CEE 690: Health and Environmental Data Science, Spring 2019

Homework 3: 100 Total Points

Due: March 3rd by 11:59 PM

Submissions required electronically through Sakai

Problem 1: Does the k-means algorithm converge? (35 points)

There are two steps in the k-means clustering:

1. Update the cluster means
2. Update the cluster assignments for each data point (choosing the cluster with the smallest distance to the mean)

The fit of the k-means algorithm is usually assessed by the sum-of-squared errors (SSE), which for data points $\boldsymbol{x}_i$ for *i*=1,…,*n* with cluster means $\boldsymbol{c}_j$ for *j=1,…k*, is defined as:

$$SSE = \sum_{i=1}^{n} \min_{j \in 1,\ldots,k} \left\| \boldsymbol{x}_i - \boldsymbol{c}_j \right\|_2^2$$

Prove the following:
The first update decreases the SSE.
The second update decreases the SSE.

Given that there is a minimum value for the SSE (it's at least 0), and each step decreases the SSE (or at least does not increase it), what can you conclude about the convergence of the algorithm?

Problem 2: Assessing Performance on Blind Predictions From HW2 (10 points)
In Homework 2, we asked you to make predictions on a blind data set. Here, we want to assess how well you did on these two problems.

For HW 2, problem 4, please assess the mean squared error between the true outcomes in "hw2_problem_4_y_blind.csv" versus your predictions.

(a) What was your hold-out MSE?
(b) A reasonable value is an MSE of about 1.3. If you did *not* achieve a value similar to this, go through the provided solutions and figure out why things did not go as planned. What went wrong?

For HW 2, problem 5, please assess the AUC between the true outcomes in "hw2_problem_4_5_blind.csv" versus your predictions.

(c) What was your hold-out AUC?
(d) A reasonable value from the given approach is about .91 AUC. If you did *not* achieve a value in this range, go through the provided solutions and figure out why things did not go as planned. What went wrong?

Problem 3: Fitting k-Means (10 points)

To get a better understanding of k-means, it is useful to visualize the results on a given dataset. Using two datasets ("iris" and the physiological data I provided), apply k-means with a varying number of clusters.

(a) Plot the MSE of the data fit vs. the number of clusters for each dataset
(b) Pick what you choose as the best number of clusters and visualize the result. Explain why you choose that number of clusters and comment on how it looks visually.

Problem 4: Assessing PCA fit (10 points)

On the physiological data that I provided, run PCA on the data and show how the reconstruction error varies as a function of the number of components.

(a) Visualize the components
(b) Visualize the data after embedding it in the top 2 principal components
(c) How many components do you think is reasonable, and why did you choose that number?

Problem 5: Combining PCA with a classifier (30 points)

In this problem, we will be combining a kernel Support Vector Machine with dimensionality reduction in the form of PCA to show how classification and dimensionality reduction can work together.

I have provided the data for this problem in three files:
"hw3_problem_5_X.csv"
"hw3_problem_5_y.csv"
"hw3_problem_5_X_blind.csv"

Built a pipeline using:
1. PCA (choosing the number of components)
2. SVM (using a Radial Basis Function kernel, choosing *gamma* and *C*)

Try evaluating validation performance using (i) a rule of thumb for choosing the number of components and (ii) choosing the number of components by evaluating the validation performance.

Note that the second strategy requires you to simultaneously determine 3 settings of the algorithms, so this is difficult to visualize. Instead, explain why you choose that setting, and predict how well you think that you will do in terms of AUC on the blind data.

Upload a saved excel file of your predictions on the blind data with your submission files. We will release the blind data's outcomes for the next assignment, and you will get to evaluate how well you actually did predicting future data.

Problem 6: (5 points)
How many hours did it take you to complete this assignment?
Affirm that you adhered to the Duke Honor Code.