CS178 Midterm Exam

Machine Learning & Data Mining: Winter 2015

Tuesday February 10th, 2014

Your name:

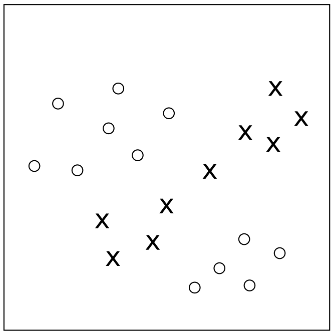
Your UCINetID (e.g., myname@uci.edu):

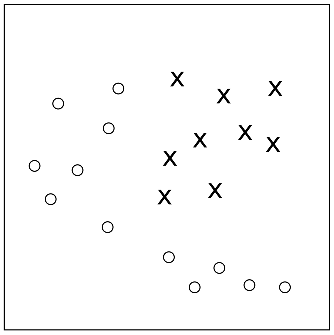
Your seat (row and number):

- Total time is 80 minutes. READ THE EXAM FIRST and organize your time; don't spend too long on any one problem.
- Please **write clearly** and **show all your work**.
- If you need clarification on a problem, please raise your hand and wait for the instructor to come over.
- Turn in any scratch paper with your exam

Problem 1: (8 points) Separability and Features

For each of the following examples of training data, **(1)** sketch a classification boundary that separates the data; **(2)** state whether or not the data are linearly separable, and if not, **(3)** give a set of features that would allow the data to be separated. (Your features do not need to be minimal, but should not contain any obviously unneeded features.)





Problem 2: (8 points) Under- and Over-fitting

Suppose that I am training a neural network classifier to recognize faces in images. Using cross-validation, we discover that my classifier appears to be overfitting the data. Give two ways I could improve my performance – be specific.

After following some of your advice, we now think that the resulting classifier is underfitting. Give two ways, **other than** reversing the methods you mentioned above, that we could improve performance; again, be specific.

Problem 3: (9 points) Bayes Classifiers and Naïve Bayes

Consider the table of measured data given at right. We will use the two observed features x_1, x_2 to predict the class y. In the case of a tie, we will prefer to predict class $y = 0$.

x_1	x_2	y
1	0	1
0	0	0
0	1	1
0	0	1
1	1	0
1	0	1
0	1	0
0	1	1

(a) Write down the probabilities necessary for a naïve Bayes classifier:

(b) Using your naïve Bayes model, what value of y is predicted given observation $(x_1, x_2) = (00)$?

(c) Using your naïve Bayes model, what is the probability $p(y = 1|x_1 = 0, x_2 = 1)$?

Problem 4: (10 points) Gradient Descent

Suppose that we have training data $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(m)}, y^{(m)})\}$ and we wish to predict y using a nonlinear regression model with two parameters:

$$\hat{y} = a + \log(b\, x_1)$$

We decide to train our model using gradient descent on the mean squared error (MSE).

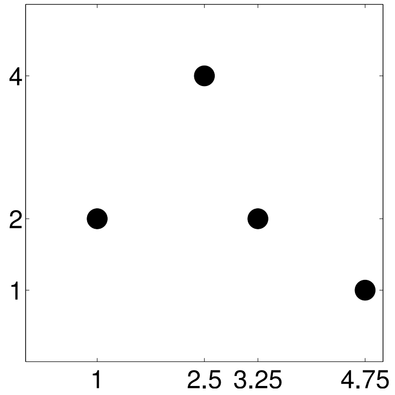
(a) Write down the expression for the MSE on our training set.

(b) Write down the gradient of the MSE.

(c) Give pseudocode for a (batch) gradient descent function `theta = train(X,Y)`, including all necessary elements for it to work.

Problem 5: (12 points) K-Nearest Neighbor Regression

Consider a regression problem for predicting the following data points, using the k-nearest neighbor regression algorithm to minimize mean squared error (MSE). In the case of ties, we will prefer to use the neighbor to the left (smaller x value). Note: if you prefer, you may leave an arithmetic expression, e.g., leave values as "$(.6)^2$".



(a) For $k = 1$, compute the training error on the provided data.

(b) For $k = 1$, compute the leave-one-out cross-validation error on the data.

(c) For $k = 3$, compute the training error on the provided data.

(d) For $k = 3$, compute the leave-one-out cross-validation error on the data.

Problem 6: (4 points) Multiple Choice

For the following questions, assume that we have m data points $y^{(i)}$, $x^{(i)}$, $i = 1 \ldots m$, each with n features, $x^{(i)} = [x_1^{(i)} \; \ldots \; x_n^{(i)}]$.

Circle one answer for each:

True or **false**: The predictions of a k-nearest neighbor classifier will not be affected if we pre-process the data to normalize the magnitude of each feature.

True or **false**: Increasing the regularization of a linear regression model will decrease the bias.

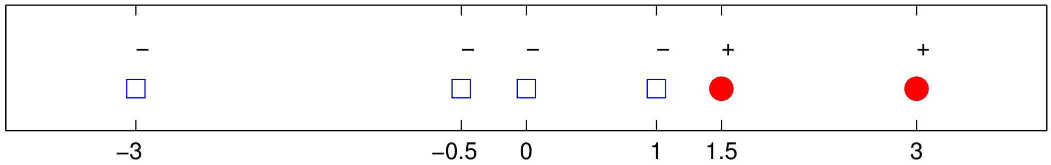
True or **false**: Linear regression can be solved using either matrix algebra or gradient descent.

True or **false**: With enough hidden nodes, a Neural Network can separate any data set.

Problem 7: (4 points) Short Answer

Give one advantage of stochastic gradient descent over batch gradient descent, **and** one advantage of batch gradient descent over stochastic.

Problem 8: (12 points) Support Vector Machines



Using the above data with one feature x (whose values are given below each data point) and a class variable $y \in \{-1, +1\}$, with filled circles indicating $y = +1$ and squares $y = -1$ (the sign is also shown above each data point for redundancy), answer the following:

(a) Sketch the solution (decision boundary) of a linear SVM on the data, and identify the support vectors.

(b) Give the solution parameters w and b, where the linear form is $wx + b$.

(c) Calculate the training error:

(d) Calculate the leave-one-out cross-validation error for these data: