CSE 251A: Homework 4 Solutions

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:

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

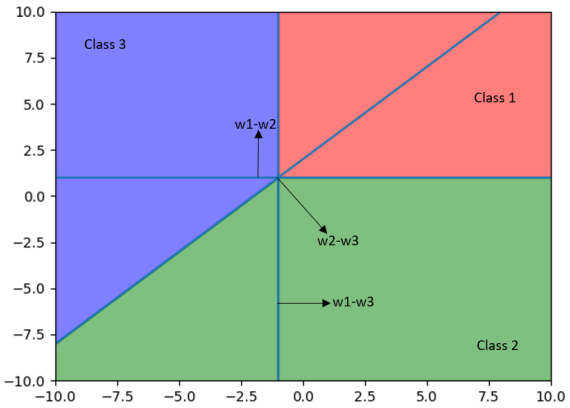
   The region where class 1 beats class 3 is:

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

   The region where class 2 beats class 3 is:

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

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



2. (a) $K(x, z) = \langle \phi(x), \phi(z) \rangle$ for the feature map ϕ, and let $\phi'(x) = \sqrt{c}\phi(x)$. Then, for all x and z,

   $$K'(x, z) = cK(x, z) = c\langle \phi(x), \phi(z) \rangle = \langle \sqrt{c}\phi(x), \sqrt{c}\phi(z) \rangle$$

   This establishes that $K'(x, z)$ is a kernel corresponding to the feature map ϕ'.

   (b) Suppose x_0 is the x for which $K(x, x) > 0$. Consider the 1×1 kernel matrix $K' = K'(x_0, x_0)$ for the kernel K' and the data point x_0. Then, $K' = cK(x_0, x_0)$. If $z = 1$, then $z^\top K' z = cK(x_0, x_0) < 0$, which violates the kernel Positive Semi Definiteness (PSD) property of kernel matrices. This proves that K' is not a kernel.

   (c) We have that $K(x, z) = \langle \phi(x), \phi(z) \rangle$ and $L(x, z) = \langle \psi(x), \psi(z) \rangle$. Then, for all x and z,

   $$\begin{aligned} K'(x, z) &= a\langle \phi(x), \phi(z) \rangle + b\langle \psi(x), \psi(z) \rangle = \langle \sqrt{a}\phi(x), \sqrt{a}\phi(z) \rangle + \langle \sqrt{b}\psi(x), \sqrt{b}\psi(z) \rangle \\ &= \langle \phi'(x), \phi'(z) \rangle \end{aligned}$$

   where $\phi'(x)$ is a concatenation of the feature maps $\sqrt{a}\phi(x)$ and $\sqrt{b}\psi(x)$. In other words, if the feature maps ϕ and ψ have m and n coordinates respectively, then ϕ' has $m + n$ coordinates; for any x, the first m coordinates of $\phi'(x)$ are $(\sqrt{a}\phi_1(x), \sqrt{a}\phi_2(x), \ldots, \sqrt{a}\phi_m(x))$ and the remaining n coordinates of $\phi'(x)$ are $(\sqrt{b}\psi_1(x), \sqrt{b}\psi_2(x), \ldots, \sqrt{b}\psi_n(x))$. Therefore $K'(x, z)$ is a kernel corresponding to the feature map ϕ'.

(d) Suppose $K(x,z) = \langle \phi(x), \phi(z) \rangle$ and $L(x,z) = \langle \psi(x), \psi(z) \rangle$. If ϕ and ψ have m and n coordinates respectively, then, for all x and z,

$$\begin{aligned} K'(x,z) &= K(x,z)L(x,z) = \langle \phi(x), \phi(z) \rangle \cdot \langle \psi(x), \psi(z) \rangle \\ &= \left(\sum_{i=1}^{m} \phi_i(x)\phi_i(z) \right) \cdot \left(\sum_{j=1}^{n} \psi_j(x)\psi_j(z) \right) \\ &= \sum_{i=1}^{m} \sum_{j=1}^{n} (\phi_i(x)\psi_j(x)) \cdot (\phi_i(z)\psi_j(z)) \end{aligned}$$

where the last step follows from using the fact that $(\sum_i a_i) \cdot (\sum_j b_j) = \sum_{i,j} a_i b_j$. Now observe that the last quantity is the dot product $\langle \phi'(x), \phi'(z) \rangle$ where ϕ' is a mn dimensional feature map. It has a coordinate $\phi'_{(i,j)}(\cdot)$ corresponding to each pair (i,j), $1 \le i \le m, 1 \le j \le n$, and $\phi'_{(i,j)}(x) = \phi_i(x)\psi_j(x)$. Thus $K'(x,z)$ is a kernel corresponding to the feature map ϕ'.

3. For the cases where K is not a kernel, we present counterexamples; where K is a kernel, we present the feature map.

(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 $v = 1$, $v^\top K v = -1 < 0$, which violates the PSD property. Thus K is not a kernel.

(b) $K(x,z)$ is not a kernel.
For $x = [2, 2, \cdots]$, we have $K(x,x) = 1 - \langle x, x \rangle = 1 - 4d$. The corresponding kernel matrix $K = 1 - 4d$. For $v = 1$, $v^\top K v = 1 - 4d < 0$, 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 $x = [1, 0, \ldots, 0], z = [2, 0, \ldots, 0], v = [1, -1]^\top$. Then the kernel matrix

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

and $v^\top A v = -2 < 0$, 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:

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

From this inequality,

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

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 $K(x,z) = \|x - 2x\|^2 = 1 > K(x,x) \cdot K(z,z)$, which leads to a contradiction! Thus K is not a kernel.

(d) $K(x,z)$ is a kernel corresponding to the feature map $\phi(x) = f(x_1, x_2)$. To show this, note that for any x and z, $K(x,z)$ can be factored as $K(x,z) = \langle \phi(x), \phi(z) \rangle$.

(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 2, 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 $\phi_0(x)$ corresponding $K_0(x,z)$ is

$$\phi_0(x) = 1 \tag{4}$$

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

$$\phi_1(x) = x \tag{5}$$

Using Problem 2, $K(x,z) = K_0(x,z) + K_1(x,z)$ is a kernel corresponding to the feature map ϕ', where for any x, $\phi'(x)$ is a concatenation of the feature maps $\phi_0(x)$ and $\phi_1(x)$.

(f) $K(x,z)$ is a kernel.
Let $K_i(x,z) = \min(x_i, z_i)$. From Problem 2, 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:

$$\phi_i(x) = [f_1(x_i), f_2(x_i), \ldots, f_{100}(x_i)]^\top \tag{6}$$

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 $\phi_i(x) = [1, \ldots, 1, 0, \ldots, 0]^\top$ where only the first x_i entries are 1. Analogously, $\phi_i(z) = [1, \ldots, 1, 0, \ldots, 0]^\top$ where only the first z_i entries are 1. Then

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

Therefore $K_i(x,z)$ is a kernel corresponding to the feature map $\phi_i(x) = [f_1(x_i), f_2(x_i), \ldots, f_{100}(x_i)]^\top$, and $K(x,z)$ is a kernel corresponding to the feature map $\phi(x)$ which is a concatenation of the feature maps $\phi_1(x), \phi_2(x), \ldots, \phi_d(x)$.

(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 2, 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 $\phi_i(x) = [1, x_i]^\top$, $K(x,z)$ is also a kernel. More specifically, $K(x,z)$ is a kernel corresponding to the feature map $\phi(x)$, where for any x, $\phi(x)$ has 2^d coordinates, one corresponding to each subset S of $\{1, 2, \ldots, d\}$. $\phi_S(x)$, the coordinate of $\phi(x)$ 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 $v = [1, -1]^\top$. 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, $v^\top A v = -1 < 0$, 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 $K(x,z)^2 \geq K(x,x) \cdot K(z,z)$.

4. (a) First, we can compute the marginal distributions of Y and Z as follows,

y	0	1
$P(Y=y)$	$\frac{2}{5}$	$\frac{3}{5}$

z	0	1
$P(Z=z)$	$\frac{9}{20}$	$\frac{11}{20}$

Then, by definition of conditional probability, i.e. $P(X=x|Y=y)=\dfrac{P(X=x,Y=y)}{P(Y=y)}$, we can get the conditional distributions of $X|Y$ as follows.

x	0	1
$P(X=x\|Y=0)$	$\frac{1}{2}$	$\frac{1}{2}$
$P(X=x\|Y=1)$	$\frac{1}{6}$	$\frac{5}{6}$

Similarly we have the conditional distributions of $X|Z$ as follows,

x	0	1
$P(X=x\|Z=0)$	$\frac{1}{3}$	$\frac{2}{3}$
$P(X=x\|Z=1)$	$\frac{3}{11}$	$\frac{8}{11}$

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

$$\begin{aligned} H(X|Y=0) &= -P(X=0|Y=0)\log P(X=0|Y=0) - P(X=1|Y=0)\log P(X=1|Y=0) \\ &= -\frac{1}{2}\log\frac{1}{2}-\frac{1}{2}\log\frac{1}{2} \\ &= \log 2 \end{aligned}$$

Similarly we have

$$\begin{aligned} H(X|Y=1) &= -P(X=0|Y=1)\log P(X=0|Y=1) - P(X=1|Y=1)\log P(X=1|Y=1) \\ &= -\frac{1}{6}\log\frac{1}{6}-\frac{5}{6}\log\frac{5}{6} \\ &= \log 6 - \frac{5}{6}\log 5 \end{aligned}$$

Thus

$$\begin{aligned} H(X|Y) &= P(Y=0)H(X|Y=0)+P(Y=1)H(X|Y=1) \\ &= \frac{2}{5}\log 2 + \frac{3}{5}\left(\log 6 - \frac{5}{6}\log 5\right) \\ &= \frac{2}{5}\log 2 + \frac{3}{5}\log 6 - \frac{1}{2}\log 5 \end{aligned}$$

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

$$\begin{aligned} H(X|Z=0) &= -P(X=0|Z=0)\log P(X=0|Z=0) - P(X=1|Z=0)\log P(X=1|Z=0) \\ &= -\frac{1}{3}\log\frac{1}{3}-\frac{2}{3}\log\frac{2}{3} \\ &= \log 3 - \frac{2}{3}\log 2 \end{aligned}$$

Similarly we have

$$\begin{aligned} H(X|Z=1) &= -P(X=0|Z=1)\log P(X=0|Z=1) - P(X=1|Z=1)\log P(X=1|Z=1) \\ &= -\frac{3}{11}\log\frac{3}{11}-\frac{8}{11}\log\frac{8}{11} \\ &= \log 11 - \frac{3}{11}\log 3 - \frac{8}{11}\log 8 \end{aligned}$$

Thus

$$\begin{aligned} H(X|Z) &= P(Z=0)H(X|Z=0) + P(Z=1)H(X|Z=1) \\ &= \frac{9}{20}\left(\log 3 - \frac{2}{3}\log 2\right) + \frac{11}{20}\left(\log 11 - \frac{3}{11}\log 3 - \frac{8}{11}\log 8\right) \\ &= -\frac{3}{2}\log 2 + \frac{3}{10}\log 3 + \frac{11}{20}\log 11 \end{aligned}$$

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

$H(X\|Y=0)$	0.693147180560
$H(X\|Y=1)$	0.450561208866
$H(X\|Y)$	0.547595597544
$H(X\|Z=0)$	0.63651416829
$H(X\|Z=1)$	0.5859526183
$H(X\|Z)$	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.

5. (a) False.
If T and T' produce zero error on the same training set S, then they have the same output on all x in this training set. However for most real problems, the training set does not include all elements in the input space, and there exist such $x_0 \in \mathcal{X} - S$; if we pick two trees T and T' such that $T(x_0) \neq T'(x_0)$, then those trees would be unequal. For example, for the following training set,

Feature 1	Feature 2	Label
0	0	0
1	1	1

the two decision trees shown in Figure 1 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.

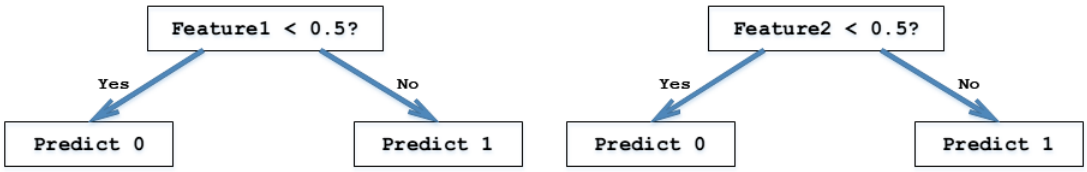


Figure 1: Two Decision Trees with Zero Error on S

(b) True.
If the trees are based on the same training set, and use the same splitting strategy (which they do since they are both ID3 Decision trees), then, the structure of the trees are not affected by the order in which the nodes are split. We can show this by induction on the nodes in order of the time at which they are split while building T.

The inductive hypothesis is as follows: for each $t = 1, 2, \ldots$, the children of node u_t (which is split at time t in tree T), and the data points associated with them are exactly the same as the children of node u_t in T' and the data points associated with them. The base case is $t = 1$, which corresponds to splitting the root in both trees. Since the splitting strategy is the same, the children nodes of the root in both T and T' as well as the data points associated with them are exactly the same as well.

In the inductive case, suppose the induction hypothesis holds until time $t - 1$. Then the structure of the subpath from the root to u_t, and the data points associated with each node on this subpath

is exactly the same in T and T'. Moreover, because the splitting strategy is the same, and because in an ID3 decision tree, the strategy for splitting u_t depends only on the data points at u_t, and is independent of the rest of the tree, when we decide to split u_t while building T', we will also split it into exactly the same nodes which have the same data points associated with them. The inductive case thus holds. Thus, T and T' have exactly the same structure and are therefore equal.

6. (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:

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

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, $1 \le i \le m$ with probability αp_i, and values x_{m+j}, $1 \le j \le n$ with probability $(1-\alpha)q_j$. Therefore,

$$\begin{aligned} H(Z) &= -\sum_{i=1}^{m} \alpha p_i \log \alpha p_i - \sum_{j=1}^{n} (1-\alpha) q_j \log (1-\alpha) q_j \\ &= -\sum_{i=1}^{m} \alpha p_i \log \alpha - \sum_{i=1}^{m} \alpha p_i \log p_i - \sum_{j=1}^{n} (1-\alpha) q_j \log (1-\alpha) - \sum_{j=1}^{n} (1-\alpha) q_j \log q_j \\ &= \alpha H(X) + (1-\alpha) H(Y) - \alpha \log \alpha - (1-\alpha) \log (1-\alpha) \end{aligned}$$

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