1. **Risk of a random classifier.**

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

   (b) If we pick a classifier that always returns label $i$, then it is correct when the input’s label is $i$, and incorrect otherwise. So the most accurate classifier of this type should return the label for which the inputs have the highest frequency, 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, ..., 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. **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 $g(x, y) = d_1(x, y) + d_2(x, y)$. To show this, we will now verify the four properties directly.

   (P1) $g(x, y) \geq 0$ because it is the sum of two nonnegative values.
   (P2) Pick any $x, y$.

   $g(x, y) = 0 \iff d_1(x, y) + d_2(x, y) = 0$
   $\iff d_1(x, y) = 0$ and $d_2(x, y) = 0$ (since both nonnegative)
   $\iff x = y$

   (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 $x, y, z$,

   $g(x, z) = d_1(x, z) + d_2(x, z)$
   $\leq (d_1(x, y) + d_1(y, z)) + (d_2(x, y) + d_2(y, z))$
   $= (d_1(x, y) + d_2(x, y)) + (d_1(y, z) + d_2(y, z))$
   $= g(x, y) + g(y, z)$

   (b) Hamming distance is a metric. We show why below by proving all four properties.

   (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 $x$. 

1
(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 $\mathbb{R}$: $x = 1, y = 4, z = 5$.

\[
\begin{align*}
    d(x, z) &= (1 - 5)^2 = 16 \\
    d(x, y) &= (1 - 4)^2 = 9 \\
    d(y, z) &= (4 - 5)^2 = 11
\end{align*}
\]

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

4. A joint distribution over data and labels.

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

(b) Let $\mu_1$ and $\mu_2$ denote the density function of $X_1$ and $X_2$ respectively, and let $\mu$ denote the joint density of $(X_1, X_2)$. Then,

\[
\mu(X_1, X_2) = \Pr(Y = 0)\mu(X_1, X_2|Y = 0) + \Pr(Y = 1)\mu(X_1, X_2|Y = 1)
\]

where $\mu(X_1, X_2|Y = i)$ is the conditional density of $(X_1, X_2)$ given that the label is $i$. For $Y = 0$, this conditional density is uniform on the square $[-2, -1] \times [-2, -1]$ and for $Y = 1$, this is uniform on $[1, 3] \times [2, 4]$. Additonally, $\Pr(Y = 0) = \Pr(Y = 1) = 1/2$. Plugging this in, we get:

\[
\mu(x_1, x_2) = \begin{cases} 
1/2 & (x_1, x_2) \in [-2, -1] \times [-2, -1] \\
(1/2) \cdot (1/4) = 1/8 & (x_1, x_2) \in [1, 3] \times [2, 4] \\
0 & \text{otherwise}
\end{cases}
\]

Observe that this density integrates to 1. Now we can calculate $\mu_1(X_1)$ as:

\[
\mu_1(x_1) = \int_{x_2 = -\infty}^{\infty} \mu(x_1, x_2) dx_2
\]

$\mu_2$ can be calculated similarly. The answers are given below.
5. Two ways of specifying a joint distribution over data and labels. We can calculate the marginal distribution \( \mu \) over \( X \) using the following relationship:

\[
\mu(X_1, X_2) = \Pr(Y = 0)\mu(X_1, X_2|Y = 0) + \Pr(Y = 1)\mu(X_1, X_2|Y = 1)
\]

Here, \( \Pr(Y = 1) = \frac{1}{4} \) and \( \Pr(Y = 0) = \frac{3}{4} \), \( \mu(X_1, X_2|Y = 0) = 1/3 \) over \([0, 3] \times [0, 1]\) and \( \mu(X_1, X_2|Y = 1) = 1/2 \) over \([-1, 1] \times [0, 1]\). Putting these all together, we can calculate the marginal distribution of \( x = (x_1, x_2) \) as follows:

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

To calculate the conditional distribution of \( y \) given \( x \), we first calculate the joint distribution of \( (x = (x_1, x_2), y) \). This is calculated as \( \mu(X_1, X_2, Y) = \Pr(Y = y)\mu(X_1, X_2|Y = y) \) for \( y = 0, 1 \). Plugging in this expression, we get:

\[
\mu(x_1, x_2, y) = \begin{cases} 
1/4 & (x_1, x_2) \in [0, 3] \times [0, 1], y = 0 \\
1/8 & (x_1, x_2) \in [-1, 1] \times [0, 1], y = 1 \\
0 & \text{otherwise} \\
\end{cases}
\]

We can now calculate the conditional distribution of \( x \) given \( y = 1 \) as \( \eta(x) = \mu(x_1, x_2, 1)/\mu(x_1, x_2) \).

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

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


(a) Recall that for a specific \( x \), the Bayes-optimal classifier predicts the label \( y \) that maximizes \( \Pr(Y = y|X = x) \) – since this will have the highest accuracy. Here, 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

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

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

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

Observe that for the distribution described in the beginning of the problem, for all $x$, $c_{01}(1 - \eta(x)) > c_{10}\eta(x)$ – hence it is always best to predict 0.

7. 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. So 1-NN will have non-zero training error for any three points where two points have this property.

(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: $R(h_n) \to 0$.

8. Bayes optimality in a multi-class setting. Recall that the Bayes optimal classifier is the one that maximizes accuracy for each $x$. If we predict label $i$ for $x$, then the accuracy is $\eta_i(x)$; this suggests that the Bayes optimal should predict the label that maximizes $\eta_i(x)$. Specifically, it predicts the label that is most likely:

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

9. Classification with an abstain option. For a given $x$, the expected cost incurred by a classifier is $\theta$ if it abstains, $\eta(x)$ if it predicts 0 and $1 - \eta(x)$ if it predicts 1. The optimal cost classifier should choose the option which has the minimum cost of the three options (predicting 0, 1 and abstaining) for each $x$; this happens when it abstains whenever the probability of error exceeds $\theta$. Putting things together, this classifier turns out to be:

$$
h^*(x) = \begin{cases} \text{abstain} & \text{if } \theta < \eta(x) < 1 - \theta \\ 1 & \text{if } \eta(x) \geq 1 - \theta \\ 0 & \text{if } \eta(x) \leq \theta \end{cases}
$$

10. The statistical learning assumption.

(a) Here, $\mu$ is the distribution over proposed songs, while $\eta$ 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 $\mu$, 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 $\mu$ is changing, and it is possible that $\eta$ 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.
11. (a) $C_1$ is generally not equal to $C_2$ in this case. For a brief counterexample, suppose $S$ has two points 
$(3,3)$ with label 0 and $(5,0)$ with label 1. Pick a test point $x = (0,0)$. 
$C_1(x) = 1$ as $\|x - (3,3)\|_1 = 6 > \|x - (5,0)\|_1 = 5$. But $C_2(x) = 0$ as 
$\|x - (3,3)\|_2 = 3\sqrt{2} = 4.2 < \|x - (5,0)\|_2 = 5$.

(b) Here, $C_1$ is equal to $C_2$. Pick any point $x$; the closest neighbor of $x$ within the training set $S$ in $L_2$-distance is going to be the closest neighbor of $x$ within $S$ in the square of the $L_2$-distance. This means that both $C_1$ and $C_2$ will output the same label for $x$. Since this holds for any test point $x$, $C_1$ and $C_2$ are equal.