6.867 Machine learning

Mid-term exam

October 8, 2003

(2 points) Your name and MIT ID:

J. J. Doe, MIT ID# 000000000

Problem 1

In this problem we use sequential active learning to estimate a linear model

$$y = w_1 x + w_0 + \epsilon$$

where the input space (x values) are restricted to be within $[-1, 1]$. The noise term ϵ is assumed to be a zero mean Gaussian with an unknown variance σ^2. Recall that our sequential active learning method selects input points with the highest variance in the predicted outputs. Figure 1 below illustrates what outputs would be returned for each query (the outputs are not available unless specifically queried).

We start the learning algorithm by querying outputs at two input points, $x = -1$ and $x = 1$, and let the sequential active learning algorithm select the remaining query points.

1. **(4 points)** Give the next two inputs that the sequential active learning method would pick. Explain why.

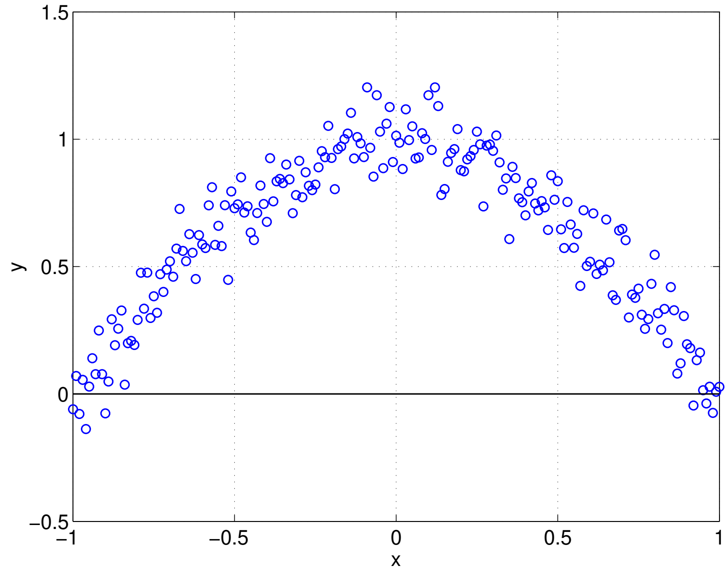

Figure 1: Samples from the underlying relation between the inputs x and outputs y. The outputs are not available to the learning algorithm unless specifically queried.

A linear function is constrained the most by the end points (the variance is always highest at the end points) and thus the next two points are -1 and 1 (order not determined).
Alternatively, you can look more closely at

$$Var\{\hat{y}(x)\} = \begin{bmatrix} 1 \\ x \end{bmatrix}^T (\mathbf{X}^T\mathbf{X})^{-1} \begin{bmatrix} 1 \\ x \end{bmatrix}$$

where, initially, $\mathbf{X} = [1, -1; 1, 1]$ *(MATLAB notation). Since* $(\mathbf{X}^T\mathbf{X})^{-1}$ *is positive (semi-)definite regardless of the points included in* $\mathbf{X}$*, the variance is a convex-up parabola and takes its largest value at (one of) the end points.*

2. **(4 points)** In the figure 1 above, draw (approximately) the linear relation between the inputs and outputs that the active learning method would find after a large number of iterations.

   Since only the points -1 and 1 are queried, the linear function is determined by the outputs at these points. Hence the line in the figure.

3. **(6 points)** Would the result be any different if we started with query points $x = 0$

and $x = 1$ and let the sequential active learning algorithm select the remaining query points? Explain why or why not.

The same argument applies here and we would start querying at x=-1 and x=1 in an alternating fashion. Note that while the variance is a function of inputs only, the resulting linear function surely depends on the outputs as well. Initially, therefore, there will be a difference due to the single non-extremal query point x=0 but as more end points are selected, the difference will vanish

Problem 2

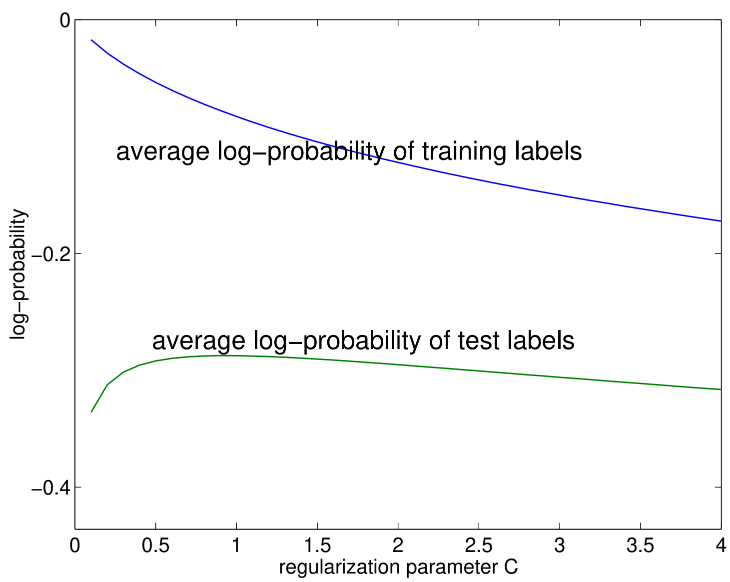

Figure 2: Log-probability of labels as a function of regularization parameter C

Here we use a logistic regression model to solve a classification problem. In Figure 2, we have plotted the mean log-probability of labels in the training and test sets after having trained the classifier with quadratic regularization penalty and different values of the regularization parameter C.

1. **(T/F – 2 points)** In training a logistic regression model by maximizing the likelihood of the labels given the inputs we have multiple locally optimal solutions.

F

The log-probability of labels given examples implied by the logistic regression model is a concave (convex down) function with respect to the weights. The (only) locally optimal solution is also globally optimal (if there are only a few examples, there might be a "flat" region with many equally optimal solutions, as in, e.g., question 3.4).

2. **(T/F – 2 points)** A stochastic gradient algorithm for training logistic regression models with a fixed learning rate will find the optimal setting of the weights exactly.

F

A fixed learning rate means that we are always taking a finite step towards improving the log-probability of any single training example in the update equation. Unless the examples are somehow "aligned", we will continue jumping from side to side of the optimal solution, and will not be able to get arbitrarily close to it. The learning rate has to approach to zero in the course of the updates for the weights to converge.

3. **(T/F – 2 points)** The average log-probability of training labels as in Figure 2 can never increase as we increase C.

T

Stronger regularization means more constraints on the solution and thus the (average) log-probability of the training examples can only get worse.

4. **(4 points)** Explain why in Figure 2 the test log-probability of labels decreases for large values of C.

As C increases, we give more weight to constraining the predictor, and thus give less flexibility to fitting the training set. The increased regularization guarantees that the test performance gets closer to the training performance, but as we over-constrain our allowed predictors, we are not able to fit the training set at all, and although the test performance is now very close to the training performance, both are low.

5. **(T/F – 2 points)** The log-probability of labels in the test set would decrease for large values of C even if we had a large number of training examples.

T

The above argument still holds, but the value of C for which we will observe such a decrease will scale up with the number of examples.

6. **(T/F – 2 points)** Adding a quadratic regularization penalty for the parameters when estimating a logistic regression model ensures that some of the parameters (weights associated with the components of the input vectors) vanish. F

   A regularization penalty for feature selection must have non-zero derivative at zero. Otherwise, the regularization has no effect at zero, and weight will tend to be slightly non-zero, even when this does not improve the log-probabilities by much.

Problem 3

Consider a training set consisting of the following eight examples:

Examples labeled "0"	Examples labeled "1"
3,3,0	2,2,0
3,3,1	1,1,1
3,3,0	1,1,0
2,2,1	1,1,1

The questions below pertain to various feature selection methods that we could use with the logistic regression model.

1. **(2 points)** What is the mutual information between the third feature and the target label based on the training set? 0

   The third feature has the same distribution conditioned on both labels, hence it is statistically independent of the target label and the mutual information between them is zero.

2. **(2 points)** Which feature(s) would a filter feature selection method choose? You can assume here that the mutual information criterion is evaluated between a single feature and the label. 1,2

   Both of these features have high, and equal, mutual information with the target label.

3. **(2 points)** Which two feature(s) would a greedy wrapper process choose? 1,3 or 2,3

First, one of the first two features will be chosen, as this allows us to easily classify correctly six of the eight features. Once one of these features is chosen, the other one does not help us. However, the third feature can help us discriminate between (2,2,1) and (2,2,0)— including it will enable us to classify all examples correctly, and will thus be included.

4. **(4 points)** Which features would a regularization approach with a 1-norm penalty $\sum_{i=1}^{3} |w_i|$ choose? Explain briefly.

Since the first two features are identical, shifting weight between them has no effect on the log-probabilities. The 1-norm penalty is oblivious to shifting weight in this way. Thus, it will not favor setting the weight of one of the first two features to zero, and such a setting will be one of many equally optimal solutions. Most optimal solutions would thus include both features, but depending on how we optimize the objective function, we might choose the rare solutions in which one of the weights is set to zero.

Including the third features improve the average log-probabilities, but it also increases the penalty. Whether or not it will be included depends on the regularization parameter C that controls the relative importance of the average log-probabilities and the regularization penalty. To check what values of C will allow inclusion of the third feature, we can check the derivative of the objective function at $w_3 = 0$: The derivative of the log probability of the sample (2,2,1) will be $g'(0)x_3 = \frac{1}{2}$. The derivatives for samples with $x_3 = 0$ will be zero, and the derivatives for the other four samples will cancel each other. Thus the total derivative of the log-probabilities term will be $\frac{1}{2}$. This will be more significant than the regularization only if $C < \frac{1}{2}$.

Problem 4

1. **(6 points)** Figure 3 shows the first decision stump that the AdaBoost algorithm finds (starting with the uniform weights over the training examples). We claim that the weights associated with the training examples after including this decision stump will be $[1/8, 1/8, 1/8, 5/8]$ (the weights here are enumerated as in the figure). Are these weights correct, why or why not?

   Do not provide an explicit calculation of the weights.

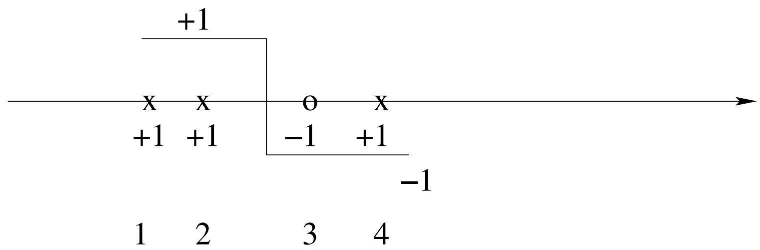

Figure 3: The first decision stump that the boosting algorithm finds.

The boosting algorithm adjusts the weights so that the current hypothesis is at chance level relative to the new weights. Since there is only one miss-classified example, it has to receive weight = 1/2. So the weights are not correct.

2. **(T/F – 2 points)** The votes that AdaBoost algorithm assigns to the component classifiers are optimal in the sense that they ensure larger "margins" in the training set (higher majority predictions) than any other setting of the votes.

   F

   The votes in the boosting algorithm are optimized sequentially and never "reoptimized" after all the hypotheses have been generated. These votes cannot therefore be optimal in the sense of achieving the largest majority predictions for the training examples.

3. **(T/F – 2 points)** In the boosting iterations, the training error of each new decision stump and the training error of the combined classifier vary roughly in concert

   F

   While the training error of the combined classifier typically decreases as a function of boosting iterations, the error of the individual decision stumps typically increases since the example weights become concentrated at the most difficult examples.

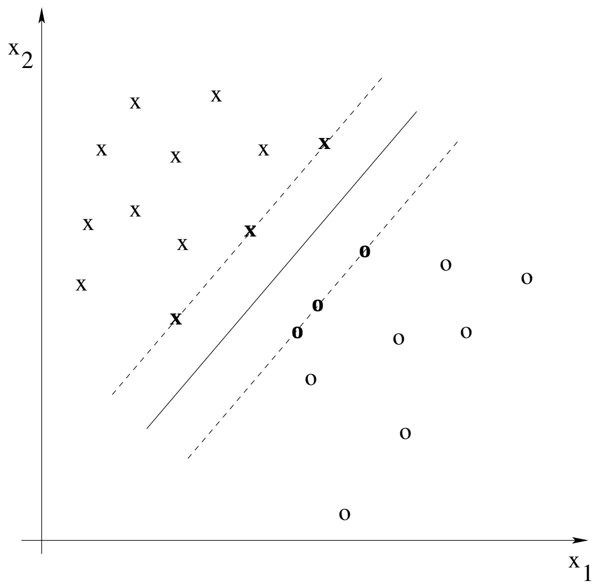

Figure 4: Training set, maximum margin linear separator, and the support vectors (in bold).

Problem 5

1. **(4 points)** What is the leave-one-out cross-validation error estimate for maximum margin separation in figure 4? (we are asking for a number) $\boxed{0}$

   Based on the figure we can see that removing any single point would not chance the resulting maximum margin separator. Since all the points are initially classified correctly, the leave-one-out error is zero.

2. **(T/F – 2 points)** We would expect the support vectors to remain the same in general as we move from a linear kernel to higher order polynomial kernels. $\boxed{\text{F}}$

   There are no guarantees that the support vectors remain the same. The feature vectors corresponding to polynomial kernels are non-linear functions of the original input vectors and thus the support points for maximum margin separation in the feature space can be quite different.

3. **(T/F – 2 points)** Structural risk minimization is guaranteed to find the model (among those considered) with the lowest expected loss $\boxed{\text{F}}$

   We are guaranteed to find only the model with the lowest upper bound on the expected loss.

4. **(6 points)** What is the VC-dimension of a mixture of two Gaussians model in the plane with equal covariance matrices? Why?

A mixture of two Gaussians with equal covariance matrices has a linear decision boundary. Linear separators in the plane have VC-dim exactly 3.

Problem 6

Using a set of 100 labeled training examples (two classes), we train the following models:

GaussI A Gaussian mixture model (one Gaussian per class), where the covariance matrices are both set to I (identity matrix).

GaussX A Gaussian mixture model (one Gaussian per class) without any restrictions on the covariance matrices.

LinLog A logistic regression model with linear features.

QuadLog A logistic regression model, using all linear and quadratic features.

1. **(6 points)** After training, we measure for each model *the average log probability of labels given examples in the training set.* Specify all the equalities or inequalities that must *always* hold between the models relative to this performance measure. We are looking for statements like "model 1 $\leq$ model 2" or "model 1 $=$ model 2". If no such statement holds, write "none".

GaussI $\leq$ ***LinLog*** *(both have logistic postiriors, and* **LinLog** *is the logistic model maximizing the average log probabilities)*
GaussX $\leq$ ***QuadLog*** *(both have logistic postiriors with quadratic features, and* **QuadLog** *is the model of this class maximizing the average log probabilities)*
LinLog $\leq$ ***QuadLog*** *(logistic regression models with linear features are a subclass of logistic regression models with quadratic functions— the maximum from the superclass is at least as high as the maximum from the subclass)*
GaussI $\leq$ ***QuadLog*** *(follows from above inequalities)*
*(***GaussX** *will have higher average log* joint *probabilities of examples and labels, then will* **GaussI***. But have higher average log joint probabilities does not necessarily translate to higher average log conditional probabilities)*

2. **(4 points)** Which equalities and inequalities must *always* hold if we instead use the *mean classification error in the training set* as the performance measure? Again use the format "model 1 $\leq$ model 2" or "model 1 = model 2". Write "none" if no such statement holds.

None. Having higher average log conditional probabilities, or average log joint probabilities, does not necessarily translate to higher or lower classification error. Counterexamples can be constructed for all pairs in both directions.

Although there is no inequalities which is always correct, it is commonly the case that ***GaussX*** $\leq$ ***GaussI*** *and that* ***QuadLog*** $\leq$ ***LinLog****. Partial credit of up to two points was awarded for these inequalities.*