EE 525X / Spring 2018

Homework 3

Please scan and upload your assignments on or before March 2, 2018.

- You are encouraged to discuss ideas and collaborate with each other; but
- you must *clearly* acknowledge your collaborator, and
- you must compose your own writeup and/or code independently, and
- you must combine all derivations, code, and results as a single PDF to be uploaded on BB.
- Maximum score: 40 points

---

1. **(10 points)** Generate a (synthetic) dataset using the same procedure as you did for Problem 6 in Assignment 2. Given this dataset $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, train a logistic regression model that tries to minimize:

$$L(w) = -\sum_{i=1}^{n} y_i \log \frac{1}{1 + e^{-\langle w, x_i \rangle}} + (1 - y_i) \log \frac{e^{-\langle w, x_i \rangle}}{1 + e^{-\langle w, x_i \rangle}}$$

   using two different methods: (i) Gradient descent (GD), and (ii) Stochastic gradient descent (SGD). For each method, plot the decay of the loss function as a function of number of iterations. Demonstrate that SGD exhibits a slower rate of convergence than GD, but is faster per-iteration, and does not suffer in terms of final quality. You may have to play around a bit with the step-sizes to get reasonable answers.

2. **(15 points)** In class, we derived a closed form expression for solving linear regression problems. This is great for finding linear behavior in data; however, if the data is nonlinear, just as in the classification case, we have to resort to the *kernel trick*, i.e., replace all inner products in the data space with kernels. Here, we formalize this idea. Suppose we are given training data $\{(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$ where each response y_i is a scalar, and each data point x_i is a vector in d dimensions.

   a. Assume that all data have been mapped into a higher dimensional space using the feature mapping $x \mapsto \phi(x)$, write down an expression for the squared error function using a linear predictor w in the high-dimensional space.

   b. Let Φ be the matrix with n rows, where row i consists of the feature mapping $\phi(x_i)$. Write down a closed form expression for the optimal linear predictor w as a function of Φ and y.

   c. For a new query data point z, the predicted value is given by $f(z) = \langle w, \phi(z) \rangle$. Plug in the closed form expression for w from the previous sub-problem to get an expression for $f(z)$.

   d. Suppose you are given access to a kernel function K where $K(x, x') = \langle \phi(x), \phi(x') \rangle$. Mathematically show that all the calculations in (b) and (c) can be performed by invoking the kernel function alone *without explicitly calculating ϕ ever*. You may want to use the Sherman-Morrison-Woodbury identity for matrices:

$$(A^{-1} + B^T C^{-1} B)^{-1} B^T C^{-1} = A B^T (B A B^T + C)^{-1}.$$

3. **(15 points)** The *Places Rated Almanac*, written by Boyer and Savageau, rates the livability of several US cities according to nine factors: climate, housing, healthcare, crime, transportation, education, arts, recreation, and economic welfare. The ratings are available in tabular form, available as a supplemental text file. Except for housing and crime, higher ratings indicate better quality of life. Let us use PCA to interpret this data better.

    a. Read the data and construct a table with 9 columns containing the numerical ratings. (Ignore the last 5 columns – they consist auxiliary information such as longitude/latitude, state, etc.)

    b. Replace each value in the matrix by its base-10 logarithm. (This pre-processing is done for convenience since the numerical range of the ratings is large.) You should now have a data matrix X whose rows are 9-dimensional vectors representing the different cities.

    c. Perform PCA on the data. Remember to center the data points first by computing the mean data vector μ and subtracting it from every point. With the centered data matrix, do an SVD and compute the principal components.

    d. Write down the first two principal components v_1 and v_2. Provide a qualitative interpretation of the components. Which among the nine factors do they appear to correlate the most with?

    e. Project the data points onto the first two principal components. (That is, compute the highest 2 scores of each of the data points.) Plot the scores as a 2D scatter plot. Which cities correspond to outliers in this scatter plot?

    f. Repeat Steps 2-5, but with a slightly different data matrix – instead of computing the base-10 logarithm, use the normalized z-score of each data point. (Recall that you used z-scores in Problem Set 2.) How do your answers change?

4. **(Optional)** How much time did you spend on this assignment?