Machine Learning *CS 4364/6364, Fall 2017*

HW1

Instructions

Collaboration: Collaboration is allowed, to discuss the problems with other students. However you must write up your own solutions for the math questions, and implement your own solutions for the programming problems. Please list all collaborators and sources consulted, at the top of your homework.

Submission: Please submit homework pdf's via blackboard. Electronic submissions preferred. (Math can be formatted in Latex (some easy editors for Latex are Lyx, TexShop, Texmaker), Word, or other editors). Written homework can be scanned or submitted in class. All code should be submitted electronically. All homeworks are due at 3:45 pm on Tuesday 10/3. 20% penalty for late homeworks. No homeworks accepted after one week after the due date.

Questions

1. **Probability**

   (a) In a box of 20 new iPhones, K of them are defective. The value of K is unknown but it is known to be in the range $\{0, 1, 2\}$, with all three values of K being equally likely. Before shipping the box out of the factory, 4 distinct iPhones are chosen at random from the box and tested. If an iPhone passes the test, then it is not defective. What is the probability that $K = 0$, given that all 4 iPhones tested passed the test? (Provide a formula consisting of all numerical values; it does not have to be simplified). *HINT: Use Bayes' Rule.*

   (b) Every day, Sam goes to the cafeteria and buys either one, two, or three cookies, with equal probability, $\frac{1}{3}$. If he buys three cookies in this first trip to the cafeteria, then he will not buy any more cookies that day. Otherwise (if he buys one or two cookies on his first trip), he will later make a second trip to the cafeteria, where he again buys either one, two, or three cookies, with equal probability, $\frac{1}{3}$.

      (i) What is the probability that Sam buys a total of exactly three cookies on any particular day?

      (ii) Let N be a random variable equal to the total number of cookies Sam buys in any particular day. What is $E[N]$?

      (iii) Let $C = \{N > 3\}$ be the event that the number of cookies Sam bought in one day is strictly greater than 3. What is $E[N|C]$?

2. **Linear Algebra**

   (a) Are the vectors $\boldsymbol{x} = \begin{bmatrix}3\\2\\0\end{bmatrix}, \boldsymbol{y} = \begin{bmatrix}2\\2\\2\end{bmatrix}, \boldsymbol{z} = \begin{bmatrix}0\\2\\1\end{bmatrix}$ linearly independent? Why or why not?

   (b) Let A be a square matrix with a real negative eigenvalue. Show that A is not positive semidefinite.

   (c) Let A be any square matrix. Let p be any polynomial. Show that the eigen values of the matrix $p(A)$ are precisely $p(\lambda_i)$, where λ_i are the eigen values of A.

3. **Principal Components Analysis (PCA) - Programming Exercise**
   You may use your programming language of choice to complete this exercise. Python or MATLAB is recommended. Please document your code with comments.

   In this exercise, we will be using PCA to reduce the dimensionality of the *Wine Data Set* from the UCI Machine Learning Data Repository: `http://archive.ics.uci.edu/ml/datasets/Wine` This data set contains 13 attributes (features) measured from each of 178 types of wine using chemical analysis. The original data set contained labels corresponding to the cultivar which produced the wine, however we will be considering these data in the unsupervised context, so you have been provided (via blackboard) with a comma-separated values (csv) file with the labels stripped out: *wine.data*

   (a) Read *wine.data* into your environment of choice. Ensure that you have 178 data points, each with 13 attributes.

   (b) Calculate the covariance matrix of the data. Perform an eigen-decomposition of the covariance matrix in order to obtain the eigenvectors and their corresponding eigenvalues. You may use a mathematical library function (e.g. numpy.cov and numpy.linalg.eig in python) for these calculations. Provide a plot of the eigenvalues, ordered from largest to smallest, with the y-axis log-scaled.

   (c) Now we want to reduce the dimesionality of the data from 13 to 2. Develop a transformation matrix A that will project a data point x, expressed as a column vector, onto the 2 dimensional eigen-basis that maximizes 2D variance. So Ax will return a 2D column vector. Provide the matrix A in your write-up. *HINT: Which eigenvector corresponds the direction of maximum variance? How would you project a data point onto this eigenvector? Which eigenvector corresponds to the direction of second-largest variance?*

   (d) Use A to project all data points onto this 2D basis.

   (e) *(Grad students only, extra credit for undergrads)* Implement the k-means clustering algorithm (do not use an existing k-means library). Run k-means, with $k = 3$ on the reduced 2D data until you reach reasonable convergence. Report your the location of your final 3 centroids and the final value of the k-means objective in your write-up.

   (f) *(All students)* Now express the reduced 2D data in the original 13D basis (*HINT: Consider the effect that transpose(A) would have on the reduced data*). For each data point calculate the L2 distance from the reconstructed 13D point to the original 13D point. In your write-up, report the average L2 distance for all data points.

In addition to submitting your source code, your write-up should contain:

1. The plot of eigenvalues
2. The matrix A
3. *(Grad students)* Your final k-means centroids and objective value
4. The average L2 distance between the original and reconstructed 13D points