University of Pittsburgh
CS 2750 Machine Learning
Professor Milos Hauskrecht

Midterm quiz
March 13, 2002

Midterm exam

Instructions

- This exam contains 5 problems. You have 80 minutes to earn 100 points.
- This exam is a closed-book exam.
- Write your solutions in the space provided. If you need more space use additional sheet. Do not put part of the answer to one problem to the space for another problem. Remember to write your name and problem number on any additional sheet you use.
- Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress.
- You will be graded not only on the correctness of your answer but also on the clarity with which you express it. Be neat.

Problem	Points	Grade
1	20	
2	20	
3	20	
4	20	
5	20	
Total	100	

Name:

Problem 1. (20 points)

A bank gave you a dataset that contains the information about its mortgage customers and asked you to predict whether a new customer applying for a mortgage is going to be a reliable or a risky customer in terms of the loan repayment.

Part a. (2 points) What type of a learning problem is this?

Part b. (2 points) Consider the binary class predictor that always picks the same class label (say 1). What is the worst case misclassification error for the predictor? Explain.

Part c. (4 points) Consider the binary class predictor that always picks the class label randomly with 50-50 % split. What is the worst case "true" mean misclassification error for this predictor? Explain.

Part d. (8 points) Assume your team came up with 3 different models (A,B, and C). You are interested in finding out how these models compare to the predictor in part c and whether they are different from it. You have decided to do statistical significance test. Explain briefly how would you construct the null hypothesis and what statistical test would you apply here. You may assume you kept a part of the original data aside for testing.

Part e. (4 points) Assume misclassification errors on the test set for models A,B and C are 25%, 26% and 80% respectively. After doing thorough statistical analysis you know that a difference of 2% in misclassification scores is sufficient (with very high probability) to rank the models. You need to come up with the best predictor but your team cannot agree which model to use and each of the three members supports different choice out of A,B and C. The arguments of the team members were summarized in the memo but you have lost it, the only thing you remember are the numbers. You need to make the final decision and come up with the best possible predictor. What is your recommendation and solution? Explain. Hint. Think in terms of the arguments members of the team could use to support their choices. Remember the better your solution the higher your premiums.

Problem 2. (20 points)

Show that the logistic regression model represents correctly the posterior of the class 1, $p(y = 1|x)$, whenever the class conditional distributions for the two classes, $p(x|y = 1)$ and $p(x|y = 0)$ are represented by distributions of the same type that come from the exponential family and that have the same scaling parameter.

Recall that the exponential family is defined as:

$$f(\mathbf{x}|\theta, \phi) = \exp\left(\frac{(\theta^T \mathbf{x} - b(\theta))}{a(\phi)} + c(\mathbf{x}, \phi)\right),$$

where θ represent the location parameters and ϕ the scaling parameter. $a(), b(), c()$ are functions that distinguish members of the exponential family.

Problem 3. (20 points)

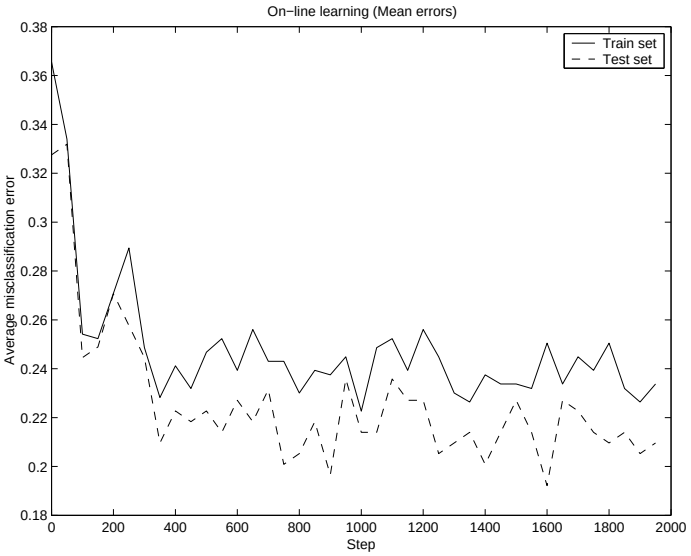
Answer true or false and justify your choice by explanation or example (counterexample). No credit will be given to the correct answer unless it is supported by the explanation.

a The gradient descent applied to a multi-layer neural network always finds the optimal set of weights.

b Non-linearities in the multi-layer neural network are modeled with weights.

c Backpropagation is an efficient method for computing the gradient of the error function for multi-layer neural models.

d Gradient descent is one of the many optimization methods one can use to learn the weights of the neural network.



e. Assume the graph above represents the evolution of the mean misclassification errors during learning with the online gradient descent and learning rate $\alpha = 2/\sqrt{i}$ on the pima dataset. Sequential, sample by sample, adjustement of weights is the only factor that can cause the observed mean error fluctuations.

f. Logistic regression unit can be modified to learn polynomial decision boundaries by using an appropriate set of feature functions.

g. Support vector machines without non-linear kernels can be applied only to linearly separable classification problems.

h. Data examples that are classified correctly by the support vector machine model do not belong to the support vector set.

i. Support vector machines model non-linear decision boundaries by expanding the input space with non-linear inputs and by finding a quadratic decision boundary in the new space.

Problem 4. (20 points) Assume that a random variable x follows a normal distribution. The normal distribution is defined as:

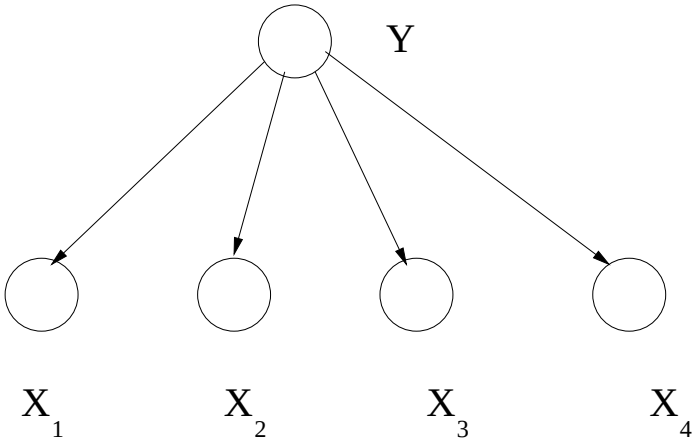
$$p(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(x-\mu)^2}{2\sigma^2}},$$

where μ, σ are the parameters. Assume we see N independent samples $x_1, x_2. \cdots x_N$ from the distribution. Derive the ML estimate of the parameter μ.

Please note that including the end result is not sufficient here. You need to show how this quantity is derived.

Problem 5 (20 points)

Naive Bayes model is a special case of a Bayesian belief network and it is used very often in the generative classification approach. Assume the Naive Bayes in the figure below.



Part a. (5 points) Assume that there are three different class values and that all attributes are binary. How many parameters are needed to define the Naive Bayes model.

Part b. (5 points) Assume we have a data set with 100 samples. There are 30% of examples of class 1, 50% of examples of class 2 and the rest are examples from class 3. All examples are independent. Assume that class priors follow Dirichlet distribution $Dir(x|\alpha_1, \alpha_2, \alpha_3)$ with $\alpha_1 = 30, \alpha_2 = 30, \alpha_3 = 40$. What is the ML estimate of class probabilities $p(Y)$? What is the Bayesian (expected value) estimate of probabilities $p(Y)$?

Part c. (5 points) The dataset consists of a set of examples and each example consists of an assignment of values to N attributes $X_1, X_2, \cdots, X_n$ and a class label (Y). Explain briefly how would you go about estimating the parameters of the conditional distribution $\theta_{i|Y=1} = \tilde{p}(X_i|Y=1)$ from the data.

Part d. (5 points) To use the Naive Bayes model as a classifier and predict class labels for new input data we need to devise appropriate discriminant functions. The most natural choice are discriminant functions based on class posterior probabilities, $p(Y=i|x,\Theta)$. Give the expression for computing such discriminant functions for the Naive Bayes model.