**CSE 251A: Machine learning** **Winter 2021**

# Homework 4 — Multiclass classification, Kernels and Decision Trees

1. A linear predictor is used to solve a classification problem with three classes. The data is two-dimensional and the linear functions for each class are:
   - Class 1: $w_1 = (1, 1)$, $b_1 = 0$
   - Class 2: $w_2 = (1, 0)$, $b_2 = 1$
   - Class 3: $w_3 = (0, 1)$, $b_3 = -1$

   Draw the resulting decision boundary and clearly mark the region corresponding to each class.

2. In the following problems, suppose that $K$ and $L$ are kernels with feature maps $\phi$ and $\psi$ respectively. For the following functions $K'(x, z)$, state if they are kernels or not. If they are kernels, write down the corresponding feature map, in terms of $\phi$, $\psi$ and constants $a$, $b$ and $c$. If they are not kernels, prove that they are not.
   (a) $K'(x, z) = cK(x, z)$, for $c > 0$.
   (b) $K'(x, z) = cK(x, z)$, where $c < 0$, and there exists some $x$ for which $K(x, x) > 0$.
   (c) $K'(x, z) = aK(x, z) + bL(x, z)$ for $a, b > 0$.
   (d) $K'(x, z) = K(x, z)L(x, z)$.

3. For each of the following functions $K(x, z)$, state if it is a kernel or not. If the function is a kernel, then write down its feature map. If it is not a kernel, prove that it is not one. For your proof, you can use the answers to Problem 2.
   (a) $x = [x_1, x_2]$, $z = [z_1, z_2]$, $x_1, x_2, z_1, z_2$ are real numbers. $K(x, z) = x_1 z_2$.
   (b) Let $x = [x_1, \ldots, x_d]$, $z = [z_1, \ldots, z_d]$, $x_i$s and $z_i$s are real numbers. $K(x, z) = 1 - \langle x, z \rangle$.
   (c) $x = [x_1, \ldots, x_d]$, $z = [z_1, \ldots, z_d]$, $x_i$s and $z_i$s are real numbers. $K(x, z) = \|x - z\|^2$.
   (d) $x = [x_1, \ldots, x_d]$, $z = [z_1, \ldots, z_d]$, and $f$ is a function. $K(x, z) = f(x_1, x_2)f(z_1, z_2)$.
   (e) $x = [x_1, \ldots, x_d]$, $z = [z_1, \ldots, z_d]$, $x_i$s and $z_i$s are real numbers. $K(x, z) = \frac{1 - \langle x, z \rangle^2}{1 - \langle x, z \rangle}$.
   (f) $x = [x_1, \ldots, x_d]$, $z = [z_1, \ldots, z_d]$, $x_i$s and $z_i$s are integers between 0 and 100. $K(x, z) = \sum_{i=1}^{d} \min(x_i, z_i)$.
   (g) $x = [x_1, \ldots, x_d]$, $z = [z_1, \ldots, z_d]$, $x_i$s and $z_i$s are real numbers.
   $$K(x, z) = (1 + x_1 z_1)(1 + x_2 z_2) \ldots (1 + x_d z_d)$$
   (h) $x = [x_1, \ldots, x_d]$, $z = [z_1, \ldots, z_d]$, $x_i$s and $z_i$s are integers between 0 and 100. $K(x, z) = \sum_{i=1}^{d} \max(x_i, z_i)$.

4. A group of biologists would like to determine which genes are associated with a certain form of liver cancer. After much research, they have narrowed the possibilities down to two genes, let us call them A and B. After analyzing a lot of data, they have also calculated the following joint probabilities.

| | Cancer | No Cancer |
|---|---|---|
| Gene A | $\frac{1}{2}$ | $\frac{1}{10}$ |
| No Gene A | $\frac{1}{5}$ | $\frac{1}{5}$ |

| | Cancer | No Cancer |
|---|---|---|
| Gene B | $\frac{2}{5}$ | $\frac{3}{20}$ |
| No Gene B | $\frac{3}{10}$ | $\frac{3}{20}$ |

(a) Let $X$ denote the 0/1 random variable which is 1 when a patient has cancer and 0 otherwise. Let $Y$ denote the 0/1 random variable which is 1 when gene $A$ is present, 0 otherwise, and let $Z$ denote the 0/1 random variable which is 1 when gene $B$ is present and 0 otherwise. Write down the conditional distributions of $X|Y = y$ for $y = 0, 1$ and $X|Z = z$, for $z = 0, 1$.

(b) Calculate the conditional entropies $H(X|Y)$ and $H(X|Z)$.

(c) Based on these calculations, which of these genes is more informative about cancer?

5. Since a decision tree is a classifier, it can be thought of as a function that maps a feature vector $x$ in some set $\mathcal{X}$ to a label $y$ in some set $\mathcal{Y}$. We say two decision trees $T$ and $T'$ are *equal* if for all $x \in \mathcal{X}$, $T(x) = T'(x)$.

   The following are some statements about decision trees. For these statements, assume that $\mathcal{X} = \mathbb{R}^d$, that is, the set of all $d$-dimensional feature vectors. Also assume that $\mathcal{Y} = \{1, 2, \ldots, k\}$. Write down if each of these statements are correct or not. If they are correct, provide a brief justification or proof; if they are incorrect, provide a counterexample to illustrate a case when they are incorrect.

   (a) If $T$ and $T'$ are any two decision trees that produce zero error on the same training set, then they are equal.

   (b) $T$ and $T'$ are two ID3 Decision Trees built on the same training dataset. No pruning is used on either. The only difference is that while building $T$, we pick the internal nodes in a Depth First Search order for splitting while for $T'$ we use a Breadth First Search order.

6. (a) A fair coin (that is, a coin with equal probability of coming up heads and tails) is flipped until the first head occurs. Let $X$ denote the number of flips required. What is the entropy $H(X)$ of $X$? You may find the following expressions useful:

$$\sum_{j=0}^{\infty} r^j = \frac{1}{1-r}, \quad \sum_{j=0}^{\infty} j r^j = \frac{r}{(1-r)^2}$$

   (b) Let $X$ be a discrete random variable which takes values $x_1, \ldots, x_m$ and let $Y$ be a discrete random variable which takes values $x_{m+1}, \ldots, x_{m+n}$. (That is, the values taken by $X$ and the values taken by $Y$ are disjoint.) Let:

$$\begin{aligned} Z &= X \text{ with probability } \alpha \\ &= Y \text{ with probability } 1-\alpha \end{aligned}$$

   Find $H(Z)$ as a function of $H(X)$, $H(Y)$ and $\alpha$.