CSE 250B: Machine learning **Winter 2019**

Homework 5 — Kernels and Decision Trees

Problem 1

In the following problems, suppose that K, K_1 and K_2 are kernels with feature maps ϕ, ϕ^1 and ϕ^2. 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 ϕ, ϕ^1, ϕ^2 and c, c_1, c_2. If they are not kernels, prove that they are not.

1. $K'(x,z) = cK(x,z)$, for $c > 0$. True.
2. $K'(x,z) = cK(x,z)$, where $c < 0$, and there exists some x for which $K(x,x) > 0$. False.
3. $K'(x,z) = c_1K_1(x,z) + c_2K_2(x,z)$ for $c_1, c_2 > 0$. True.
4. $K'(x,z) = K_1(x,z)K_2(x,z)$. True. Product of two kernels is a kernel.

Problem 2

For 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 1.

1. $x = [x_1, x_2]$, $z = [z_1, z_2]$, x_1, x_2, z_1, z_2 are real numbers. $K(x,z) = x_1z_2$. False. x = [1,-1] => K(x,x)<0
2. Let $x = [x_1, \ldots, x_d]$, $z = [z_1, \ldots, z_d]$, x_is and z_is are real numbers. $K(x,z) = 1 - \langle x, z\rangle$. False. x = [1,0,0] z = [2,0,0,0]
3. $x = [x_1, \ldots, x_d]$, $z = [z_1, \ldots, z_d]$, x_is and z_is are real numbers. $K(x,z) = \|x - z\|^2$. False. Cauchy-schwarz inequality
4. $x = [x_1, \ldots, x_d]$, $z = [z_1, \ldots, z_d]$, and f is a function. $K(x,z) = f(x_1, x_2)f(z_1, z_2)$. True
5. $x = [x_1, \ldots, x_d]$, $z = [z_1, \ldots, z_d]$, x_is and z_is are real numbers. $K(x,z) = \frac{1-\langle x,z\rangle^2}{1-\langle x,z\rangle}$. True
6. $x = [x_1, \ldots, x_d]$, $z = [z_1, \ldots, z_d]$, x_is and z_is are integers between 0 and 100. $K(x,z) = \sum_{i=1}^{d} \min(x_i, z_i)$. True
7. $x = [x_1, \ldots, x_d]$, $z = [z_1, \ldots, z_d]$, x_is and z_is are real numbers. True
$$K(x,z) = (1 + x_1z_1)(1 + x_2z_2)\ldots(1 + x_dz_d)$$
8. $x = [x_1, \ldots, x_d]$, $z = [z_1, \ldots, z_d]$, x_is and z_is are integers between 0 and 100. $K(x,z) = \sum_{i=1}^{d} \max(x_i, z_i)$.

False. Contradictory with Cauchy-Swartz inequality.

Problem 3

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	$\frac{3}{10}$	$\frac{3}{20}$

1. 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 $z = 0, 1$.
2. Calculate the conditional entropies $H(X|Y)$ and $H(X|Z)$.
3. Based on these calculations, which of these genes is more informative about cancer?

Problem 4

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.

1. If the decision trees T and T' do not have exactly the same structure, then they can never be equal.
2. If T and T' are any two decision trees that produce zero error on the same training set, then they are equal.

Problem 5

In this problem, we will formally examine how transforming the training data in simple ways can affect the performance of common classifiers. Transforming training features by scaling is equivalent to measuring these features in different units; in practice, we frequently have to combine multiple homogeneous or heterogeneous features, and it is important to understand how changing units in which features are measured can affect machine learning algorithms.

Suppose we are given a training data set $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ where each feature vector x_i lies in d-dimensional space. Suppose each $x_i = [x_i^1, x_i^2, \ldots, x_i^d]$, so coordinate j of x_i is denoted by x_i^j.

For each x_i, suppose we transform it to z_i by rescaling each axis of the data by a fixed factor; that is, for every $i = 1, \ldots, n$ and every coordinate $j = 1, \ldots, d$, we write:

$$z_i^j = \alpha^j x_i^j$$

Here α^js are real, non-zero and positive constants. Thus, our original training set S is transformed after rescaling to a new training set $S' = \{(z_1, y_1), \ldots, (z_n, y_n)\}$. For example, if we have two features, and if $\alpha^1 = 3$, and $\alpha^2 = 2$, then, a feature vector $x = (x^1, x^2)$ gets transformed by rescaling to $z = (z^1, z^2) = (3x^1, 2x^2)$.

A classifier $C(x)$ in the original space (of x's) is said to be equal to a classifier $C'(z)$ in the rescaled space (of z's) if for every $x \in \mathbb{R}^d$, $C(x) = C'(z)$, where z is obtained by transforming x by recaling. In our previous example, the classifier C in the original space:

$$C(x) : \textit{Predict } 0 \textit{ if } x^1 \leq 1, \textit{ else predict } 1.$$

is equal to the classifier C' in the rescaled space:

$$C'(z)\text{: } \textit{Predict } 0 \textit{ if } z^1 \leq 3, \textit{ else predict } 1.$$

This is because if $C(x) = 0$ for an $x = (x^1, x^2)$, then $x^1 \leq 1$. This means that for the transformed vector $z = (z^1, z^2) = (3x^1, 2x^2)$, $z^1 = 3x^1 \leq 3$, and thus $C'(z) = 0$ as well. Similarly, if $C(x) = 1$, then $x^1 > 1$ and $z^1 > 3$ and thus $C(z) = 1$. Now, answer the following questions:

1. First, suppose that all the α^i values are equal; that is, $\alpha^1 = \ldots = \alpha^d$. Suppose we train a k-NN classifier C on S and a k-NN classifier C' on S'. Are these two classifiers equal? What if we trained C and C' on S and S' respectively using the ID3 Decision Tree algorithm? What if we trained C and C' on S and S' respectively using the Perceptron algorithm? If the classifiers are equal, provide a *brief* argument to justify why; if they are not equal, provide a counterexample.

2. Repeat your answers to the questions in part (1) when the α_is are different. Provide a *brief* justification for each answer if the classifiers are equal, and a counterexample if they are not.

3. From the results of parts (1) and (2), what can you conclude about how k-NN, decision trees and perceptrons behave under scaling transformations?