CS178 Midterm Exam

Machine Learning & Data Mining: Winter 2017

Wednesday February 15th, 2017

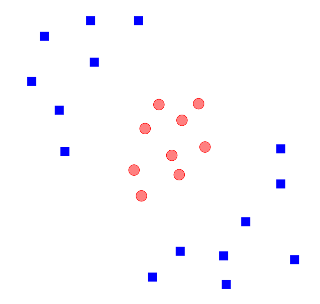
Your name: **Use Row/Seat:**

Your ID # and UCINetID:
(e.g., 123456789, myname@uci.edu)

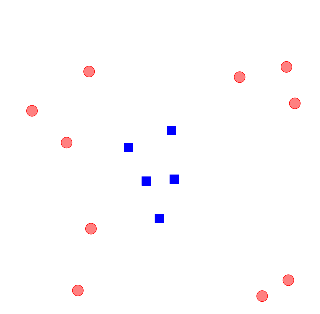
- Total time is 50 minutes. READ THE EXAM FIRST and organize your time; don't spend too long on any one problem.
- Please **write clearly** and **show all your work**.
- If you need clarification on a problem, please raise your hand and wait for the instructor or TA to come over.
- You may use **one** sheet of your own, handwritten notes for reference, as well as blank scratch paper and a (basic) calculator.
- Turn in your notes and any scratch paper with your exam.

Problem 1: (8 points) Separability and Features

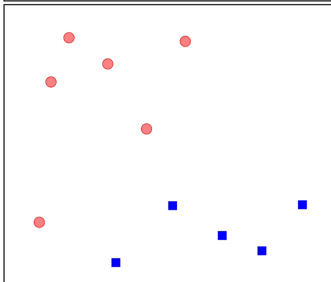
For each of the following examples of training data and classifiers, state whether there exists a set of parameters that can separate the data and justify your answer briefly (∼1 sentence).



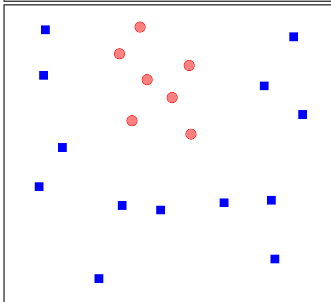
Perceptron with quadratic features, $[x_1 \;\; x_2 \;\; x_1x_2 \;\; x_1^2 \;\; x_2^2]$:



Perceptron with quadratic features, $[x_1 \;\; x_2 \;\; x_1x_2 \;\; x_1^2 \;\; x_2^2]$:



Depth-two decision tree:



Depth-two decision tree:

Problem 2: (10 points) Bayes Classifiers

Consider the table of measured data given at right. We will use the three observed features x_1, x_2, x_3 to predict the class y. In the case of a tie, we will prefer to predict class $y = 0$.

x_1	x_2	x_3	y
0	0	0	1
1	1	0	1
1	0	1	1
1	0	1	1
0	1	0	0
0	0	1	0
1	1	1	0
0	1	0	0

(a) Write down the probabilities used by a naïve Bayes classifier:

$p(y = 0)$: $p(y = 1)$:

$p(x_1 = 1|y = 0)$: $p(x_1 = 1|y = 1)$:

$p(x_1 = 0|y = 0)$: $p(x_1 = 0|y = 1)$:

$p(x_2 = 1|y = 0)$: $p(x_2 = 1|y = 1)$:

$p(x_2 = 0|y = 0)$: $p(x_2 = 0|y = 1)$:

$p(x_3 = 1|y = 0)$: $p(x_3 = 1|y = 1)$:

$p(x_3 = 0|y = 0)$: $p(x_3 = 0|y = 1)$:

(b) Using your naïve Bayes model, compute:

$p(y = 0|x_1 = 1, x_2 = 0, x_3 = 0)$: $p(y = 1|x_1 = 1, x_2 = 0, x_3 = 0)$:

(c) Give an example of a problem setting in which we might want to use naïve Bayes even though the model assumptions may not be correct.

Problem 3: (10 points) Decision Trees

Consider the table of measured data given at right. (Note that some data points are repeated.) We will use a decision tree to predict the outcome y using the three features, $x_1, \ldots, x_3$. In the case of ties, we prefer to use the feature with the smaller index (x_1 over x_2, etc.) and prefer to predict class 1 over class 0. You may find the following values useful (although you may also leave logs unexpanded):

$\log_2(1) = 0 \quad \log_2(2) = 1 \quad \log_2(3) = 1.59 \quad \log_2(4) = 2$
$\log_2(5) = 2.32 \quad \log_2(6) = 2.59 \quad \log_2(7) = 2.81 \quad \log_2(8) = 3$

x_1	x_2	x_3	y
1	1	1	1
1	0	0	1
1	0	0	1
0	0	0	0
0	1	1	0
1	0	1	0

(a) What is the entropy of y?

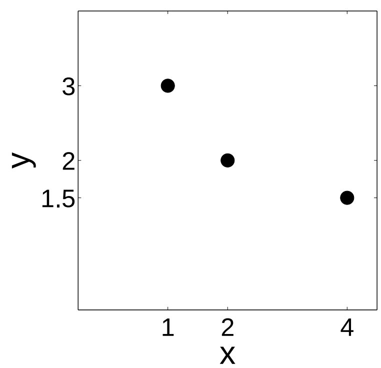
(b) Which variable would you split first? Justify your answer.

(c) What is the information gain of the variable you selected in part (b)?

(d) Draw the rest of the decision tree learned on these data.

Problem 4: (12 points) Nearest Neighbor Regression

Consider the data points shown at right, for a regression problem to predict y given a scalar feature x. We wish to use a k-nearest neighbor learner to minimize the mean squared error (MSE) of our predictions.



(a) Compute the **training** error (MSE) of a 1-nearest neighbor predictor.

(b) Compute the **leave-one-out** cross-validation error (MSE) of a 1-nearest neighbor predictor.

(c) Compute the **training** error (MSE) of a 2-nearest neighbor predictor.

(d) Compute the **leave-one-out** cross-validation error (MSE) of a 2-nearest neighbor predictor.

(e) Based on just these data & results, which value of k would you choose, and why?

Problem 5: (10 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, $x^{(i)} = [x_1^{(i)} \; \ldots \; x_n^{(i)}]$.

True or **false**: The predictions of a decision tree classifier will not be affected if we pre-process the data to normalize the magnitude of each feature (e.g., rescale each feature to the range [-1,1]).

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

True or **false**: Using backpropagation to train a neural network will avoid getting stuck in local optima.

True or **false**: With sufficient depth, a decision tree can approximate any boolean function.

True or **false**: if two models have the same VC dimension, they are equally likely to overfit the data.

True or **false**: Given sufficently many data m, the 1-nearest neighbor classifier error rate approaches the Bayes optimal error rate.

True or **false**: Stochastic gradient descent is often preferred over batch when the number of data points m is very large.

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

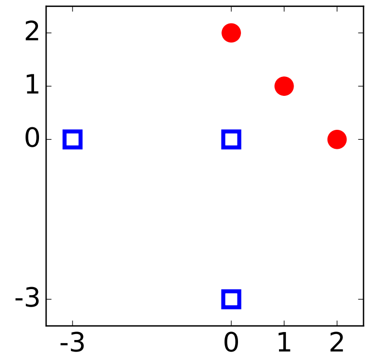
A quadratic program (such as SVM learning) is characterized by a quadratic objective function and a set of **convex** **linear** **quadratic** **soft** constraints.

Suppose that we are training a linear regressor (perceptron). Adding extra features that are not related to the target (e.g., randomly generated values) before training will typically **increase** **not change** **decrease** our training error.

Problem 6: (10 points) Support Vector Machines

Suppose we are learning a linear support vector machine with two real-valued features x_1, x_2 and binary target $y \in \{-1, +1\}$. We observe training data (pictured at right):

x_1	x_2	y
-3	0	-1
0	-3	-1
0	0	-1
2	0	+1
0	2	+1
1	1	+1



Our linear classifier takes the form

$$f(x; w_1, w_2, b) = \text{sign}(w_1 x_1 + w_2 x_2 + b).$$

(a) Sketch the decision boundary of the trained SVM, and identify the support vectors

(b) Give the parameter values of the trained SVM.

(c) What is the **training** error rate on these data?

(d) What is the **leave-one-out** cross-validation error rate on these data?