1. To classify a point $x$, we evaluate the three linear functions and pick the one with the highest value. The region where class 1 beats class 2 is:

$$w_1 \cdot x + b_1 > w_2 \cdot x + b_2 \iff (w_1 - w_2) \cdot x + (b_1 - b_2) > 0 \iff x_2 > 1$$

The region where class 1 beats class 3 is:

$$w_1 \cdot x + b_1 > w_3 \cdot x + b_3 \iff (w_1 - w_3) \cdot x + (b_1 - b_3) > 0 \iff x_1 > -1$$

The region where class 2 beats class 3 is:

$$w_2 \cdot x + b_2 > w_3 \cdot x + b_3 \iff (w_2 - w_3) \cdot x + (b_2 - b_3) > 0 \iff x_1 - x_2 > -2$$

So class 1 is predicted in the intersection of the first two regions, etc. This is summarized in the figure below.

2. Pointwise product of positive semidefinite matrices.
   
   (a) Because $X$ and $Y$ are independent, $E(Z) = E(XY) = E(X)E(Y) = 0$.

   $$\text{Cov}(Z_i, Z_j) = \text{Cov}(X_iY_i, X_jY_j)$$
   $$= E(X_iX_jY_iY_j) - E(X_iY_i)E(X_jY_j)$$
   $$= E(X_iX_j)E(Y_iY_j) - E(X_i)E(Y_i)E(X_j)E(Y_j)$$
   $$= E(X_iX_j)E(Y_iY_j)$$
   $$= E((X_i - E(X_i))(X_j - E(X_j)))E((Y_i - E(Y_i))(Y_j - E(Y_j)))$$
   $$= \text{Cov}(X_i, X_j)\text{Cov}(Y_i, Y_j)$$
   $$= M(i, j)N(i, j)$$

   So, the covariance matrix of $Z$ is the pointwise product of $M$ and $N$.

   (b) Since covariance matrices are always positive semidefinite, it follows that $Q$ is PSD.

3. Closure properties of kernels.

   In each case, we will establish that $k(x, x')$ is a kernel function by invoking Mercer’s condition. That is, we will show that for any finite set of points $x_1, \ldots, x_m \in \mathcal{X}$, the $m \times m$ matrix $K$ given by

   $$K_{ij} = k(x_i, x_j)$$

   is positive semidefinite.
Let $\Phi(x)$ be a function that maps $x$ to a point in $\mathcal{X}$ and define matrix $K$ as above. Also define $m \times m$ matrices $K^{(1)}$ and $K^{(2)}$ by

$$K^{(1)}_{ij} = k_1(x_i, x_j), \quad K^{(2)}_{ij} = k_2(x_i, x_j).$$

Since $k_1$ and $k_2$ are kernel functions, we know that $K^{(1)}$ and $K^{(2)}$ are PSD. And since the set of PSD matrices is closed under addition and under multiplication by a nonnegative scalar, it follows that $K = \alpha_1 K^{(1)} + \alpha_2 K^{(2)}$ is also PSD.

(b) Define $K, K^{(1)}, K^{(2)}$ as above. This time $K$ is the pointwise product of $K^{(1)}$ and $K^{(2)}$; by the previous problem, $K$ is PSD.

4. Let $\Phi(x) = (x_1^2, \ldots, x_m^2, \sqrt{2}x_1x_2, \ldots, \sqrt{2}x_{d-1}x_d)$, where all pairs of coordinates are included. Then

$$\Phi(x) \cdot \Phi(z) = \sum_{i=1}^{d} x_i^2 z_i^2 + 2 \sum_{i \neq j} x_i z_i z_j = (x_1 z_1 + x_2 z_2 + \ldots + x_d z_d)^2 = (x \cdot z)^2 = k(x, z).$$

5. (a) $K(x, z)$ is not a kernel.

For $x = [1, -1]$, we have $K(x, x) = 1 \times -1 = -1$. The corresponding kernel matrix $K = -1$. For $v = 1$, $v^T K v = -1 < 0$, which violates the PSD property. Thus $K$ is not a kernel.

(b) $K(x, z)$ is not a kernel.

For $x = [2, 2, \ldots]$, we have $K(x, x) = 1 - \langle x, x \rangle = 1 - 4d$. The corresponding kernel matrix $K = 1 - 4d$. For $v = 1$, $v^T K v = 1 - 4d < 0$, which violates the kernel PSD property for $d > 0$. Thus $K$ is not a kernel.

(c) $K(x, z)$ is not a kernel.

One way to prove that $K$ is not a kernel is to show a counterexample to the PSD property. Pick $x = [1, 0, \ldots, 0]$, $z = [2, 0, \ldots, 0]$, $v = [1, -1]^T$. Then the kernel matrix

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

and $v^T A v = -2 < 0$, which violates positivity.

A nice, second way to prove this is through contradiction. Suppose $K$ a kernel, such that $K(x, z) = \langle \phi(x), \phi(z) \rangle$. Recall the Cauchy-Schwarz Inequality for inner product, that we discussed in Lecture 2:

$$\langle \phi(x), \phi(z) \rangle^2 \leq \langle \phi(x), \phi(x) \rangle \cdot \langle \phi(z), \phi(z) \rangle \tag{1}$$

From this inequality,

$$K(x, z)^2 \leq K(x, x) \cdot K(z, z) \tag{2}$$

Suppose $x$ is any vector with norm 1 and let $z = 2x$. By the definition of $K$, we have $K(x, x) = \|x - x\|^2 = 0$ and $K(x, x) \cdot K(z, z) = 0$. However $K(x, z) = \|x - x\|^2 = 1 > K(x, x) \cdot K(z, z)$, which leads to a contradiction! Thus $K$ is not a kernel.

(d) $K(x, z)$ is a kernel corresponding to the feature map $\phi(x) = f(x_1, x_2)$.

(e) $K(x, z)$ is a kernel.

Recall that $a^2 - b^2 = (a - b) \cdot (a + b)$

Hence, we have

$$\frac{1 - \langle x, z \rangle^2}{1 - \langle x, z \rangle} = 1 + \langle x, z \rangle \tag{3}$$

In the above equation, we can rewrite 1 as $\langle x, z \rangle^0$
Thus, we can now write, $K(x, z) = K_0(x, z) + K_1(x, z)$. In Problem 1, we saw that the sum or product of two kernels is also a kernel. We know that $K_0(x, z)$ and $K_1(x, z)$ are both kernels. The feature map $φ_0(x)$ corresponding $K_0(x, z)$ is

$$φ_0(x) = 1$$ (4)

$K_1(x, z)$ corresponds to the feature map

$$φ_1(x) = x$$ (5)

Using Problem 3, $K(x, z) = K_0(x, z) + K_1(x, z)$ is a kernel corresponding to the feature map $φ'$, where for any $x$, $φ'(x)$ is a concatenation of the feature maps $φ_0(x)$ and $φ_1(x)$.

(f) $K(x, z)$ is a kernel.

Let $K_i(x, z) = \min(x_i, z_i)$. From Problem 3, we know that the sum of two kernels $K_1$ and $K_2$ is also a kernel whose corresponding feature map is the concatenation of the feature maps corresponding to $K_1$ and $K_2$. Thus if we can find the feature maps for all $K_i(x, z)$, then we can get the feature map for $K(x, z)$ by concatenating these maps. Consider following feature map:

$$φ_i(x) = [f_1(x_i), f_2(x_i), \ldots, f_{100}(x_i)]^T$$ (6)

where $f_k(t) = I(t \geq k) = \begin{cases} 1 & t \geq k \\ 0 & t < k \end{cases}$ . Without loss of generality, suppose that $x_i \leq z_i$. Then

$φ_i(x) = [1, \ldots, 1, 0, \ldots, 0]^T$ where only the first $x_i$ entries are 1. Analogously,

$φ_i(z) = [1, \ldots, 1, 0, \ldots, 0]^T$ where only the first $z_i$ entries are 1. Then

$$\langle φ_i(x), φ_i(z) \rangle = \sum_{i=1}^{x_i} 1 \cdot 1 + \sum_{i=x_i+1}^{z_i} 0 \cdot 1 + \sum_{i=z_i+1}^{100} 0 \cdot 0 = x_i = \min(x_i, z_i)$$

Therefore $K_i(x, z)$ is a kernel corresponding to the feature map $φ_i(x) = [f_1(x_i), f_2(x_i), \ldots, f_{100}(x_i)]^T$, and $K(x, z)$ is a kernel corresponding to the feature map $φ(x)$ which is a concatenation of the feature maps $φ_1(x), φ_2(x), \ldots, φ_d(x)$.

(g) $K(x, z)$ is a kernel.

Let $K_i(x, z) = 1 + x_i z_i$, then $K(x, z) = \prod_{i=0}^{d} K_i(x)$. From Problem 1, we know that the product of two kernels is also a kernel. Since $K_i(x, z)$ is a kernel corresponding to the feature map $φ_i(x) = [1, x_i]^T$, $K(x, z)$ is also a kernel. More specifically, $K(x, z)$ is a kernel corresponding to the feature map $φ(x)$, where for any $x$, $φ(x)$ has $2^d$ coordinates, one corresponding to each subset $S$ of $\{1, 2, \ldots, d\}$. $φ_S(x)$, the coordinate of $φ(x)$ corresponding to the set $S$ is $\prod_{i \in S} x_i$. This kernel is called the All Subsets kernel.

(h) $K(x, z)$ is not a kernel.

One way to prove this is by showing a violation of the PSD property. Let $x = [0, \ldots, 0], z = [1, 0, \ldots, 0]$ and $v = [1, -1]^T$. Then the kernel matrix

$$K = \begin{bmatrix} K(x, x) & K(x, z) \\ K(z, x) & K(z, z) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}$$

Thus, $v^T A v = -1 < 0$, which violates positivity.

Another nice way is through a violation of the Cauchy-Schwartz inequality. Consider $x = [0, \ldots, 0]$ and $z = [1, 0, \ldots, 0]$. Then $K(x, x) = 0$, $K(x, z) = K(z, z) = 1$, which violates Cauchy-Schwarz inequality – that is $K(x, z)^2 \geq K(x, x) \cdot K(z, z)$.

6. (a) First, we can compute the marginal distributions of $Y$ and $Z$ as follows,
Then, by definition of conditional probability, i.e. \( P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} \), we can get the conditional distributions of \( X|Y \) as follows:

\[
\begin{array}{c|cc}
  & 0 & 1 \\
  X & \frac{2}{5} & \frac{3}{5} \\
\end{array}
\]

Similarly we have the conditional distributions of \( X|Z \) as follows,

\[
\begin{array}{c|cc}
  & 0 & 1 \\
  X & \frac{1}{3} & \frac{2}{3} \\
\end{array}
\]

(b) By the definition of conditional entropy, \( H(X|Y) = P(Y = 0)H(X|Y = 0) + P(Y = 1)H(X|Y = 1) \).

\[
H(X|Y = 0) = -P(X = 0|Y = 0) \log P(X = 0|Y = 0) - P(X = 1|Y = 0) \log P(X = 1|Y = 0) \\
= -\frac{1}{2} \log \frac{1}{2} - \frac{1}{2} \log \frac{1}{2} \\
= \log 2
\]

Similarly we have

\[
H(X|Y = 1) = -P(X = 0|Y = 1) \log P(X = 0|Y = 1) - P(X = 1|Y = 1) \log P(X = 1|Y = 1) \\
= -\frac{1}{6} \log \frac{1}{6} - \frac{5}{6} \log \frac{5}{6} \\
= \log 6 - \frac{5}{6} \log 5
\]

Thus

\[
H(X|Y) = P(Y = 0)H(X|Y = 0) + P(Y = 1)H(X|Y = 1) \\
= \frac{2}{5} \log 2 + \frac{3}{5} \left( \log 6 - \frac{5}{6} \log 5 \right) \\
= \frac{2}{5} \log 2 + \frac{3}{5} \log 6 - \frac{1}{2} \log 5
\]

For \( H(X|Z) \), we can get

\[
H(X|Z = 0) = -P(X = 0|Z = 0) \log P(X = 0|Z = 0) - P(X = 1|Z = 0) \log P(X = 1|Z = 0) \\
= -\frac{1}{3} \log \frac{1}{3} - \frac{2}{3} \log \frac{2}{3} \\
= \log 3 - \frac{2}{3} \log 2
\]

Similarly we have

\[
H(X|Z = 1) = -P(X = 0|Z = 1) \log P(X = 0|Z = 1) - P(X = 1|Z = 1) \log P(X = 1|Z = 1) \\
= -\frac{3}{11} \log \frac{3}{11} - \frac{8}{11} \log \frac{8}{11} \\
= \log 11 - \frac{3}{11} \log 3 - \frac{8}{11} \log 8
\]
Thus
\[ H(X|Z) = P(Z = 0)H(X|Z = 0) + P(Z = 1)H(X|Z = 1) \]
\[ = \frac{9}{20} \left( \log 3 - \frac{2}{3} \log 2 \right) + \frac{11}{20} \left( \log 11 - \frac{3}{11} \log 3 - \frac{8}{11} \log 8 \right) \]
\[ = -\frac{3}{2} \log 2 + \frac{3}{10} \log 3 + \frac{11}{20} \log 11 \]

Using natural logarithm, the numerical values are shown as follows.

|        | \(H(X|Y)\)  | \(H(X|Y\) = 0) | 0.693147180560 |
|--------|--------------|----------------|---------------|
|        | \(H(X|Y)\)  | \(H(X|Y\) = 1) | 0.450561208866 |
|        | \(H(X|Z)\)  | \(H(X|Z\) = 0) | 0.54795597544 |
|        | \(H(X|Z)\)  | \(H(X|Z\) = 1) | 0.63651416829 |
|        | \(H(X|Z)\)  | \(H(X|Z\) | 0.5859526183 |
|        | \(H(X|Z)\)  | \(H(X|Z\) | 0.6087053158 |

(c) From the table above, \(H(X|Y) < H(X|Z)\). This suggests that there is less uncertainty in \(X\) when given \(Y\) than when given \(Z\). Therefore gene A is more informative about the cancer.

7. (a) False.

Two trees are said to be different in structure if
i. the trees are different in shape, or
ii. the splitting rule at some node is different.

Consider the following example. The input has two boolean features \(x_1\) and \(x_2\), and the class label is \(XOR(x_1, x_2)\), i.e.

<table>
<thead>
<tr>
<th>Feature 1</th>
<th>Feature 2</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

The following two trees do not have the same structure as they differ in splitting rules at each node, yet both predict the label correctly.

![Figure 1: Two trees with different structures but same prediction on boolean features \(x_1\) and \(x_2\).](image)

(b) False.

If \(T\) and \(T'\) produce zero error on the same training set \(S \subseteq \mathcal{X}\), then, \(\forall x \in S, T(x) = T'(x)\).

However, the training set typically does not include all elements in feature space \(\mathcal{X}\). Thus, there exist such \(x_0 \in \mathcal{X} - S\) that \(T(x_0) \neq T'(x_0)\). For example, consider the following training set:
For training set above, the two decision trees shown in Figure 2 both produce zero error. However, for the point $x_1 = (0, 1)$ or the point $x_2 = (1, 0)$, these two trees would give different predictions. Hence they are not equal.

Figure 2: Two Decision Trees with Zero Error on $S$

8. (a) Observe that $X$ is a random variable which takes values $k = 1, 2, 3, \ldots$. For a fixed integer $k$, we need $k$ flips to get the first head if the first $k - 1$ tosses come up tails, and the $k$-th toss comes up a head. Therefore,

$$p_k = \Pr(X = k) = \frac{1}{2^{k-1}} \cdot \frac{1}{2} = \frac{1}{2^k}$$

Therefore,

$$H(X) = -\sum_{k=1}^{\infty} p_k \log p_k = -\sum_{k=1}^{\infty} \frac{1}{2^k} \log \frac{1}{2^k} = \sum_{k=1}^{\infty} \log 2 \cdot \frac{k}{2^k}$$

The last step follows because $\log \frac{1}{2^k} = -k \log 2$. From the expressions given above, the sum is:

$$\sum_{k=1}^{\infty} \frac{k}{2^k} = \sum_{k=0}^{\infty} \frac{k}{2^k} = \frac{1}{(1 - \frac{1}{2})^2} = 2$$

Thus, $H(X) = 2 \log 2$.

(b) Let $p_i = \Pr(X = x_i)$ and let $q_j = \Pr(Y = x_{m+j})$. Then, $H(X) = -\sum_{i=1}^{m} p_i \log p_i$ and $H(Y) = -\sum_{j=1}^{n} q_j \log q_j$. By definition of $Z$, $Z$ takes values $x_i, 1 \leq i \leq m$ with probability $\alpha p_i$, and values $x_{m+j}, 1 \leq j \leq n$ with probability $(1 - \alpha)q_j$. Therefore,

$$H(Z) = -\sum_{i=1}^{m} \alpha p_i \log \alpha p_i - \sum_{j=1}^{n} (1 - \alpha)q_j \log(1 - \alpha)q_j$$

$$= -\sum_{i=1}^{m} \alpha p_i \log \alpha - \sum_{i=1}^{m} \alpha p_i \log p_i - \sum_{j=1}^{n} (1 - \alpha)q_j \log(1 - \alpha) - \sum_{j=1}^{n} (1 - \alpha)q_j \log q_j$$

$$= \alpha H(X) + (1 - \alpha)H(Y) - \alpha \log \alpha - (1 - \alpha) \log(1 - \alpha)$$

Here the last step follows from the observation that $\sum_{i=1}^{m} p_i = 1$ and $\sum_{j=1}^{n} q_j = 1$. 

<table>
<thead>
<tr>
<th>Feature 1</th>
<th>Feature 2</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feature1 &lt; 0.5?</th>
<th>Predict 0</th>
<th>Feature2 &lt; 0.5?</th>
<th>Predict 1</th>
</tr>
</thead>
</table>