**CS 378** **Introduction to Data Mining** **Spring 2009**

# Homework 2

Instructor: Inderjit Dhillon Date Due: Feb 24, 2009

**Keywords:** *Probability, Principal Components Analysis, Exploratory Data Analysis, Netflix Data*

Show work for all problems. Print out all code written in Matlab and attach to your homework solutions.

1. (5 points) **Probability:**

   Suppose that we have three colored boxes $r$ (red), $b$ (blue), and $g$ (green). Box $r$ contains 2 apples, 3 oranges, and 5 limes, box $b$ contains 2 apple, 1 orange, and 1 limes, and box $g$ contains 3 apples, 4 oranges, and 3 limes. If a box is chosen at random with probability $p(r) = 0.3$, $p(b) = 0.5$, $p(g) = 0.2$, and a piece of fruit is removed from the box (with equal probability of selecting any of the items in the box).

   (a) (1 point) What is the probability of selecting an apple?

   (b) (2 points) If we observe that the selected fruit is in fact an orange, what is the probability that it came from the green box?

   (c) (2 points) Construct the complete $3 \times 3$ contingency table (joint probability distribution) for the random variables that correspond to selecting a box and selecting a fruit.

2. (5 points) **Principal Components Analysis:**

   In this problem, you are given the Iris dataset at `http://www.cs.utexas.edu/~wtang/cs378/iris.tar.gz`. The file "iris.data" contains all 150 instances in the Iris dataset. The first four columns of this file correspond to various attributes of the iris plant (sepal length, sepal width, petal length, and petal width). The last column denotes the species type, which is encoded as either 0, 1, or 2 (see the file iris.classes for the mapping from class id to species name). The goal of this problem is visualize the Iris dataset using Pricipal Component Analysis (PCA).

   Complete the following steps in Matlab:

   (a) (2 points) Load the file iris.data into Matlab. (Use the Matlab command load to do this: X=load('iris.data');.) Compute the $4 \times 4$ sample covariance matrix of the dataset. Make sure to omit the species type (the last column in the dataset) from the covariance calculation. Note that the $(i, j)$-th entry of the covariance matrix equals the covariance between the $i$-th and $j$-th attributes.

   (b) (1 point) Compute the eigenvectors and eigenvalues of this sample covariance matrix. (Use the Matlab command eig.)

   (c) (2 points) Project the Iris data onto the plane spanned by the two eigenvectors corresponding to the 2 largest eigenvalues. (To do this, you will need to take the *inner product* between each instance in the data set and each of the two eigenvectors. Each projected instance in the new space will be a vector of dimension two.) Plot this projection using the Matlab command plot. Identify the three classes in the plot.

3. (10 points) **Exploratory Analysis of Netflix Data**

   In this problem, you will analyze a subset of the Netflix Prize data. This subset contains about 2M ratings, which can be downloaded from `http://www.cs.utexas.edu/~wtang/cs378/netflix_subset.tar.gz`. For the detailed information about the data format and usage, please check the "README" file in the package.

   You need to complete the following questions to get familiar with the data set:

(a) (1 point) Plot a histogram showing the distribution of the number of users for each possible rating. For each possible rating score (i.e. 1 through 5), you should count the number of users. (Hint: Use the Matlab command hist to do this.)

(b) (1 point) Compute the average rating across all users and movies.

(c) (2 points) Find the user who rates the largest number of movies. What is the average rating for that user? Find the user who rates the smallest number of movies. What is the average rating for that user? Report the ID of the user and the corresponding average rating in both case.

(d) (3 points) What are the top 10 most highly rated movies according to the number of ratings for each movie? Report the movie name, the number of ratings, its average rating and the standard deviation for each of the top 10 movies.

(e) (3 points) Let's consider pre-processing the data by removing some inherent bias in the data. To demonstrate the effect of removing those bias, the ratings are split into training set and test set. The row/column indices of the ratings in the test set are given in the file "probe.index". We will compute the global effects on the training set and use those effects as predictions for the ratings in the test set. Complete the following pre-processing steps:

**(1)** Compute the global mean $r_g$ of the known ratings in the training set and remove $r_g$ from the training set. Use $r_g$ as the prediction for the ratings in the test set and compute the *Root Mean Squared Error* (RMSE). Report the RMSE value that you get.
The RMSE can be computed as follows:

$$RMSE = \sqrt{\sum_{i=1}^{N} (r_{target}^{(i)} - r_{predict}^{(i)})^2 / N},$$

where $N$ is the total number of ratings, $r_{target}^{(i)}$ is the $i$th actual rating in the test set and $r_{predict}^{(i)}$ is the corresponding $i$th prediction.

**(2)** Based on the resulting training set (after removing $r_g$), compute the user bias $r_u$ (the average of known values) for user $u$ and remove $r_u$ for user $u$ in the training set. Use $r_g + r_u$ as the combined prediction and compute the RMSE again on the test set. Report the RMSE value that you get.

**(3)** Based on the resulting training set (after removing $r_u$ for each user $u$), compute the movie bias $r_m$ (the average of known values) for movie $m$. Use $r_g + r_u + r_m$ as the combined prediction and compute the RMSE again on the test set. Report the RMSE value that you get.