CSci 5521-002 Homework 1 Due Sept 28, 2018

1. After your yearly checkup, the doctor has good news and bad news. The bad news is that you tested positive for a serious disease and that the test is very accurate: the probability of testing positive when you do have the disease is 0.983, and the probability of testing negative when you don't have the disease is 0.945. The good news is that this is a rare disease, striking only one in ten thousand people in your demographic.

   (a) What are the chances you have the disease?

   (b) Now assign a cost to the errors: deciding to seek treatment for the cancer when in fact you are healthy will cost you $1000 in unnecessary tests and the recovery therefrom. Deciding to forgo treatment when in fact you have the cancer will cost you and your family $1,000,000 in loss of life/income etc. Assume a correct decision (seek treatment if you have cancer, forgo treatment if you are healthy) has no cost, for simplicity.

   (c) What is the expected cost (i.e., "risk") assuming the cancer test comes out positive and you undergo treatment?

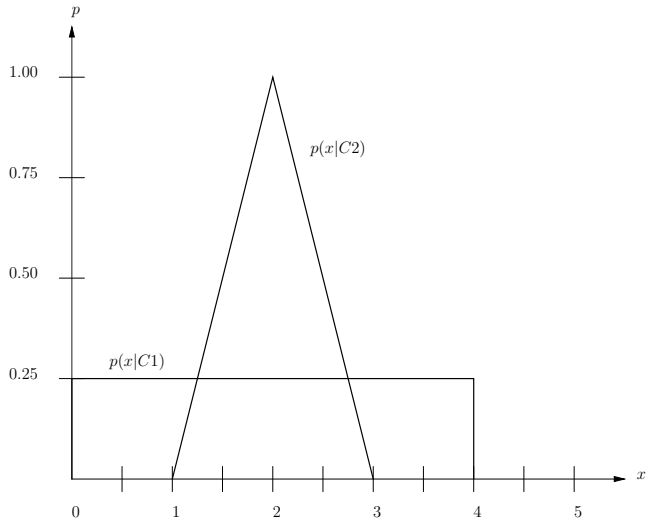
   (d) What should your decision be after a positive test? (Is this different from the answer to part (a)?)

   (e) What is the expected cost if the cancer test is negative and you do not undergo treatment?

2. We want to build a pattern classifier with a continuous attribute using Bayes' Theorem. The object to be classified has one feature, x in the range $0 \leq x \leq 4$. The conditional probability density functions for each class are, respectively,

$$p(x|C_1) = \begin{cases} \frac{1}{4} & \text{if } 0 \leq x < 4 \\ 0 & \text{otherwise} \end{cases}$$

$$p(x|C_2) = \begin{cases} x-1 & \text{if } 1 \leq x < 2 \\ 3-x & \text{if } 2 \leq x < 3 \\ 0 & \text{otherwise} \end{cases}$$



   (a) Assuming equal priors, $P(C_1) = P(C_2) = 0.5$, classify an object with the attribute value $x = 1.5$.

   (b) Assuming unequal priors, $P(C_1) = 0.75$, $P(C_2) = 0.25$, classify the object with the attribute value $x = 1.5$

   (c) Consider a decision function $\phi(x)$ of the form $\phi(x) = (|x-2|) - \alpha$ with one free parameter α in the range $0 \leq \alpha \leq 1$. You choose Class 2 for a given input x if and only if $\phi(x) < 0$, or equivalently $2 - \alpha < x < 2 + \alpha$, otherwise you choose class 1. What is the optimal decision boundary – that is, what is the value of α which minimizes the probability of misclassification? What is the resulting probability of misclassification with this optimal value for α? Assume equal priors. Hint: take advantage of the symmetry around $x = 2$.

   (d) Assume equal priors. Also assume there are penalties when choosing a class as follows:

	true class is 1	true class is 2
you classify object as Class 1	−5	+1
you classify object as Class 2	+3	−5

   What is the decision boundary (optimal value for α) that would minimize the expected penalty?

(e) Compute the estimated means and standard deviations for the conditional probability density for each class separately [use the unbiased estimates]. Plot corresponding normal (Gaussian) density functions using these estimated means and variances.

3. Consider a sample training set in one dimension with attribute values in the interval $[0, b]$, and 2 classes. Suppose your space of possible classifiers ("hypothesis space" $\mathcal{H}_k$) consists of "bucket" classifiers constructed by dividing the interval $[0, b]$ into k equal subintervals and assigning class 1 or 2 to each subinterval. Your only choices are the number k and the class assignment for each subinterval. The learning process is to determine which class to associate with each subinterval. Assume the number k of sub-intervals is given and fixed.

   (a) How many different classifiers are there in the Hypothesis space $\mathcal{H}_k$?

   (b) What is the VC dimension of $[0, b]$ with respect to $\mathcal{H}_k$?

4. Implement a program to fit two multivariate Gaussian distributions to the 2-class data in "training_data.txt" and classify the test data in "test_data.txt" by computing the log odds $log\frac{P(C_1|x)}{P(C_2|x)}$ with $P(C_1) = 0.6$ and $P(C_2) = 0.4$. Your program should display the quantities μ_1, μ_2, $\mathbf{S_1}$ and $\mathbf{S_2}$, the sample means and sample covariance matrices obtained for each class separately, assuming they are independent.

   You should then apply the classifier to the test set and show the resulting contigency table (confusion matrix):

number of C_1 samples classified as C_1	number of C_2 samples classified as C_1
number of C_1 samples classified as C_2	number of C_2 samples classified as C_2

   What is the resulting error rate on the test set?

Instructions

- **All solutions must be submitted electronically via Canvas.**
- **Things to submit: one PDF and one ZIP file:**

  1. hw1_sol.pdf: A document which contains the solutions to Problems 1, 2, 3, and 4, your name, student ID, email, any assumptions you are making, and any other necessary details. The solution to 4 should include the formulas for the parameters and their corresponding numerical values. The PDF file should include all the numerical values requested in Problem 4.
  2. For Problem 4 also submit a zip file containing the Matlab source file `classify.m` and any associated files needed to make this run. The function `classify` reads in the training and test data files, computes and returns the parameters estimated from the training set and the error rate on the test set. It should be a function which begins as follows:

```
function [mu1,mu2,S1,S2,ConfusionMatrix,ErrorRate]=classify(TrainingSet, TestSet);
% Solve Hw1 Q4, student name:.....
% Input Parameters:  TrainingSet, TestSet:  file names (strings), in "csv" format.
% .  .  .  more comments explaining the contents .  .  .
TrainingData=dlmread(TrainingSet);
class1=find(TrainingData(:,end)==1); % indices of observations in class 1.
class2=find(TrainingData(:,end)==2); % indices of observations in class 2.
TestData=dlmread(TestSet);
   .  .  .
```

- Do not include the data files downloaded from the class web site. Do not include the PDF file within the ZIP file. Rather, the PDF document should be submitted as a separate document.