CMPS 1500 Introduction to Computer Science I – Fall 13

11/25/13

Extra Credit Homework

This homework is not mandatory.

Programming portion due **Friday 12/6/13** at 11:55pm on Blackboard.

In order to receive any credit for the programming portions, you are required to thoroughly comment and test your code.

In this homework we are going to take baseball player data and use techniques from supervised machine learning to predict player positions and salaries. Please download the file `hwX.zip`; this includes the following files: `hwX.py, baseball_data.csv, training_set.csv, prediction_data.csv`. Please rename `baseball.py` to `lastName_firstName_hwX.py`. `baseball_data.csv` contains the full baseball player data set, where each element of data is of the form: [AB, H, HR, First Name, Last Name, Salary, Position]. In `training_set.csv` the data for 50 random players has been removed, but otherwise it contains the same information as `baseball_data.csv`. The file `prediction_data.csv` contains the data of those 50 players, but without their positions (those are to be predicted).

1. **Classification (10 points)**
   Consider the data in `training_set.csv` as two-dimensional data using home runs (HR) and batting average (H/AB). The goal is to design a nearest-neighbor classifier that can predict the position for any given player in `prediction_data.csv` for which HR and H/AB is given.

   (a) (0 points) Define a suitable distance measure for this task, based on the Euclidean distance $d((x_1, y_1), (x_2, y_2)) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$. Here, the x-coordinate is HR and the y-coordinate is H/AB, but since these two values have different ranges it might be more appropriate to normalize the contribution of each coordinate. There is no right or wrong answer here, the point is simply to use a suitable distance measure.

   (b) (6 points) Design a nearest-neighbor classifier to predict the position for a player with given home runs (HR) and batting average (H/AB). First, test your code with 30 home runs and with a batting average of 0.300. Then for the 50 random players in `prediction_data`, predict their position and check for overall accuracy of your prediction.

   (c) (4 points) Develop a more advanced prediction by building a k-nearest neighbor classifier for some value $k > 1$; here you can have multiple nearest neighbors "vote" for the player's position for example. Again, test your code with 30 home runs and with a batting average of 0.300, and then for the 50 random players in `prediction_data`, predict their position and check for overall accuracy of your prediction.

FLIP OVER TO BACK PAGE ⟹

2. **Regression (10 points)**
   In this problem we will use the data in `baseball_data.csv` to compute a linear regression to predict the salary of the players. (We won't use the other two data sets.)

   (a) (6 points) First, we will use simple linear regression with a single independent variable. This time, we will consider the data in `baseball_data.csv` as two-dimensional data using home runs (HR) and salary. Apply the formula we covered in class to compute a simple linear regression line for home runs $x = HR$ and salary y, in order to predict salary (y) from home runs (x). Answer the following three questions:

      i. If a player had 25 home runs, what would you expect his salary to be?
      ii. Which player, based on home runs, is expected to have had the largest salary?
      iii. Which player was the most underpaid player?

   (b) (4 points) Now, the task is to develop a more advanced prediction that uses a more advanced linear regression with two independent variables ($x_1 = HR$ and $x_2 = (H/AB)$ to predict salary y. Look up the formula for this case on the internet and implement it. Answer the following three questions:

      i. What is the expected salary for a player with 30 home runs and a batting average of .300?
      ii. What is the expected salary for a player with 5 home runs and a batting average of .200?
      iii. Find the ten most underpaid players.