CS 188 HW 2

Due on Friday, February 14 at 11:59PM

1 Instructions:

You may form small groups (e.g. of up to four people) to work on this assignment, but you must write up all solutions by yourself. List your study partners for the homework on the first page, or "none" if you had no partners.

Keep all responses brief, a few sentences at most. Show all work for full credit.

Start each problem on a new page, and be sure to clearly label where each problem and subproblem begins. All problems must be submitted in order (all of P1 before P2, etc.).

No late homeworks will be accepted. This is not out of a desire to be harsh, but rather out of fairness to all students in this large course.

2 Perceptron Training

Assume a three input perceptron plus bias (it outputs 1 if $b + \sum_i w_i * x_i > 0$, else 0). Assume a learning rate c of 1 and initial weights all 1: $\Delta w_i = c(t - z) * x_i$, where t is the true label and z is the predicted label.

Show weights after each pattern in Table 1 until the result converges. Use an Excel sheet (attach your Excel sheet to the homework). Iterate over the training samples from top to bottom.

x_1	x_2	x_3	t
1	0	1	0
1	1	0	0
1	0	1	1
0	1	1	1

Table 1: Train Set

3 Input Validation

A SickBit health sensor produces a stream of readings from 20 different sensors (think blood pressure, heart rate body temperature, etc.). List two techniques you could use to check whether the stream of data coming from the sensors are valid or not. Write one or two sentences to describe each approach.

4 Distributions

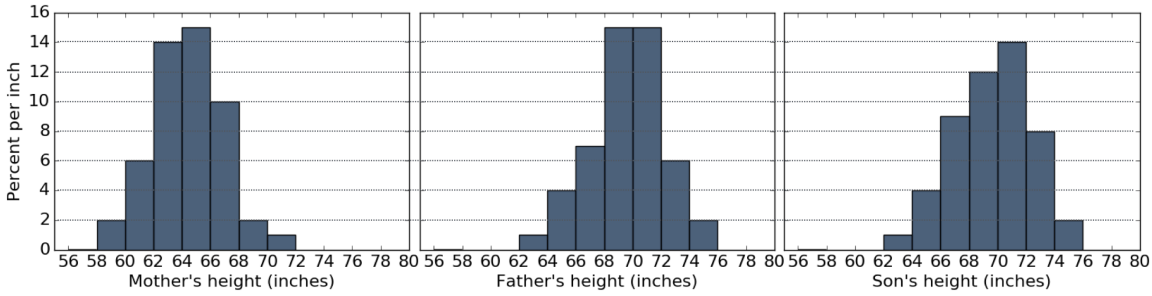

Figure 1: Height Distributions

Galton measured the heights of individuals in 200 families, each of which included one mother, one father, and a varying number of adult sons. The three histograms of heights in Figure 1 depict the distributions for all mothers, fathers, and adult sons. All bars are 2 inches wide. All bar heights are integers. The heights of all people in the data set are included in the histograms.

(a) Calculate each quantity described below or write Unknown if there is not enough information above to express the quantity as a single number (not a range). Show your work!

(i) The percentage of mothers that are at least 60 inches but less than 64 inches tall.

(ii) The percentage of fathers that are at least 64 but less than 67 inches tall.

(iii) The number of sons that are at least 70 inches tall.

(iv) The number of mothers that are at least 60 inches tall.

(b) If the father's histogram were redrawn, replacing the two bins from 72-to-74 and from 74-to-76 with one bin from 72-to-76, what would be the height of its bar? If it's impossible to tell, write Unknown.

(c) The percentage of sons that are taller than all of the mothers is between _____ and _____. Fill in the blanks in the previous sentence with the smallest range that can be determined from the histograms, then explain your answer below.

5 Voronoi

Draw the Voronoi diagram of 10 points all on a line. Draw separately the Voronoi diagram of 10 points all on a circle. What do these two diagrams have in common?

6 Augmentation

Many methods for making predictions from data, such as linear regression, are limited in terms of the transformations that they can apply to input data before making a prediction. As linear regression assumes that the output is the sum of coefficients multiplied by input features, it is unable to account for cases where the impact of two features together is greater than the sum of their parts. For example, a house that both has > 5 bedrooms and is in California may be worth four times more than would be expected from the learned price impact of each feature on its own.

Feature Crosses are synthetic features you can form by crossing two or more features together, and they can help to improve the predictive power of techniques such as linear regression. Expanding on the above housing example, you could generate a new feature that indicates a combination of both a home's number of bedrooms and location.

(a) Describe two pairs of features from Project 2 that might be interesting to cross together, and explain why.

(b) You have latitude and longitude for homes, and you think feature crosses may allow you to make better predictions. However, your latitude and longitude are continuously valued. How might you do a feature cross in this case?

(c) Think up a dataset consisting of features X and Y and associated labels Z that is shaped such that a linear model would perform poorly without feature crosses. Provide a table with at least 7 points from your dataset.