website. Note, you may have to download this first and then open it from your own computer. Set X to be the pixel intensity associated with the red color in the image using code like that performed in class. Answer the following questions:

   (a) What are the dimensions of X? Plot a histogram of the pixel intensities within the image.

   (b) Ensure that the columns X are centered and perform PCA on this image. Plot the scree plots for this data, which illustrate the percentage variation explained against the number of principal components and the cumulative percentage variation explained against the number of principal components. How many PCs are needed to explain 90% of the total variation of X?

   (c) For $d = 1, 5, 10, 15, 20, 30, 50, 100, 200$ project the image onto the first d principal components and plot the resulting compressed image for each d. For each of the nine plots, include the cumulative percentage variation explained by the projection.

2. We will now use PCA and PCR to analyze data from the Cancer Genome Atlas (TCGA). The data contains the gene expression of 217 patients who were classified as either having a "Luminal A" breast cancer tumor or a "Basal" breast cancer tumor. The data we will examine contains a sample of 2000 randomly selected genes associated with each patient.

   Load the TCGA data from the TCGA example .txt file in the Data folder on the course website. Set the first column of the matrix aside as a variable *Tumor.Type*. Further, set $y =$ to expressions of the first gene. Keep the remaining as the design matrix X. Answer the following questions.

   (a) First, randomly select 80% of the rows (patients) and keep them as the training set. Set the remaining 20% aside as the test set. Perform PCA on the training set. Plot the scree plots for the resulting PCs. What number of PCs are needed to explain 85% of the variation in the training set?

   (b) Plot a pairwise scatter plot of the first 4 PCs on the training data and color the scores according to the breast cancer *Tumor.Type* of the patient. Discuss any trends that the pairwise scatter plot reveal, if any.

(c) Run a principal component regression of the first gene on the remaining genes in the training set. Use 10-fold cross-validation to determine the number of PCs to use, and plot the associated MSPE of the cross validation trials against the number of PCs. What is your chosen model? (i.e how many principal components are you using?)

(d) Apply your model from (c) on the test set and calculate the MSPE.

Conceptual Problems

The following questions are to make sure that you understand some of the linear algebra theory needed to say something about what exactly PCA is doing.

1. Consider a data set consisting of four points in $\mathbb{R}^2$

$$\mathbf{x}_1 = (1,2),\ \mathbf{x}_2 = (-1,2),\ \mathbf{x}_3 = (2,-1),\ \mathbf{x}_4 = (2,1)$$

(a) Write down the data matrix $\mathbf{X}_0$ having rows $\mathbf{x}_1, \dots, \mathbf{x}_4$.

(b) Column center $\mathbf{X}_0$ so that each column has mean zero. This is equivalent to replacing each observation $\mathbf{x}_i$ by the centered observation $\tilde{\mathbf{x}}_i = \mathbf{x}_i - 4^{-1}\sum_{i=1}^{4}\mathbf{x}_i$. Check that $\sum_{i=1}^{4}\tilde{\mathbf{x}}_i = \mathbf{0}$, and draw a plot (by hand) of the points $\tilde{\mathbf{x}}_i$. Call the recentered data matrix $\mathbf{X}$.

(c) Calculate the 2×2 empirical covariance matrix $\mathbf{\Sigma} = \frac{1}{4}\mathbf{X}^T\mathbf{X}$.

(d) Calculate the eigenvalues of $\mathbf{\Sigma}$. Is $\mathbf{\Sigma}$ invertible? If so, find $\mathbf{\Sigma}^{-1}$. *Note: it is OK to use software for this.*

(e) Find orthonormal eigenvectors of $\mathbf{\Sigma}$. *Note: it is OK to use software for this.*

(f) What is the best one-dimensional subspace (line) for approximating $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$, $\mathbf{x}_4$? Draw this vector on your plot in (a).

2. Let $\mathbf{u}$ and $\mathbf{v}$ be $n \times 1$ vectors. Let $||\cdot||^2$ denote the squared Euclidean norm where

$$||\mathbf{x}||^2 = \sum_{i=1}^{n} x_i^2$$

Prove the following two assertions:

(a) $||\mathbf{u}-\mathbf{v}||^2 = ||\mathbf{u}||^2 - 2\mathbf{u}^T\mathbf{v} + ||\mathbf{v}||^2$

(b) Now, assume that $||\mathbf{v}||^2 = 1$. Show that $||\mathbf{u} - (\mathbf{u}^T\mathbf{v})\mathbf{v}||^2 = ||\mathbf{u}||^2 - (\mathbf{u}^T\mathbf{v})^2$

3. Let X be an observed $n \times p$ data matrix. Answer the following questions:

   (a) Suppose that we would like to fit a model:

   $$y = X\beta + \epsilon$$

   Describe two situations in which it would be beneficial to use principal component regression. Describe these situations in terms of the matrix X.

   (b) Suppose that we perform principal component analysis on the matrix X and thus calculate

   $$\mathbf{z}_m = \sum_{j=1}^{p} \phi_{jm} \mathbf{x}_j$$

   for $m = 1, \ldots, d \leq p$. In terms of the covariance matrix $\Sigma = n^{-1} X^T X$, $\mathbf{z}_m$ and ϕ_{jm}, define what I mean by a) principal components, b) principal component scores, and c) principal component loadings associated with X.

   (c) In terms of X, name two important features of the first d principal components that explain why they are desirable low-dimensional representations of X.

   (d) In terms of Σ, what is the total variation in X explained by the first d principal components.