CS273a Midterm Exam

Introduction to Machine Learning: Fall 2012

Tuesday December 11th, 2012

Your name:

Name of the person in front of you (if any):

Name of the person to your right (if any):

- Total time is 1:15. 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: VC Dimension

Argue by example / counterexample what is the VC dimension of each of the following classifiers.

(a) A perceptron on two *binary* features

(b) A decision stump on two *binary* features

Problem 2: Decision Trees

We plan to use a decision tree to predict an outcome y using four features, $x_1, \ldots, x_3$. We observe six training patterns, each of which we represent as $[x_1, x_2, x_3]$ (so, "010" means $x_1 = 0$, $x_2 = 1$, $x_3 = 0$). We observe the training data,

$y = 0:$ [100], [111], [001]

$y = 1:$ [110], [110], [011]

You may find the following values useful (although you may also leave logs unexpanded):

$\log_2(1) = 0 \quad \log_2(2) = 1 \quad \log_2(3) = 1.59 \quad \log_2(4) = 2$

$\log_2(5) = 2.32 \quad \log_2(6) = 2.59 \quad \log_2(7) = 2.81 \quad \log_2(8) = 3$

(a) What is the entropy of y?

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

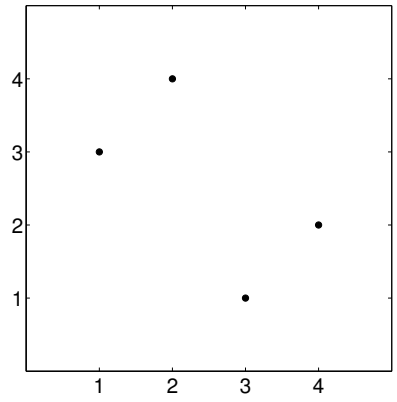
(c) What is the information gain of the variable you selected in part (b)?

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

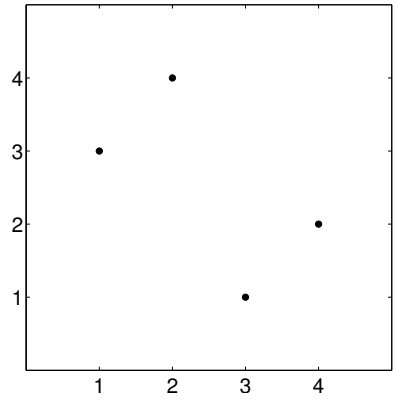
Problem 3: Gradient Boosting

Consider the following data set consisting of four points; for convenience, the data are repeated in each part.

(a) Compute the best single decision stump regressor function, to minimize mean squared error.



(b) Now, we wish to create a gradient boosted ensemble of decision stumps to minimize MSE. Starting from the decision stump learned in 9a), and using a learning rate of 1, what is the next predictor? Show your work.



(c) Is the resulting ensemble the best possible ensemble of two decision stumps? If yes, why? If not, give a better ensemble.

Problem 4: Naïve Bayes

We plan to use a naïve Bayes classifier to predict an outcome y using four features, $x_1, \ldots, x_3$. We observe five training patterns, each of which we represent as $[x_1, x_2, x_3]$ (so, "010" means $x_1 = 0$, $x_2 = 1$, $x_3 = 0$). We observe the training data,
$y = 0:$ [000], [111]
$y = 1:$ [100], [010], [001]

(a) Compute (& show) all the necessary probabilities for a naïve Bayes model.

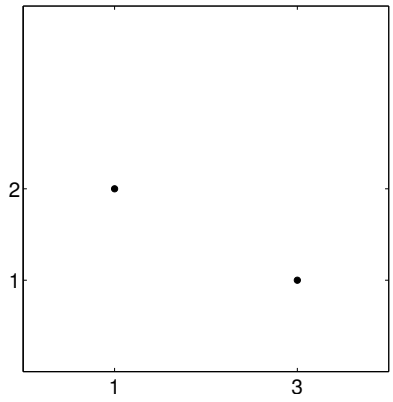
(b) Suppose you observe $x = [110]$. What class (value of y) would you predict? Show your work.

(c) Suppose you observe $x = [100]$. Compute the posterior probability, $p(y|x = 100)$.

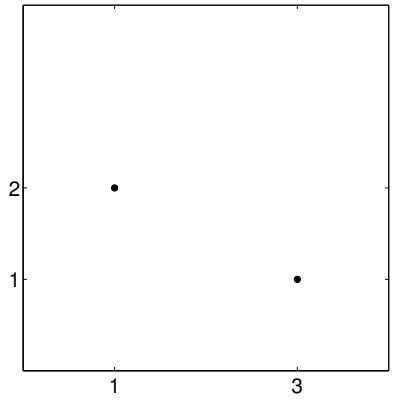
Problem 5: Bagging

Consider the data set, consisting of two data points, given in each part.

(a) Draw the regression function (predicted values for all x) using a nearest-neighbor regressor. Label any necessary values on your graph.



(b) Suppose that we create a very large ensemble of *bagged* nearest-neighbor regressors, using data set draws of size two. Compute the regression function of the complete ensemble, again labeling any necessary values.

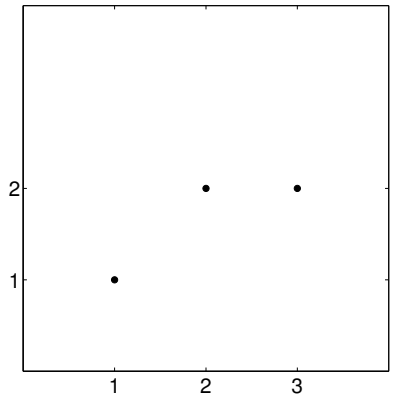


(c) Is the model in (b) simpler or more complex than the model in (a)? Why?

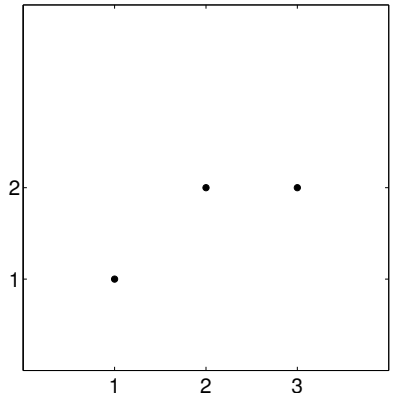
Problem 6: Cross-validation

Consider the following data points, copied in each part. We wish to perform linear regression to minimize mean squared error.

(a) Compute the leave-one-out cross-validation error of a zero-order (constant) predictor.



(b) Compute the leave-one-out cross-validation error of a first-order (linear) predictor.



Problem 7: Latent space models

Suppose that, as in HW5, we wish to model a collection of text documents using a latent space model. For interpretability, we would like our latent representation to be non-negative. As one solution, we use an exponential transform to ensure positive values, giving the model

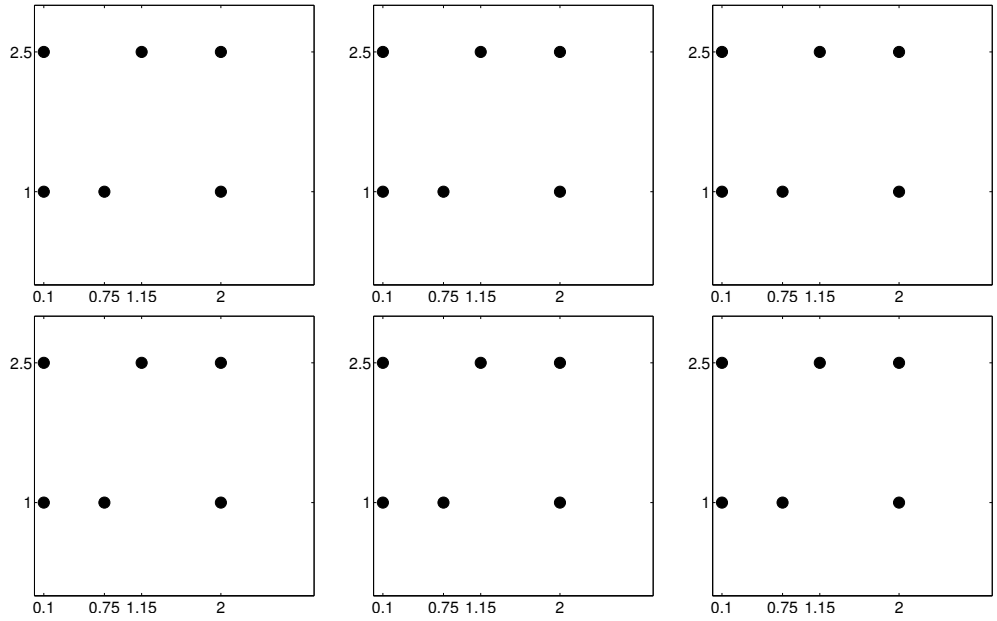
$$x_j^{(i)} \approx \sum_k \exp(U_{ik}) \exp(V_{kj})$$

Give a stochastic gradient descent algorithm to learn this model, minimizing the mean squared error in the predicted values. Include all necessary details for the implementation.

Problem 8: Clustering

Consider the two-dimensional data points plotted in each panel. In this problem, we will cluster these data using two different algorithms, where each panel is used to show an iteration or step of the algorithm.

(a) Execute the hierarchical agglomerative clustering (linkage) algorithm on these data points, using "single linkage" for the cluster scores. Stop when converged, or after 6 steps, whichever is first. Show each step separately in a panel.



(b) Now execute hierarchical agglomerative clustering on the data points, but use "complete linkage" for the cluster scores. Stop when converged, or after 6 steps, whichever is first. Show each step separately in a panel.