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. *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
   $d(x, z) = |C| \leq |A| + |B| = d(x, y) + d(y, z)$.

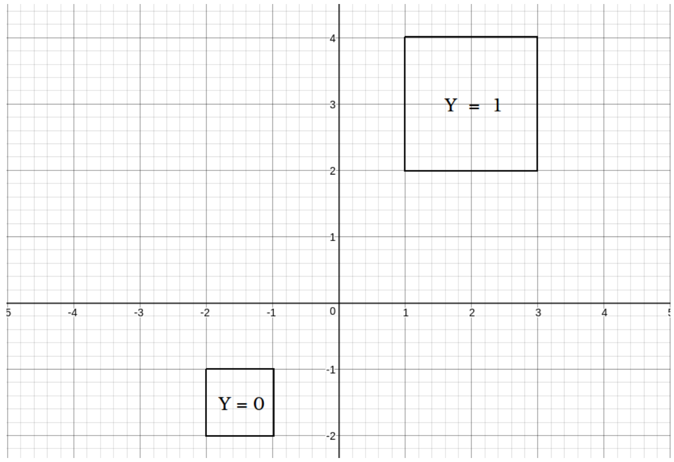
   (c) Squared Euclidean distance is not a metric as it does not satisfy the triangle inequality. Consider the following three points in $\mathbb{R}$: $x = 1, y = 4, z = 5$.

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

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

3. *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.

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

4. *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

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

5. *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:

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

6. *Error rate of 1-NN classifier.*

(a) Consider a training set in which the same point x appears twice, but with different labels. The training error of 1-NN on this data will not be zero.

(b) We mentioned in class that the risk of the 1-NN classifier, $R(h_n)$, approaches $2R^*(1 - R^*)$ as $n \to \infty$ where R^* is the Bayes risk. If $R^* = 0$, this means that the 1-NN classifier is consistent: $R(h_n) \to 0$.

7. *Bayes optimality in a multi-class setting.* The Bayes-optimal classifier predicts the label that is most likely:

$$h^*(x) = \arg\max_{i \in |\mathcal{Y}|} \eta_i(x)$$

8. *The statistical learning assumption.*

(a) Here, μ is the distribution over proposed songs, while η tells us which songs will be successful. Both are likely to change with time, violating the statistical learning assumption. However, the drift might be quite slow, so a classifier trained today may work well for another year or two before needing to be re-trained.

(b) In this example, the bank's data set consists only of loans it *accepted*. It is not a random sample from μ, which is the distribution over all loan applications. This is a severe violation of the i.i.d. sampling requirement.

(c) The move from the west coast to the entire country means that μ is changing, and it is possible that η is changing as well. Technically, this violates the statistical learning assumption; but it is possible that the change in distribution may not be very severe.