COMP 6321, Fall 2017 **CONCORDIA UNIVERSITY**

MACHINE LEARNING.
Assignment 3.
Due: November 10, 2017.

1. [20 points] **A midterm preparation question**

   For each of the learning problems outlined below, specify what is the best learning algorithm to use and why. Note that you should give *one* algorithm for each problem, even if there are several correct answers.

   (a) You have about 1000 training examples in a 6-dimensional continuous feature space. You only expect to be asked to classify 100 test examples.

   (b) You are going to develop a classifier to recommend which children should be assigned to special education classes in kindergarten. The classifier has to be justified to the board of education before it is implemented.

   (c) You are working for a huge retailing company. You are trying to predict whether customer X will like a particular item, as a function of the input which is a vector of 1 million bits specifying whether each of the other customers liked the item. You will train a classifier on a very large data set of items, where the inputs are everyone else preferences for that item, and the output is customer Xs preference for that item. The classifier will have to be updated frequently and efficiently as new data comes in.

   (d) You are working for an oil company which is trying to decide where to drill. You have 40 attributes, both discrete and continuous, that describe a plot of land. Some of these attributes are noisy. For previous sites, you know whether they contained oil or not, but you only have data about 50 such sites.

2. [20 points] **Properties of entropy**

   (a) Define joint probability mass function for random variables X, Y as follows

   $$p(0,0) = 1/3, p(0,1) = 1/3, p(1,0) = 0, p(1,1) = 1/3.$$

   Evaluate the following quantities

   i. $H[x]$ (entropy)

   ii. $H[y]$

   iii. $H[y|x]$ (conditional entropy)

   iv. $H[x|y]$

   v. $H[x,y]$

   vi. $I[x,y]$ (mutual information).

   (b) Prove that maximum entropy of a discrete probability distribution is achieved for a uniform distribution.

(c) Prove that maximum entropy of a continuous distribution with a given mean and variance is achieved for Gaussian distribution.

(d) Prove rigorously using information gain that test T1 wins (see Lecture 4, slide 90).

3. [25 points] **Kernels**

In this problem, we consider constructing new kernels by combining existing kernels. Recall that for some function $K(\mathbf{x}, \mathbf{z})$ to be a kernel, we need to be able to write it as a dot product of vectors from some high-dimensional feature space:

$$K(x, z) = \phi(\mathbf{x})^T \phi(\mathbf{z})$$

Mercer's theorem gives a necessary and sufficient condition for a function K to be a kernel: its corresponding kernel matrix has to be symmetric and positive semi-definite.

Suppose that $K_1(\mathbf{x}, \mathbf{z})$ and $K_2(\mathbf{x}, \mathbf{z})$ are kernels over $\Re^n \times \Re^n$. For each of the cases below, state whether K is also a kernel. If it is, prove it. If it is not, give a counterexample. You can use either Mercer's theorem, or the definition of a kernel as needed

(a) $K(\mathbf{x}, \mathbf{z}) = aK_1(\mathbf{x}, \mathbf{z}) + bK_2(\mathbf{x}, \mathbf{z})$, where $a, b > 0$ are real numbers

(b) $K(\mathbf{x}, \mathbf{z}) = aK_1(\mathbf{x}, \mathbf{z}) - bK_2(\mathbf{x}, \mathbf{z})$, where $a, b > 0$ are real numbers

(c) $K(\mathbf{x}, \mathbf{z}) = K_1(\mathbf{x}, \mathbf{z})K_2(\mathbf{x}, \mathbf{z})$

(d) $K(\mathbf{x}, \mathbf{z}) = f(\mathbf{x})f(\mathbf{z})$ where $f : \Re^n \to \Re$ is a real-valued function

(e) $K(\mathbf{x}, \mathbf{z}) = p(\mathbf{x})p(\mathbf{z})$ where p is a probability density function.

4. [5 points] **Nearest neighbor vs decision trees**

Consider a classification problem with instances consisting of two real-valued attributes. Do the decision boundaries that are possible for 1-nearest neighbor classification rule and decision trees coincide, or are they different? Justify your answer.

5. [10 marks] It is easy to see that the nearest-neighbor error rate P can equal the Bayes rate P^*. if $P^* = 0$ (the best possibility) or if $P^* = (c-1)/c$ (the worst possibility). One might ask whether or not there are problems for which $P = P^*$ when P^* is between these extremes.

(a) Show that the Bayes rate for the one-dimensional case where $P(\omega_i) = 1/c$ and

$$p(x|\omega_i) = \begin{cases} 1 & 0 \le x \le \frac{cr}{c-1} \\ 1 & i \le x \le i + 1 - \frac{cr}{c-1} \\ 0 & \text{elsewhere} \end{cases}$$

is $P^* = r$, where $P^* = 1 - \int \max_{1 \le i \le c} P(\omega_i|x)p(x)dx$
(see Duda, Hart, Stork, *Pattern Classification*, Wiley, 2000, p. 68, ex. 12).

(b) Show that for this case that the nearest-neighbor rate is $P = P^*$.
Hint: Use the formula for the asymptotic error rate of the 1-nearest-neighbor rule

$$L_{NN} = \int \left[1 - \sum_{i=1}^{c} P^2(\omega_i|x)\right] p(x)dx.$$

6. [25 points] **Implementation**

   In this problem, you are asked to perform *one* of the implementation and data analysis tasks below (of your choosing). Either one is worth 25 points.

   (a) Implement Adaboost with decision stumps, and run it on the Wisconsin data set, using 10-fold cross-validation. Plot training and testing error curves as a function of the number of training rounds.

   (b) Implement k-nearest neighbor on the Wisconsin data set, using Euclidean distance. Perform cross-validation for different values of k and plot training and testing error curves as a function of k.