CSE 151: Introduction to Machine Learning Spring 2018

Problem Set 6

Instructor: Kamalika Chaudhuri **Due on:**

Problem 1: 6 points

In class, we saw that if we have two labels, then the error of a classifier which guesses completely randomly is 0.5. In this problem, we look at what happens when there are $k > 2$ labels.

1. Random guesser Geser knows that there are k labels, and for each example, selects a label out of $\{1, \ldots, k\}$ uniformly at random. What is the error of Geser ?

2. Now suppose we have a more sophisticated random guesser Zavulon who knows that w_1 fraction of the data distribution has label 1, w_2 fraction has label 2, and so on. For each example, Zavulon also selects a label out of $\{1, \ldots, k\}$ at random, but he selects label 1 with probability w_1, label 2 with probability w_2 and so on. What is the error of Zavulon?

Problem 2: 14 points

Consider the following two data distributions D_1 and D_2 over labeled examples. There is a single feature, denoted by X which takes values in the set $\{1, 2, 3, 4\}$ and a binary label $Y \in \{0, 1\}$. D_1 is described as follows:

$$\begin{aligned} \Pr(X = i) &= \frac{1}{4}, \;\; i \in \{1, 2, 3, 4\} \\ \Pr(Y = 1 | X = i) &= 1, \;\; i \in \{1, 4\} \\ \Pr(Y = 0 | X = i) &= 1, \;\; i \in \{2, 3\} \end{aligned}$$

D_2 is described as follows.

$$\begin{aligned} \Pr(X = i) &= \frac{1}{4}, \;\; i \in \{1, 2, 3, 4\} \\ \Pr(Y = 1 | X = i) &= \frac{i}{10}, \;\; i \in \{1, 2, 3, 4\} \end{aligned}$$

1. Consider the following classifier h: $h(x) = 1$ if $x > 1.5$ and 0 otherwise. What is the true error of h when the true data distribution is D_1?

2. Suppose our concept class C is the set of all classifiers of the form: $h_t(x) = 1$ if $x > t$ and 0 otherwise. Write down a classifier in this concept class that minimizes the true error when the data distribution is D_1. What is the true error of this classifier? Do we have a non-zero bias when the concept class is C and the data distribution is D_1?

3. Repeat parts (1) and (2) for the data distribution D_2.