ENGN 2520 Midterm exam

NAME:

BANNER ID:

BROWN EMAIL:

(LEAVE THIS BLANK)

problem	grade	memo
1		
2		
3		
4		
total		

Problem 1

We want to classify fish into two types: bass and salmon. We have 2 real valued features: length and weight.

Suppose we assume the length and weight of a fish are independent conditional on the fish type, and each is distributed according to a gaussian distribution with mean and variance that depends on the fish type.

We estimate the parameters of

- $p(length|type = bass)$,
- $p(weight|type = bass)$,
- $p(length|type = salmon)$,
- $p(weight|type = salmon)$,
- $p(type = salmon)$,
- $p(type = bass)$,

using maximum likelihood estimation with a training set T.

(a) What is the classification rule defined by the Bayes optimal classifier in terms of the six distributions specified above. You should use a 0/1 loss to derive classifier.

(b) Suppose the resulting classifier has poor performance on a test set. Give 3 different reasons why this might happen in practice.

Problem 2

Given short answers for the questions below.

(a) Suppose we want to use regression to fit a polynomial function to some training data T. How can we pick the degree of the polynomial to use in the regression?

(b) What is an advantage of a multilayer neural network over a linear classifier?

(c) What is an advantage of a linear classifier over a multilayer neural network?

(d) Consider a binary classification problem with a set of training examples T. Let H_1 and H_2 be two sets of classifers and suppose the VC dimension of H_1 is smaller than the VC dimension of H_2. Suppose we find classifiers $c_1 \in H_1$ and $c_2 \in H_2$ that both correctly classify all examples in T. How should we pick between c_1 and c_2? Give a brief justification.

Problem 3

Consider a binary classification problem where the input is a point in $\mathbb{R}$.

Suppose we are interested in learning classifiers that are defined in terms of intervals $[a, b]$ with $a < b$. Examples inside the interval are classified as positive, and examples outside the interval are classified as negative.

Define a feature map $\phi : \mathbb{R} \to \mathbb{R}^D$ such that every interval classifier can be represented by a linear classifier in the feature space defined by ϕ. Justify your answer.

Problem 4

Suppose you have an instrument that measures the speed of a running animal. You install the instrument in a habitat that contains two kinds of animals: tigers and cheetahs. The speed of a tiger is a random variable with probability density $p_1(x)$. The speed of a cheetah is a random variable with probability density $p_2(x)$. Suppose p_1 and p_2 are known. These densities are different but they have significant overlap, so it is hard to decide if an animal is a tiger or a cheetah based on a single observation of its speed.

(a) Let x be the speed of a random animal in the habitat. Let w_1 be the fraction of animals in the habitat that are tigers, and w_2 be the fraction of animals in the habitat that are cheetahs (since there are no other animals in this habitat we have $w_1 + w_2 = 1$). What is the probability density of x?

(b) Suppose we record the speed of n animals $x_1, \ldots, x_n$. Assume each measurement comes from a random animal in the habitat. How can we estimate w_1 and w_2 from these measurements? Give an algorithm for estimating w_1 and w_2. Justify your answer.