CSE 250B: Homework 1 Solutions

1. *Risk of a random classifier.*

   (a) No matter what the correct label is, the probability that a random classifier selects it is 0.25. Therefore, this classifier has risk (error probability) 0.75.

   (b) We should return the label with the highest probability, which is A. The risk of this classifier is the probability that the label is something else, namely 0.5.

2. *Discrete and continuous distributions.*

   (a) Another example of a discrete distribution with infinite support is the *geometric distribution*. The simplest case of this has possible outcomes $0, 1, 2, \ldots$, where the probability of outcome i is $1/2^{i+1}$.

   (b) If X follows a uniform distribution over $[a, b]$ (where $a < b$), the probability that X takes on any specific value is 0.

3. *Complexity analysis for k-d tree with defeatist search.*

   (a) Let's assume that we are given the d-dimensional data points in the form of an $n \times d$ matrix. We will construct a tree data structure whose leaves each contain at most n_o of these data points (more precisely, a list of the row indices corresponding to the points).

   At each internal node of the tree, containing (say) m points:

   - The time taken to choose a coordinate for splitting is $O(md)$, if we pick the coordinate with highest variance.
   - We can use a linear-time median-finding algorithm to find the split point.
   - We partition the points into left and right groups, also in linear time.

   Therefore, the total time taken for this node is $O(md)$ and thus the time for constructing an entire level of the tree is $O(nd)$.

   There are n points in the data set and with each successive level, the number of points per cell is halved, until we reach leaf nodes with $\leq n_o$ points. So the height of the tree is $\log(n/n_o)$.

   Therefore, the total complexity of building a k-d tree as specified in the problem is $O(nd \log(n/n_o))$.

   (b) To answer a query, we first move to the appropriate leaf of the tree, which takes time $O(\log(n/n_o))$ (constant time per internal node along the way), and we then look for the nearest neighbor within that leaf, which takes time $O(n_o d)$. The total query time is thus $O(n_o d + \log(n/n_o))$.

4. *Properties of metrics.* Recall that d is a distance metric if and only if it satisfies the following properties:

   (P1) $d(x, y) \geq 0$

   (P2) $d(x, y) = 0 \iff x = y$

   (P3) $d(x, y) = d(y, x)$ (symmetry)

   (P4) $d(x, z) \leq d(x, y) + d(y, z)$ (triangle inequality)

   (a) If d_1 and d_2 are metrics, then so is $g(x, y) = d_1(x, y) + d_2(x, y)$. All four properties can be verified directly.

   (P1) $g(x, y) \geq 0$ because it is the sum of two nonnegative values.

   (P2) Pick any x, y.

$$
\begin{aligned}
g(x, y) = 0 &\iff d_1(x, y) + d_2(x, y) = 0 \\
&\iff d_1(x, y) = 0 \text{ and } d_2(x, y) = 0 \text{ (since both nonnegative)} \\
&\iff x = y
\end{aligned}
$$

(P3) $g(x,y) = d_1(x,y) + d_2(x,y) = d_1(y,x) + d_2(y,x) = g(y,x)$.

(P4) For any x, y, z,

$$\begin{aligned} g(x,z) &= d_1(x,z) + d_2(x,z) \\ &\leq (d_1(x,y) + d_1(y,z)) + (d_2(x,y) + d_2(y,z)) \\ &= (d_1(x,y) + d_2(x,y)) + (d_1(y,z) + d_2(y,z)) \\ &= g(x,y) + g(y,z) \end{aligned}$$

(b) Hamming distance is a metric.

(P1) $d(x,y) \geq 0$ because number of positions at which two strings differ can't be negative.

(P2) $d(x,x) = 0$ because a string differs from itself at no positions. Also, if $x \neq y$, there will be at least one position where x and y differ and hence $d(x,y) \geq 0$.

(P3) $d(x,y) = d(y,x)$ because x differs from y at exactly the same positions where y differs from x.

(P4) Pick any $x, y, z \in \Sigma^m$. Let A denote the positions at which x, y differ: $A = \{i : x_i \neq y_i\}$, so that $d(x,y) = |A|$. Likewise, let B be the positions at which y, z differ and let C be the positions at which x, z differ.
Now, if $x_i = y_i$ and $y_i = z_i$, then $x_i = z_i$. Thus $C \subseteq A \cup B$, whereupon
$d(x,z) = |C| \leq |A| + |B| = d(x,y) + d(y,z)$.

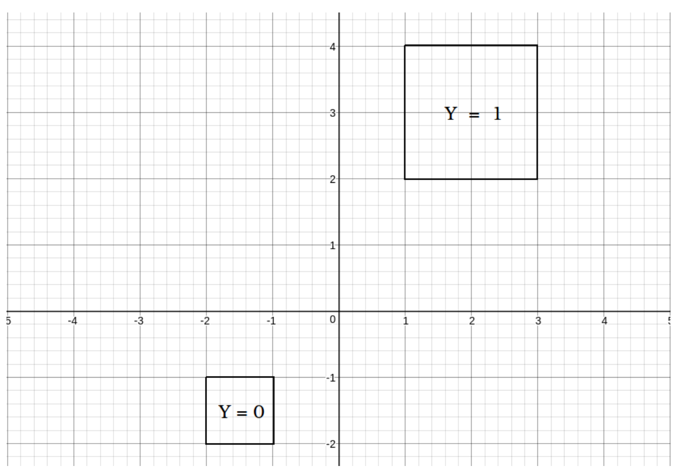
(c) Squared Euclidean distance is not a metric as it does not satisfy the triangle inequality. Consider the following three points in $\mathbb{R}$: $x = 1, y = 4, z = 5$.

$$\begin{aligned} d(x,z) &= (1-5)^2 = 16 \\ d(x,y) &= (1-4)^2 = 9 \\ d(y,z) &= (4-5)^2 = 11 \end{aligned}$$

Here $d(x,z) > d(x,y) + d(y,z)$.

5. *A joint distribution over data and labels.*

(a) Graph with regions where (x_1, x_2) might fall.



(b) Let μ_1 denote the density function of X_1.

$$\mu_1(x_1) = \begin{cases} 1/2 & \text{if } -2 \leq x_1 \leq -1 \\ 1/4 & \text{if } 1 \leq x_1 \leq 3 \\ 0 & \text{elsewhere} \end{cases}$$

(c) Let μ_2 denote the density function of X_2.

$$\mu_2(x_2) = \begin{cases} 1/2 & \text{if } -2 \leq x_2 \leq -1 \\ 1/4 & \text{if } 2 \leq x_2 \leq 4 \\ 0 & \text{elsewhere} \end{cases}$$

6. *Two ways of specifying a joint distribution over data and labels.*

The marginal distribution of $x = (x_1, x_2)$ is given by the following density function:

$$\mu(x_1, x_2) = \begin{cases} 1/8 & \text{if } -1 \leq x_1 < 0 \\ 3/8 & \text{if } 0 \leq x_1 < 1 \\ 1/4 & \text{if } 1 \leq x_1 \leq 3 \end{cases}$$

The conditional distribution of y given $x = (x_1, x_2)$ is

$$\eta(x) \;=\; \Pr(Y = 1 | X = (x_1, x_2)) \;=\; \begin{cases} 1 & \text{if } -1 \leq x_1 < 0 \\ 1/3 & \text{if } 0 \leq x_1 < 1 \\ 0 & \text{if } 1 \leq x_1 \leq 3 \end{cases}$$

7. *Bayes optimality.*

(a) The Bayes-optimal classifier predicts 1 when $-0.5 \leq x \leq 0.5$, and 0 elsewhere. Its risk (probability of being wrong) is:

$$R^* \;=\; \int_{-1}^{1} \min(\eta(x), 1 - \eta(x))\, \mu(x)\, dx \;=\; \int_{-1}^{0.5} 0.2|x|\, dx \;+\; \int_{0.5}^{1} 0.4|x|\, dx \;=\; 0.275.$$

(b) The 1-NN classifier based on the four given points predicts as follows:

$$h(x) \;=\; \begin{cases} 1 & \text{if } -0.6 \leq x \leq 0.5 \\ 0 & \text{if } x < -0.6 \text{ or } x > 0.5 \end{cases}$$

Notice that this differs slightly from the Bayes optimal classifier. The risk of rule h is

$$\begin{aligned} R(h) &= \int_{-1}^{1} \Pr(y \neq h(x) \mid x)\, \mu(x)\, dx \\ &= \int_{-1}^{-0.6} 0.2|x|\, dx + \int_{-0.6}^{-0.5} 0.8|x|\, dx + \int_{-0.5}^{0.5} 0.2|x|\, dx + \int_{0.5}^{1} 0.4|x|\, dx \;=\; 0.308. \end{aligned}$$

(c) The cost of predicting 1 when the true label is 0 is ten times the cost of predicting 0 when the true label is 1. The best thing to do is to simply predict 0 everywhere.

(d) The classifier with smallest cost-sensitive risk is:

$$h^*(x) = \begin{cases} 1 & \text{if } c_{01}(1 - \eta(x)) \leq c_{10}\eta(x) \\ 0 & \text{if } c_{01}(1 - \eta(x)) > c_{10}\eta(x) \end{cases}$$