6.891 Fall 1999: Problem Set #1

Due: Thursday, 9/16/99

Pre-questions

Howdy. This part of the problem set is intended as a diagnostic test of your knowledge of probability and linear algebra – the background material necessary for Machine Learing. Unfortunately we will not be able to review much of this material in 6.891, so this problem set is designed to let you know what sorts of things you should find easy. If you find that this material is really hard, *do not panic*. You have a week or two to catch up. Also, Kinh will be teaching a review session on Monday afternoon (stay tuned to schedule this).

Problem 1: Probability, Densities and Bayes' Rule

Assume that x is distributed according to the Gaussian density with mean $\mu = 0$ and variance $\sigma^2 = 1$.

- (A) What is the probability that $x = 0$?

Assume that y is distributed according to the Gaussian density with mean $\mu = 1$ and variance $\sigma^2 = 1$.

- (B) What is the probability that $y = 0$?

Given a distribution: $Pr(Z = z) = 1/2Pr(X = z) + 1/2Pr(Y = z)$ known as a mixture (i.e. 1/2 of the time points are generated by the X process and 1/2 of the time by the Y process).

- (C) If $Z = 0$, what is the probability that the X process generated this data point?

Problem 2; Linear Algebra

A line can be written as $ax + by = 1$. You are given a number of example points:

$$P = \begin{bmatrix} x_1 & y_1 \\ \vdots & \vdots \\ x_n & y_n \end{bmatrix} \tag{1}$$

. Let:

$$M = \begin{bmatrix} a \\ b \end{bmatrix} \tag{2}$$

- (A) Write a single matrix equation that expresses the constraint that each of these points lies on a single line.

- (B) Is it always the case that some M exists?

- (C) Write an expression for M assuming it does exist.

Some Slightly Harder Problems

Problem 3: High Dimensional Weirdness

Consider the line passing through the origin whose minimum angle to any axis is maximal. That is, this line maximizes the angle between it and all axes. In two dimensions, two such lines exist, both at $45°$ from each axes. Let α be the angle between this line and any axis (notice that the angles between this line and all axes will be equal).

- (A) Find $\alpha(d)$ in closed form, where d is the dimensionality of the space. What is $\alpha(d)$ for $d = 2, 3, 256, \infty$?

Problem 4: Apples and Worms

The task of finding and removing apples that house worms before they get to the grocery store is a big problem (well, imagine that it is for the remainder of this question). Consumers have been shown to react poorly upon finding worms in their apples. To combat this problem, Wormfinder Inc. has developed an amazing new non-intrusive test for worms in apples. This test, called WormScan has the incredible false negative rate of exactly 0 (i.e. if an apple is declared by WormScan to be free of worms, it is guaranteed to have no worms in it). Unfortunately, such a performance comes with a cost; the false positive rate is 3% (i.e. 3% of all good apples are marked as having a worm inside). Statistically, it has been found that 0.2% (1 in 500) apples have worms.

- (A) What percentage of the apples will test as having worms?

- (B) Given that an apple has tested as having worms, what is the probability that there is a worm inside?

Two other companies, Worm Busters Ltd. and FFT (Fast Fruit Testers), have each released a product to test for worms in apples. Both of these companies' tests perform with 0% false negative and 3% false positive rates just as the Wormfinder Inc's technology. Sun Grocery Stores would like to use all three tests together to boost their ability to detect worms in apples.

- (C) Is this necessarily possible? Why?

- (D) If the three tests are statistically independent given the status of the apple, what will be the false positive rate of the boosted combination?

- (E) After combining the three, what percentage of the apples that test as being bad have worms in them?

Unfortunately, Sun Groceries has found that these three products are far too expensive. Instead, they have set many teams of engineers to design cheap worm testers. The problem is that each of these testers is rather poor. They correctly identify whether there is a worm inside only 73% of the time (this rate is for positives and negatives). However, Sun is fairly certain that the tests are independent since each of the teams came up with a different design. They have decided that a false negative rate of 0.01% and a false positive rate of 1% are acceptable.

- (F) How should they boost these tests and how many will they need to achieve the desired performance? [Hint: Do *not* try to solve this analytically; write a program to solve the necessary equations.]