CS383 - Machine Learning

Assignment 1 - Dimensionality Reduction & Clustering
Summer 2016

Introduction

In this assignment you'll work on visualizing data, reducing its dimensionality and clustering it.

Your code should be generalizable. That is, it should not just run on the provided dataset but on any dataset of it's form. You'll be given the format of the data in the associated sections.

As a reminder, make sure to clear out old variables prior to running your script.

Grading

Part 1	10pts
Part 2	20pts
Part 3	20pts
Part 4	30pts
Report	20pts
Code doesn't generalize	-5pts

Table 1: Grading Rubric

1 Dimensionality Reduction via PCA

Download the dataset *diabetes.csv* from Blackboard. This dataset has eight features ($D = 8$) and 768 samples ($N = 78$). The first column is the class label {-1, 1}. **However** your script should be able to work on any dataset that lacks a header row and then has data observations, one per row, in the format:

$$(r^i, \chi_1^i, \chi_2^i, ..., \chi_d^i)$$

where $r^i \in \{-1, 1\}$ and d is the number of features.

Write a script that:

1. Reads in the data
2. Standardizes the data (except for the first column of course)
3. Reduces data (except for the first column of course) to 2D using PCA
4. Graphs the data for visualization
   (a) Even though we're not using class labels to do the dimensionality reduction, plot the -1 data in blue and the +1 data in red

2 Dimensionality Reduction via LDA

Now we'll actually use the labels to reduce our data to 1-dimension using LDA.
Using the same data as in Section 1, write a script that:

1. Reads in the data
2. Standardizes the data (except for the first column of course)
3. Reduces data (except for the first column of course) to 1D using LDA
4. Graphs the data for visualization
   (a) Here since we have the class labels, plot the -1 data in blue and the +1 data in red.
   (b) Place the -1 data samples on the y=0 axis
   (c) Place the +1 data samples on the y=1 axis

3 Ranking Features by Information Gain

Next we'll use the label information to rank the features by information gain.

Using the same data as in Section 1, write a script that:

1. Reads in the data
2. Standardizes the data (except for the first column of course)
3. Computes the information gain for each feature.

For your submission you should organize the results into a sorted table of the form:

$$(InformationGain, FeatureNumber)$$

4 Clustering

Next we're going to cluster this same data using k-means!

First here's some details for your implementation to ensure that everyone gets the same results:

Details on k-means implementation

1. Seed the random number generator with zero (do this right before running your k-means)
2. Randomly select two data instances and use them for the initial seeds (since we'll do $k = 2$)
3. Terminate the EM process when the sum of the L1 distance between the prior seeds and the new ones is less than *eps* (which is a MatLab defined variable related to the possible precision).

Now write a script that:

1. Reads in the data
2. Standardizes the data
3. Performs k-means clustering **using just the 6th and 7th feature of the data with k=2**

In addition, in your k-means code you'll want to visualize the progress of the algorithm (this will be part of your report):

1. Plot the initial setup
   (a) Data points are red 'x'
   (b) Cluster centers are blue 'o'
2. Plot the initial cluster assignments
   (a) Cluster 1 = red
   (b) Cluster 2 = blue
   (c) Data points are as 'x' (according the their assigned color)
   (d) Cluster centers are as 'o' (according the their assigned color)
3. Plot the final cluster assignments
   (a) Cluster 1 = red
   (b) Cluster 2 = blue
   (c) Data points are as 'x' (according the their assigned color)
   (d) Cluster centers are as 'o' (according the their assigned color)
   (e) Title should indicate how many iterations it took to get there

Submission

For your submission, upload to Blackboard a single zip file containing:

1. PDF Writeup
2. Source Code
3. readme.txt file

The readme.txt file should contain information on how to run your code to reproduce results for each part of the assignment.

The PDF document should contain the following:

1. Part 1: The visualization of the PCA result
2. Part 2: The visualization of the LDA result
3. Part 3: The Information Gain table sorted according to Information Gain
4. Part 4: The visualization of the k-means clustering process including:
   (a) The initial setup visualization
   (b) The initial cluster assignment visualization
   (c) The final cluster assignment visualization

   and report how many iterations it took for your algorithm to terminate.