EECS 127/227AT Optimization Models in Engineering
Spring 2019 Homework 3

Release date: 9/19/19.

Due date: 9/26/19, 23:00 (11 pm).

Please LaTeX or handwrite your homework solution and submit an electronic version.

Submission Format
Your homework submission should consist of a single PDF file that contains all of your answers (any handwritten answers should be scanned) as well as your IPython notebook saved as a PDF.

If you do not attach a PDF "printout" of your IPython notebook, you will not receive credit for problems that involve coding. Make sure that your results and your plots are visible. Assign the IPython printout to the correct problem(s) on Gradescope.

1. Interpreting the data matrix

In several areas such as machine learning, statistics and data analysis you come across a data matrix X. Sometimes this matrix has dimensions $\mathbb{R}^{m\times n}$ while other times it has dimensions $\mathbb{R}^{n\times m}$ and it can get really confusing as to what exactly it represents. In this problem, we describe a way of interpreting the data matrix.

First, what exactly is a data matrix? As the name suggests, it is a collection of data points. Suppose you are collecting data about courses offered in EECS department in Fall 2019. Each course has certain attributes or features that you are interested in. Possible examples of features are the number of students in the course, the number of GSIs in the course,the number of units the course is worth, the size of the classroom that the course was taught in, the difficulty rating of the course in numerical (1-5) scale and so on. Suppose there were a total of 20 courses and for each course we have 10 features. Then we have 20 data points, with each data point being a 10-dimensional vector. We can arrange this in a matrix of size 20×10, where each row corresponds to values of different features for the same point, and each column corresponds to values of same feature for different points.

Generalizing this, suppose we have n data points with each point containing values for m features (i.e each point lies in m-dimensional space) then our data matrix X would be of size $n \times m$, i.e. $X \in \mathbb{R}^{n\times m}$. We can interpret X in the following two ways:

(a) $X = \begin{bmatrix} \leftarrow \mathbf{x}_1^\top \rightarrow \\ \leftarrow \mathbf{x}_2^\top \rightarrow \\ \vdots \\ \leftarrow \mathbf{x}_n^\top \rightarrow \end{bmatrix}$.

Here $\mathbf{x}_i \in \mathbb{R}^m,\ i = 1, 2, \ldots, n$, and $x_i^\top$ is a row vector that contains values of different features for the ith data point.

(b) $X = \begin{bmatrix} \uparrow & \uparrow & \cdots & \uparrow \\ \mathbf{x}^1 & \mathbf{x}^2 & \ldots & \mathbf{x}^m \\ \downarrow & \downarrow & \cdots & \downarrow \end{bmatrix}$.

Here $\mathbf{x}^j \in \mathbb{R}^n, j = 1, 2, \ldots, m$ and $\mathbf{x}^j$ is a column vector that contains values of the the jth feature for different data points. Note that in several places you will encounter the case where the columns are referred to as $\mathbf{x}_1, \mathbf{x}_2, \ldots$ instead but it is important to understand the context and be clear on what the column represents.

Consider the matrix X as described above. We explore how we can manipulate the data matrix to get some desirable properties.

(a) Suppose we want to compute a vector that contains the mean value for each feature. What is the length of the vector containing mean value of the features? Which of the following python commands will give us the mean value of the features:

i. feature_means = numpy.mean(X, axis = 0)
ii. feature_means = numpy.mean(X, axis = 1)

(b) Suppose we want to compute the standard deviation of each feature. What is the dimension of the vector containing standard deviation of the features? Which of the following python commands will give us the standard deviation of the features:

i. feature_stddevs = numpy.std(X, axis = 0)
ii. feature_stddevs = numpy.std(X, axis = 1)

(c) Suppose we want every feature *"centered"*, i.e. the feature is zero mean. How would you achieve this?

(d) Suppose we want every feature *"standardized"*, i.e the feature is zero mean and has unit variance. How would you achieve this?

(e) Another metric of interest is the covariance matrix, which tells us how different features are related to each other. What is the size of the covariance matrix?:

i. $n \times n$
ii. $m \times m$.

Hint: Is the number of features m or n?

(f) For rest of the problem assume that the data matrix is centred so every feature is zero mean. Let C denote the covariance matrix. Show that C can be represented in the following ways:

$$C = \frac{X^\top X}{n}$$

$$C = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top.$$

Recall that $\mathbf{x}_i^\top$ is the ith row of X.

(g) Let the (i, j) entry of C (c_{ij}) denote the covariance between feature i and feature j. Then which of the following is true?

i. $c_{ij} = \frac{1}{m}(\mathbf{x}^i)^\top \mathbf{x}^j$
ii. $c_{ij} = \frac{1}{n}(\mathbf{x}^i)^\top \mathbf{x}^j$.

Recall that $\mathbf{x}^i$ is the ith column of X.

(h) Recall that our data points are the rows of X and these lie in a m-dimensional space. Suppose we are interested in taking the projection of the points onto a one-dimensional subspace in $\mathbb{R}^m$ spanned by the unit vector u. Sometimes this is referred to informally as "Projecting points along direction u". Then which of the following is true:

i. $\mathbf{u} \in \mathbb{R}^m$

ii. $\mathbf{u} \in \mathbb{R}^n$

Hint: Think about how many points we have and what dimension a single point lies in.

(i) Note there are three different interpretations of the term "projection" and these are used interchangeably with abuse of notation which can make it confusing at times.
Consider vectors $\mathbf{a}$ and $\mathbf{b}$ in $\mathbb{R}^n$. Let $\mathbf{b}$ be unit norm (i.e $\mathbf{b}^\top \mathbf{b} = 1$). Then we have:

i. The **vector projection** of $\mathbf{a}$ on $\mathbf{b}$ is given by $(\mathbf{a}^\top \mathbf{b})\mathbf{b}$. Note that is a vector in $\mathbb{R}^n$.

ii. The **scalar projection** of $\mathbf{a}$ on $\mathbf{b}$ is given $\mathbf{a}^\top \mathbf{b}$. This is a scalar but can take both positive and negative values.

iii. The **projection length** of $\mathbf{a}$ on $\mathbf{b}$ is given by $|\mathbf{a}^\top \mathbf{b}|$, and is the absolute value of the scalar projection.

Recall that our data points are the rows of X. Suppose we want to obtain a column vector, $\mathbf{z} \in \mathbb{R}^n$ containing scalar projections of points along the direction given by the unit vector $\mathbf{u}$. Show that this is given by,

$$z = X\mathbf{u}.$$

(j) Suppose we treat $\mathbf{z} = (z_1, z_2, \ldots z_n)$ as samples of a random variable $Z \in \mathbb{R}$ corresponding to the scalar projection along direction $\mathbf{u}$. We are interested in calculating empirical variance of the scalar projections. Show that this can be calculated as

$$\sigma_z^2 = \frac{1}{n}\mathbf{u}^\top X^\top X\mathbf{u} = \mathbf{u}^\top C\mathbf{u}.$$

Hint: The empirical variance is given by, $\sigma_z^2 = \frac{1}{n}\sum_{i=1}^{n}(z_i - \mu_z)^2$, *where* $\mu_z = \frac{1}{n}\sum_{i=1}^{n} z_i$, *is the empirical mean. Recall that X is assumed to be centered.*

2. Computation and Geometric Interpretation of Singular Value Decomposition (SVD)

Consider the 2×2 matrix

$$A = \frac{1}{\sqrt{10}} \begin{pmatrix} 2 \\ 1 \end{pmatrix} \begin{pmatrix} 1 & -1 \end{pmatrix} + \frac{2}{\sqrt{10}} \begin{pmatrix} -1 \\ 2 \end{pmatrix} \begin{pmatrix} 1 & 1 \end{pmatrix}.$$

(a) What is an SVD of A? Express it as $A = USV^\top$, with S the diagonal matrix of singular values ordered in decreasing fashion. Make sure to check all the properties required for U, S, V.

(b) Find the semi-axis lengths and principal axes of the ellipsoid

$$\mathcal{E}(A) = \left\{ Ax \ : x \in \mathbb{R}^2, \ \ \|x\|_2 \le 1 \right\}.$$

Hint: Use the SVD of A to show that every element of $\mathcal{E}(A)$ is of the form $y = U\bar{y}$ for some element $\bar{y}$ in $\mathcal{E}(S)$. That is, $\mathcal{E}(A) = \{U\bar{y} \ : \ \bar{y} \in \mathcal{E}(S)\}$. (In other words the matrix U maps $\mathcal{E}(S)$ into the set $\mathcal{E}(A)$.) Then analyze the geometry of the simpler set $\mathcal{E}(S)$.

(c) What is the set $\mathcal{E}(A)$ when we append a zero vector after the last column of A, that is A is replaced with $\tilde{A} = [A, 0] \in \mathbb{R}^{2\times 3}$?

(d) Same question when we append a row after the last row of A, that is, A is replaced with $\tilde{A} = [A^\top, 0]^\top \in \mathbb{R}^{3\times 2}$. Interpret geometrically your result.

3. PCA and low-rank compression

We are given a $m \times n$ matrix $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$, with $\mathbf{x}_i \in \mathbb{R}^m$, $i = 1, \ldots, n$ being the data points. (Note that this is the transpose of the $n \times m$ data matrix seen in the previous exercise.) We assume that the data matrix is centered, in the sense that $\mathbf{x_1} + \ldots + \mathbf{x_n} = \mathbf{0}$. In lecture, it was asserted that there is equivalence between three problems:

(P_1) Finding a line going through the origin that maximizes the variance of the points projected on the line.

(P_2) Finding a line going through the origin that minimizes the sum of squares of the distances from the points to their projections;

(P_3) Finding a rank-one approximation to the data matrix.

In this exercise, you are asked to show the equivalence between these three problems.

(a) Consider the problem of projecting a point $\mathbf{x}$ on a line $\mathcal{L} = \{\mathbf{x}_0 + v\mathbf{u} \,:\, v \in \mathbb{R}\}$, with $\mathbf{x}_0 \in \mathbb{R}^m$, $\mathbf{u}^T\mathbf{u} = 1$, given.
Show that the projected point $\mathbf{z}$ is given by

$$\mathbf{z} = \mathbf{x}_0 + v^*\mathbf{u},$$

where we define

$$v^* = (\mathbf{x} - \mathbf{x}_0)^\top \mathbf{u},$$

and that the minimal squared distance $\|\mathbf{z} - \mathbf{x}\|_2^2$ is equal to $\|\mathbf{x} - \mathbf{x}_0\|_2^2 - ((x - x_0)^T u)^2$.

(b) Show that problems P_1, P_2 are equivalent.

(c) Show that P_3 is equivalent to P_1.
Hint: Show that the data matrix is rank-one if and only if it can be expressed as the outer product of two vectors. Recall that P_3 involves minimization of a Frobenius norm, and try to use some properties of this norm.

Bonus. Find the rank-one approximation of the entire EECS127/227AT material. What are the subjects that are the most likely to be at the finals?

4. PCA and face analysis

We have seen how to represent images as matrices. Similarly, we may represent an image as a vector. In this exercise we will explore analyzing a dataset of images represented as vectors.

We will analyze a dataset of human faces to see if we can learn anything meaningful from it. One way to analyze a dataset is to look for directions of maximal variation. That is, in the space of the data we look for the directions in which the data samples change value the most.

Let $\mathbf{x}_i \in \mathbb{R}^m, i = 1, \ldots, n$, be the given data points to analyze, denote $\hat{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i$, the mean of the data points with $\tilde{\mathbf{x}_i} = \mathbf{x}_i - \hat{\mathbf{x}}$, the centered datapoint. Let

$$\tilde{X} = \begin{bmatrix} \leftarrow \tilde{\mathbf{x}}_1^\top \rightarrow \\ \leftarrow \tilde{\mathbf{x}}_2^\top \rightarrow \\ \vdots \\ \leftarrow \tilde{\mathbf{x}}_n^\top \rightarrow \end{bmatrix} \in \mathbb{R}^{n \times m}$$

be the matrix such that the ith row is $\tilde{\mathbf{x}}_i^\top$.
Then the components of the centered data along a direction (by direction, we mean unit norm vector) $\mathbf{z}$ are given by:

$$\alpha_i = \tilde{\mathbf{x}}_i^\top \mathbf{z},\ \ i = 1, \ldots, n$$

The mean square variation of the data along direction $\mathbf{z}$ is given by

$$\frac{1}{n}\sum_{i=1}^n \alpha_i^2 = \frac{1}{n}\sum_{i=1}^n \mathbf{z}^\top \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^\top \mathbf{z} = \frac{1}{n}\mathbf{z}^\top \tilde{X}^\top \tilde{X}\mathbf{z}$$

The direction $\mathbf{z}$ along which the data has the largest variation can thus be found as the solution to the following optimization problem:

$$\max_{\substack{\mathbf{z}\in\mathbb{R}^m \\ \text{s.t. } ||\mathbf{z}||_2=1}} \quad \mathbf{z}^\top \tilde{X}^\top \tilde{X}\mathbf{z}$$

(a) Show that the direction of maximal variation is the direction of the eigenvector corresponding to the largest eigenvalue of the matrix $\tilde{X}^\top \tilde{X}$.

(b) How is this related to the singular value decomposition of $\tilde{X}$?

(c) Suppose we wish to find the direction with the second-largest variance, and that v_1 is the direction of maximal variation. We can construct the following transformation to the data points:

$$\tilde{\mathbf{x}}_i^{(1)} = \tilde{\mathbf{x}}_i - \mathbf{v}_1(\mathbf{v}_1^\top \tilde{\mathbf{x}}_i) : i = 1, \ldots, n.$$

This transformation simply removes the components along the direction of maximal variation from the data points.
We can then define the matrix

$$\tilde{X}^{(1)} = \begin{bmatrix} \leftarrow \tilde{\mathbf{x}}_1^{(1)\top} \rightarrow \\ \leftarrow \tilde{\mathbf{x}}_2^{(1)\top} \rightarrow \\ \vdots \\ \leftarrow \tilde{\mathbf{x}}_n^{(1)\top} \rightarrow \end{bmatrix} \in \mathbb{R}^{n \times m}$$

The direction of second largest variation can then be found by solving the optimization problem:

$$\max_{\substack{\mathbf{z}\in\mathbb{R}^m \\ \text{s.t. } ||\mathbf{z}||_2=1}} \quad \mathbf{z}\tilde{X}^{(1)\top}\tilde{X}^{(1)}\mathbf{z}$$

Show that the direction with the second largest variation is the direction of the eigenvector corresponding to the second largest eigenvalue of the matrix $\tilde{X}^\top\tilde{X}$ (Solutions that show this by doing the algebra are preffered).

We can repeat the procedure to find more directions of variation. We will use this procedure to find the directions of maximal variation in our dataset of images.

(d) Loading images: In the IPython notebook for this assignment, complete the function loadImage. The function cv2.imread(filename), takes in a file name and returns the image at the file location as a matrix. Complete the function so that the matrix becomes a vector. (Hint: You may use numpy's flatten)

Remember to `conda activate ee127` before starting!

(e) Finding directions of largest variation: We will not make you write the code to solve the optimization problem. Rather, this time we will make use of `sklearn.decomposition.PCA` to obtain these directions. Use the `PCA` object's `fit` method to run SVD, then complete the IPython code to obtain the mean of X and the first 15 directions of largest variation in the dataset using the `PCA` object. What do you expect these directions to look like intuitively?

(f) Visualizing directions of maximal variation: We will now try to see what each of the directions looks like. To do this, we will take an image and vary it along these directions, one at a time, visualizing what it does to the image. An image has been pre-selected for you. Complete the line of code to reshape the image from a vector representation to a matrix. (Hint: You may use numpy's reshape)

(g) Each of the sliders represents a direction of variation with the first slider being the direction with the largest variation, the next being the second largest and so on. Vary the sliders one at a time, while keeping the other sliders fixed, observing what happens to the pre-selected image. What about the image does the direction of maximal variation change? Does this seem reasonable to you, and why? Take two screenshots of you varying the direction of maximal variation (first slider) only. One screenshot with the slider all the way to the left and another screenshot with the slider all the way to the right, with other sliders held fixed. You may also experiment with creating new images by varying the different sliders. Have fun creating new faces!

5. Eigenvectors of a symmetric matrix

Let $\mathbf{p}, \mathbf{q} \in \mathbb{R}^n$ be two linearly independent vectors, with unit norm ($\|\mathbf{p}\|_2 = \|\mathbf{q}\|_2 = 1$). Define the symmetric matrix $A \doteq \mathbf{p}\mathbf{q}^\top + \mathbf{q}\mathbf{p}^\top$. In your derivations, it may be useful to use the notation $c \doteq \mathbf{p}^\top \mathbf{q}$.

(a) Show that $\mathbf{p}+\mathbf{q}$ and $\mathbf{p}-\mathbf{q}$ are eigenvectors of A, and determine the corresponding eigenvalues.

(b) Determine the nullspace and rank of A.

(c) Find an eigenvalue decomposition of A, in terms of $\mathbf{p}, \mathbf{q}$. *Hint:* use the previous two parts.

(d) What is the answer to the previous part if $\mathbf{p}, \mathbf{q}$ are not normalized? Write A as a function $\mathbf{p}, \mathbf{q}$ and their norms and the new eigenvalues as a function of $\mathbf{p}, \mathbf{q}$ and their norms.