**CSE 151: Introduction to Machine Learning** Spring 2018

# Problem Set 2

*Instructor: Kamalika Chaudhuri* **Due on:**

## Problem 1 (14 points)

In class, we mentioned that for some applications, it may make sense to do $k$-nearest neighbors with respect to a distance other than the usual Euclidean distance. In this problem, we will look at the $k$-nearest neighbor problem when the distance between the points is the following modified form of the Euclidean distance. Given two vectors $x = (x_1, x_2)$ and $z = (z_1, z_2)$, the modified distance measure $d_M(x, z)$ is defined as:

$$d_M(x, z) = \sqrt{\frac{1}{2}(x_1 - z_1)^2 + (x_2 - z_2)^2}$$

1. Consider the following labelled training dataset:

   $$((0, 0), 1), ((2, 2), 2), ((4, 0), 3)$$

   First, consider the 1-nearest neighbor classifier on these points with respect to the usual Euclidean distance, and draw the decision boundary for this classifier. Write down the equations for the different sections (or segments) of the decision boundary. Clearly mark each region in your drawing with the label assigned by the classifier to a test example in this region.

2. Now, consider the 1-nearest neighbor classifier on labelled datasets in part (1) with respect to the modified Euclidean distance $d_M$. In a separate figure, draw the decision boundary for this classifier. Again, write down the equations for the different segments of the decision boundary, and clearly mark each region in your drawing with the label assigned by the classifier to a test example in this region.

3. Repeat parts (1) and (2) (namely, drawing the decision boundary with respect to the Euclidean distance and the modified Euclidean distance $d_M$) for the following labelled training dataset:

   $$((0, 0), 1), ((1, 1), 1), ((-1, 1), 2)$$

## Problem 2 (6 points)

In class, we talked about how $k$-nearest neighbor classifiers are robust to errors or noise in the data when $k > 1$. In this problem, we will look more closely at how exactly this error correction happens.

Suppose that we have two labels 0 and 1. Suppose we are given any $k$, and any test point $x$; let $z_1, \ldots, z_k$ be the $k$ closest neighbors of $x$ in the training data. For the rest of the question, we make the assumption that for all $i = 1, \ldots, k$, the probability that the label of $z_i$ is not equal to the label of $x$ is $p = 0.2$; moreover, for $i \neq j$, the events that the label of $z_i$ is not equal to the label of $x$ and the label of $z_j$ is not equal to the label of $x$ are independent. In reality of course, this assumption will not hold for very large $k$, but for small $k$, this is a fairly reasonable assumption to make.

1. What is the probability that the 1-nearest neighbor classifier makes a mistake on $x$?

2. Now calculate the probability that 3-nearest neighbor classifier and the 5-nearest neighbor classifier make a mistake on $x$. What can you conclude from these calculations about the robustness of these classifiers?