CSE 517a (Machine Learning): Midterm

Wednesday, March 3rd, 2010

.

First Name	
Last Name	
cec login	
Question 1	/
Question 2	/ 16
Question 3	/ 20
Question 4	/ 22
Question 5	/ 20
TOTAL	/

This page was unintentionally left blank. (Seriously, how do you get rid of it?)

1. Short questions

1 Point for correct answer, +2 Points for explanation (if applicable)
In case of "false" provide a small justification of your answer.

T/F A classifier trained on less training data is less likely to overfit.

T/F As the amount of training data increases, the training error goes down.

T/F The best machine learning algorithms make no assumptions.

T/F The regularization constant should be chosen on the test data set.

T/F In AdaBoost, weights of the misclassified examples go up by the same multiplicative factor per iteration.

T/F The training error of Adaboost decreases exponentially.

T/F Linear models are too restricted to overfit.

T/F Boosting is used to reduce the bias of low-variance classifiers.

T/F Minimizing the log-loss for linear classifiers is equivalent to choosing the weight vector w with the Maximum Likelihood Estimate for the conditional probability $P(y|x;w) = \frac{1}{1+e^{-w^\top xy}}$ and $y \in \{+1, -1\}$.

T/F The zero-one loss can be minimized with the gradient descent algorithm.

2. Nearest Neighbor Classification

[16 Points]

1. How does the *bias* of $k-NN$ change as k *increases*? [2]

2. How does the *variance* of $k-NN$ change as k *decreases*? [2]

3. Joe implemented a function $predictions = knnclassify(xTr, xTe, k)$. He wants to know the leave-one-out training error of $1-$nearest neighbors and runs $predictions = knnclassify(xTr, xTr, 1)$. He is amazed and full of joy about how well his algorithm performs. What could have gone wrong? How would you fix it? [4]

4. How does the curse of dimensionality affect kNN classification of uniformly sampled data points within a $[0,1]^d$ space? [4]

5. Why does kNN still work on high-dimensional images of hand-written digits despite the curse of dimensionality? [2]

6. When would you use $KD-Trees$ over $Ball-trees$ and vice versa? [2]

3. Decision Trees

[20 Points]

1. In the id3tree-algorithm what are the criteria to decide when to stop recursing and create a leaf? [4]

2. Why are decision trees called myopic? [2]

3. If you build a single decision tree, how can you reduce the chance of overfitting (no need to state any algorithm)? [3]

4. Assume you are given the following data set:

person	side-kick	ears	smokes	height	class
Batman	n	y	n	180	good
Robin	y	n	n	176	good
Alfred	n	n	n	185	good
Penguin	n	n	y	140	evil
Catwoman	n	y	n	170	evil
Joker	n	n	n	179	evil

5. (a) What is the expected class entropy after a binary split at $height \leq 160$? (No need to compute the actual number, just write down the expression you could type into a calculator.) [3]

   (b) What is the first feature that the id3 algorithm would pick to split on (using expected entropy)? (No justification required.) [2]

   (c) Draw a full decision tree that the id3 algorithm would learn. [4]

   (d) Given the following validation data, what would your validation error be (in terms of misclassified examples)? [2]

person	side-kick	ears	smokes	height	class
Riddler	n	n	n	170	evil
Batgirl	y	n	n	150	good
Timo	n	n	n	60	good

4. Bias, Variance

[22 Points]

1. Write down the definitions of (squared) bias, variance and noise. [6]

2. Describe a method to detect if your classifier suffers from too much bias or too much variance. [6]

3. Describe *bagging*, and explain what effect it has on bias / variance. [4]

4. The Riddler uses linear regression (ordinary least squares) on a low dimensional data set. He is trying to reduce the classifier's bias with *boosting*. It is easy to change OLS to incorporate weights – but strangely he observes that his training error does not decrease even after many iterations. Write down the expression of the final boosted classifier after T iterations. What shape does the boosted decision boundary have? Can you explain his findings? [6]

5. Linear Models

[20 Points]

1. Given a linear classifier $h_{w,b}(x) = sign(w^\top x + b)$. Show how a change of variable makes the explicit treatment of b unnecessary. Write down a new definition h_w. [2]

2. Define the separating hyperplane $\mathcal{H}$ in terms of h_w. [2]

3. How would you change your function for regression? [2]

4. Batgirl has n data points $(x_1, y_1), \ldots, (x_n, y_n)$ with real valued labels $y_i \in \mathcal{R}$. She uses the following loss function to learn w:

$$w = \arg\min_w \sum_{i=1}^{n} (w^\top x_i - y_i)^2 + \lambda w^\top w.$$

   (a) Do you recognize the loss function (ignoring the $\lambda w^\top w$ for now)? For what data sets might this choice of loss function become problematic? [3]

   (b) Why would she add the term $\lambda w^\top w$? For what kind of data could this term for $\lambda > 0$ improve the test error? [3]

   (c) Derive a closed-form solution for the vector w (You can use: $\frac{\partial trace(w^\top A)}{\partial w} = A$, $\frac{\partial trace(w^\top Bw)}{\partial w} = Bw + B^\top w$ and $w^\top w = w^\top Iw$ where I is the identity matrix). [8]

Space for calculations or doodling.