CS 475 Machine Learning: Midterm

Wednesday March 12, 2014

240 Points Total

Name (print): _____

JHED: _____

If you think a question is unclear or multiple answers are reasonable, please write a brief explanation of your answer, to be safe. Also, show your work if you want wrong answers to have a chance at some credit: it lets us see how much you understood.

I have neither given nor received any unauthorized aid on this exam. I understand that this exam must be taken without the aid of notes, textbooks, the use of the internet, or any other aid. The work contained herein is wholly my own. I understand that violation of these rules, including using an authorized aid or copying from another person, may result in my receiving a 0 on this exam and an additional 240 point reduction in my overall grade.

_____

Signature Date

Good luck!

True/False (75 points)

For each question, circle either True or False. If False, explain why.
5 points for True, 2 points for False, 3 points for explanation for False.

1) There is an efficient (polynomial time) algorithm that can learn the optimal (smallest) decision tree that accurately captures the training data given 1000 examples of training data with 200 binary features and binary labels.
True False
Explanation if False:

2) You are given two machine learning algorithms: A and B. Algorithm A yields a binary classifier with accuracy 90% on training data. Algorithm B yields a binary classifier with accuracy 70% on test data. You should use the binary classifier produced by Algorithm A.
True False
Explanation if False:

3) Finding the optimal solution for linear regression can only be found through the use of a numerical convex optimization algorithm, such as gradient descent.
True False
Explanation if False:

4) You are given a convex function and a fixed step size η. Running stochastic gradient descent with step size η you will eventually converge to the optimal solution.
True False
Explanation if False:

5) Using maximum a posteriori (MAP) estimation for linear regression will return the same solution as minimizing the squared error.
True False
Explanation if False:

6) Logistic regression cannot learn a non-linear decision boundary.
True False
Explanation if False:

7) The commonality between linear regression and logistic regression is that they both optimize the squared loss.
True False
Explanation if False:

8) The objective of logistic regression is to maximize the probability of the training data: $p(Y, X)$.
True False
Explanation if False:

9) The Perceptron algorithm will always make a finite number of errors when trained on an infinite stream of linearly separable data.
True False
Explanation if False:

10) Online algorithms are fundamentally mistake driven algorithms, which means that they only update the current hypothesis when an error has been made in prediction.
True False
Explanation if False:

11) For linearly separable data sets, a Perceptron and SVM will learn the same parameters.
True False
Explanation if False:

12) The Kernel trick is effective since it can be used in both the primal and dual formulation of the SVM.
True False
Explanation if False:

13) As we increase the parameter K in kNN, we increase bias in the predictions, which means that we will obtain worse training accuracy.
True False
Explanation if False:

14) A feed forward artificial neural network with a single hidden layer containing a finite number of variables is a universal approximator among continuous functions on compact subsets of $\mathcal{R}^n$.
True False
Explanation if False:

15) A maximum likelihood estimate is always inferior to a maximum a posteriori estimate (i.e., performs worse on training data) since maximum likelihood estimates underestimate the true variance.
True False
Explanation if False:

Short Answer (45 points)

16) (9 points) For each of the following problems, state which learning setting is the best fit: supervised learning, unsupervised learning, or reinforcement learning.

(a) Frank would like to build a computer game for checkers. One of the features of the game is that a user can play against the computer. Frank wants to ensure that the computer player is very good at checkers. He turns to machine learning to teach the computer how to play checkers.

_____

(b) Nancy works for a major search engine: Yahdu. She has access to millions of queries entered by users and which webpages the users visited after the search. Nancy wants Yahdu to return the webpages visited by users as the top answer for each query.

_____

(c) Deloris is working on a self-driving car. The car uses a camera to watch the road. Deloris is responsible for the steering component, which makes sure that the car stays on the road. To build this component, Deloris has access to thousands of images of the road taken from the car's camera and volunteers who have indicated where the road is located in each image.

_____

17) (10 points) State the name of the algorithm that optimizes the following loss functions.

(a) 0/1 Loss _____

(b) Hinge Loss _____

(c) Squared error loss _____

(d) Logistic loss _____

Which loss function is inappropriate for classification? Why?

18) (10 points) After years of searching, I found a biased coin! I called in my two friends, Mr. MLE and Ms. MAP. I then showed them my biased coin by flipping it 20 times and getting 15 heads and 5 tails. I claim it is biased.

(a) Does Mr. MLE (maximum likelihood estimation) think my coin is biased? Why?

(b) Does Ms. MAP (maximum a posteriori estimation) think my coin is biased? Why?

19) (8 points) Suppose we have an SVM using each of the following kernels. Which of these kernels does not have an equivalent closed-formed SVM primal problem? Why?

(a) RBF Kernel: $K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{||\mathbf{x}-\mathbf{x}'||_2^2}{2\sigma^2}\right)$

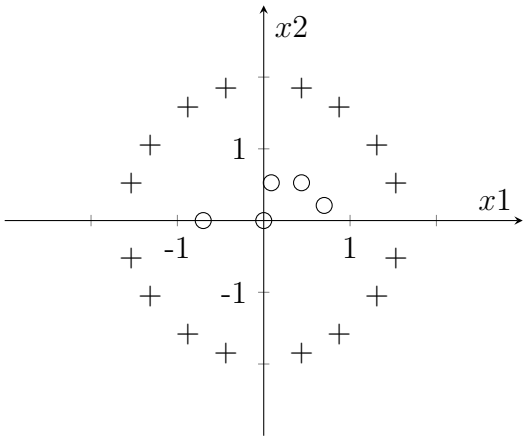
(b) Polynomial Kernel: $K(\mathbf{x}, \mathbf{x}') = (1 + \mathbf{x}^T\mathbf{x}')^d$

(c) Linear Kernel: $K(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T\mathbf{x}'$

20) (8 points) Suggest two modifications to the standard k-Nearest Neighbors algorithm that could alleviate over-fitting.

Long Answer (120 points)

21) (30 points) We decide to use logistic regression to fit the following training data.



Suppose a hypothesis class parameterized by $\mathbf{w}$ given by:

$$h(\mathbf{w}^T\mathbf{x}) = g(w_{(0)} + w_{(1)}x_{(1)} + w_{(2)}x_{(2)} + w_{(3)}x_{(1)}^2 + w_{(4)}x_{(2)}^2)$$

where $v_{(0)}$ is the 0th position of vector v, and

$$g(z) = \frac{1}{1+e^{-z}}$$

(a) Give values for the parameters $\mathbf{w}$ that correctly separates the given data.

(b) Write an equation in terms of vector $\mathbf{x}$ that represents this decision boundary.

(c) What is the purpose of $w_{(0)}$ in the hypothesis class?

(d) Suppose you are given a test data point of type 'o' at location $(0, 2)$. Would this point be correctly classified by your decision boundary?

22) (30 points) Consider the difference between the SVM primal problem its dual. Suppose we have N training examples ($\mathbf{x}$,y), and each $\mathbf{x}$ is an M dimensional feature vector $\mathbf{x} \in \mathcal{R}^M$ with a binary label $y \in \{-1, 1\}$.

(a) How many parameters are used in the primary problem? What are they?

(b) How many parameters are used in the dual problem? What are they?

(c) One of the formulations (primal or dual) could have more parameters than the other. If this is the case, does one formulation have a greater likelihood of over fitting the data than the other?

(d) When considering a formulation for a linear SVM, which formulation would be prefer to ensure computational efficiency in terms of N and M?

(e) Given a training dataset $\{\mathbf{X}, \mathbf{Y}\}$, write the prediction rule for both the primal and dual for a test example $\mathbf{x}$.

23) (30 points) Consider a generative model that takes example $\mathbf{x}$ containing M binary features and outputs a *vector* Y containing K elements, where each element k is binary label. For example, consider an image classification problem in which each element of Y is a different type of binary label, such as *"does this picture have a dog?"*, *"was this picture taken outside?"*, etc.

(a) Write the data likelihood for this model. How many parameters does it have in terms of the number of elements K and features M? Make no independence or conditional independence assumptions about the variables.

(b) Use the Naive Bayes assumption to assume that each element of $\mathbf{x}$ is conditionally independent given the entire vector Y. Write the data likelihood for this model. How many parameters are there in this model in terms of output elements K and input features M? Derive the maximum likelihood solution for these parameters.

(c) Assume that each element of Y is independent and that each element of $\mathbf{x}$ is conditionally independent given the entire vector Y. How many parameters are there in this model in terms of output elements K and input features M? Derive the maximum likelihood solution for these parameters.

24) (30 points) A brand new computer game called DetectiveX is taking the world by storm. In this game, you play a character that moves around in a world solving mysteries. You play the game by taking turns. On each turn, your character can choose between several actions:

1. Move (4 actions): left, right, up, down. These actions move you around the 2-d map in the world.
2. Talk: talk to a nearby character
3. Take: take an item that is next to you. If you take an item, you place it in your bag. At all times, you know what items are currently in your bag.
4. Look: look closely at something nearby. When you look, you can gain information about the world (i.e., you see an item.) You always have access to the global map of the world and it becomes filled in as you move and look.

The goal of the game is to maximize your score (points), which are displayed in the top right of the screen at all times. After every move, you either receive points or lose points, the number of which depends on the action you took on your turn.

You decide you'd like to build a machine learning algorithm to play this game and, with luck, enter the algorithm in the world wide DetectiveX competition (for big money!) Luckily, you've taken CS 475 so you have some good ideas about how to build an algorithm. Also, you have obtained a special simulator for this game which can be used to train your model. The simulator allows you to instantiate a state of the game and then take an action (and see the resulting points awarded.) Since you can reload the simulator to any state, you can go back and forth in the game to consider the consequences of each action.

Design a Perceptron algorithm that learns to play DetectiveX. The goal will be to take actions within the game that achieves the highest score. Make sure to include the following details:

1. The pseudo-code of the algorithm. Describe how you train the model and how you would use it during the competition.
2. The parameters you are learning and the update equation for the parameters.
3. The loss function you will use.
4. Examples of features your model will use. Be sure that you provide an example of at least one feature for every one of the 7 actions. The same feature can be used for multiple actions, and you need only give one example for each category. We aren't looking for an exhaustive list of features.

(Space to answer question)

(Scratch space)

(Scratch space)