CS 109B: Midterm Exam 2

April 6, 2017

Honor Code

The midterm must be completed entirely on your own, and may not be discussed with anybody else.

- You have to write your solutions entirely on your own.
- You cannot share written materials or code with anyone else.
- You may not provide or make available solutions to individuals who take or may take this course in the future.

Your submitted code will be automatically checked for plagiarism. If you are using external resources, make sure to indicate the sources.

The Harvard College Honor Code

Members of the Harvard College community commit themselves to producing academic work of integrity – that is, work that adheres to the scholarly and intellectual standards of accurate attribution of sources, appropriate collection and use of data, and transparent acknowledgement of the contribution of others to their ideas, discoveries, interpretations, and conclusions. Cheating on exams or problem sets, plagiarizing or misrepresenting the ideas or language of someone else as one's own, falsifying data, or any other instance of academic dishonesty violates the standards of our community, as well as the standards of the wider world of learning and affairs.

Introduction

In this exam we're asking you to work with measurements of genetic expression for patients with two related forms of cancer: Acute Lymphoblastic Leukemia (ALL) and Acute Myeloid Leukemia (AML). We ask you to perform two general tasks: (1) Cluster the patients based only on their provided genetic expression measurements and (2) classify samples as either ALL or AML using Support Vector Machines.

In the file `MT2_data.csv`, you are provided a data set containing information about a set of 72 different tissue samples. The data have already been split into training and testing when considering the SVM analyses, as the first column indicates. The first 34 samples will be saved for testing while the remaining 38 will be used for training. Columns 2-4 contain the following general information about the sample:

- ALL.AML: Whether the patient had AML or ALL.
- BM.PB: Whether the sample was taken from bone marrow or from peripheral blood.
- Gender: The gender of the patient the sample was obtained from.

Note that some of the samples have missing information in these columns. Keep this in mind when conducting some of the analyses below. The remaining columns contain expression measurements for 107 genes. You should treat these as the features. The genes have been pre-selected from a set of about 7000 that are known to be relevant to, and hopefully predictive of, these types of cancers.

Problem 1: Clustering [60 points]

For the following, **you should use all 72 samples** – you will only use the genetic information found in columns 5-111 of the dataset. The following questions are about performing cluster analysis of the samples

using only genetic data (not columns 2-4).

(a) (10 points) Standardize the gene expression values, and compute the Euclidean distance between each pair of genes. Apply multi-dimensional scaling to the pair-wise distances, and generate a scatter plot of genes in two dimension. By visual inspection, into how many groups do the genes cluster? If you were to apply principal components analysis to the standardized data and then plot the first two principal components, how do you think the graph would differ? Briefly justify. (you do not need to perform this latter plot)

(b) (10 points) Apply **Partitioning around medoids** (PAM) to the data, selecting the optimal number of clusters based on the Gap, Elbow and Silhouette statistics – if the they disagree, select the largest number of clusters indicated by the statistics. Summarize the results of clustering using a principal components plot, and comment on the quality of clustering using a Silhouette diagnostic plot.

(c) (10 points) Apply **Agglomerative clustering** (AGNES) with Ward's method to the data. Summarize the results using a dendrogram. Determine the optimal number of clusters in a similar way as in (b), and add rectangles to the dendrograms sectioning off clusters. Comment on the ways (if any) the results of PAM differ from those of AGNES.

(d) (10 points) Apply **Fuzzy clustering** (FANNY) to the data, determining the optimal number of clusters as in (b). Summarize the results using both a principal components plot, and a correlation plot of the cluster membership weights. Based on the cluster membership weights, do you think it makes sense to consider summarizing the results using a principal components plot? Briefly justify.

(e) (20 points) For the clusters found in parts (b)-(d), select just one of the clusterings, preferably with the largest number of clusters. For this clustering, what proportion of each cluster are ALL (Acute Lympohblastic Leukemia) samples? In each cluster, what proportion are samples belonging to female subjects? In each cluster, what proportion of the samples were taken from bone marrow as opposed to peripheral blood? What, if anything, does this analysis imply about the clusters you discovered?

Problem 2: Classification [40 points]

For the following problem, we will not be using the general information about the sample due to missing values. Subset the columns keeping only the ALL.AML and the 107 genetic expression values. Then split the samples into two datasets, one for training and one for testing, according to the indicator in the first column. There should be 38 samples for training and 34 for testing.

The following questions essentially create a diagnostic tool for predicting whether a new patient likely has Acute Lymphoblastic Leukemia or Acute Myeloid Leukemia based only on their genetic expression values.

(a) (15 points) Fit two SVM models with linear and RBF kernels to the training set, and report the classification accuracy of the fitted models on the test set. Explain in words how linear and RBF kernels differ as part of the SVM. In tuning your SVMs, consider some values of `cost` in the range of 1e-5 to 1 for the linear kernel and for the RBF kernel, `cost` in the range of 0.5 to 20 and `gamma` between 1e-6 and 1. Explain what you are seeing.

(b) (10 points) Apply principal component analysis (PCA) to the genetic expression values in the training set, and retain the minimal number of PCs that capture at least 90% of the variance in the data. How does the number of PCs identified compare with the total number of gene expression values? Apply to the test data the rotation that resulted in the PCs in the training data, and keep the same set of PCs.

(c) (15 points) Fit a SVM model with linear and RBF kernels to the reduced training set, and report the classification accuracy of the fitted models on the reduced test set. Do not forget to tune the regularization and kernel parameters by cross-validation. How does the test accuracies compare with the previous models from part (a)? What does this convey? *Hint*: You may use similar ranges for tuning as in part (a), but for the RBF kernel you may need to try even larger values of `cost`, i.e. in the range of 0.5 to 40.