6.867 Machine learning

Mid-term exam

October 28, 2001

(2 points) Your name and MIT ID:

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.

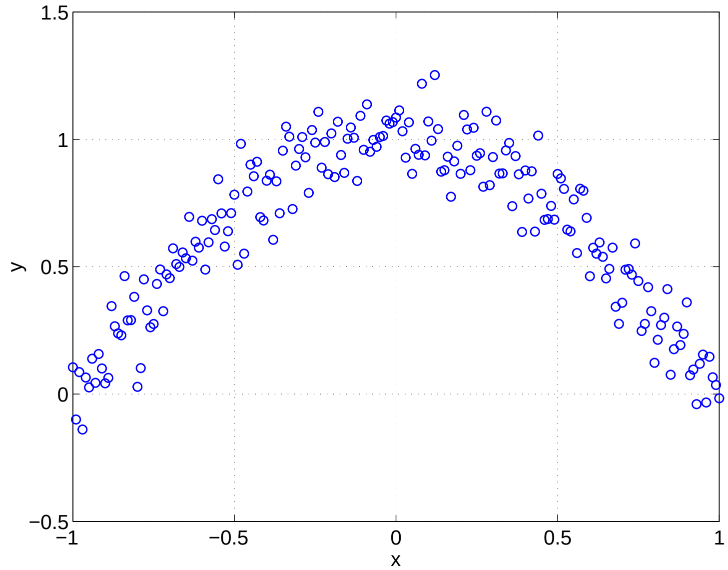

Figure 1: Samples from the underlying relation between the inputs x and outputs y. The outputs are not available to the learning algorithm unless specifically queried.

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.

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.

Problem 2

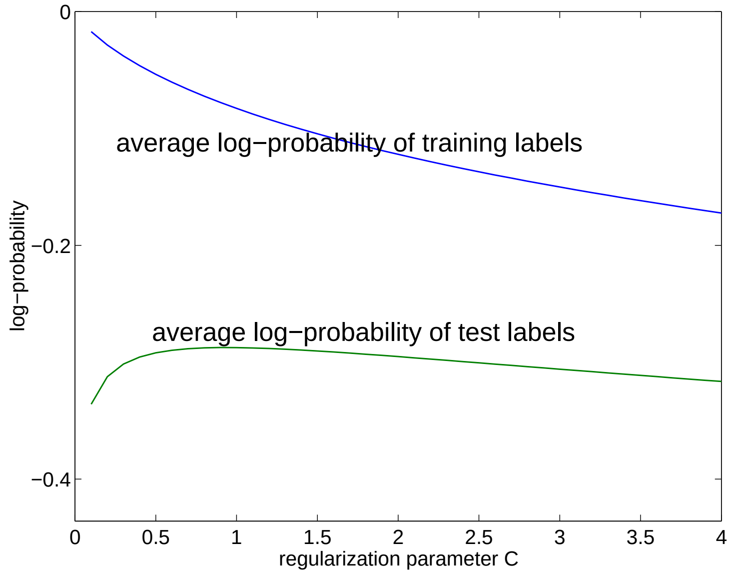

Figure 2: Log-probability of labels as a function of regularization parameter C

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. ☐

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. ☐

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

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

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.

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.

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?

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.

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

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

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.

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.

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 .

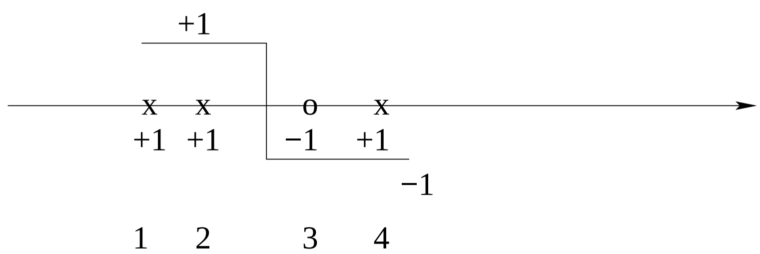

Figure 3: The first decision stump that the boosting algorithm finds.

Problem 5

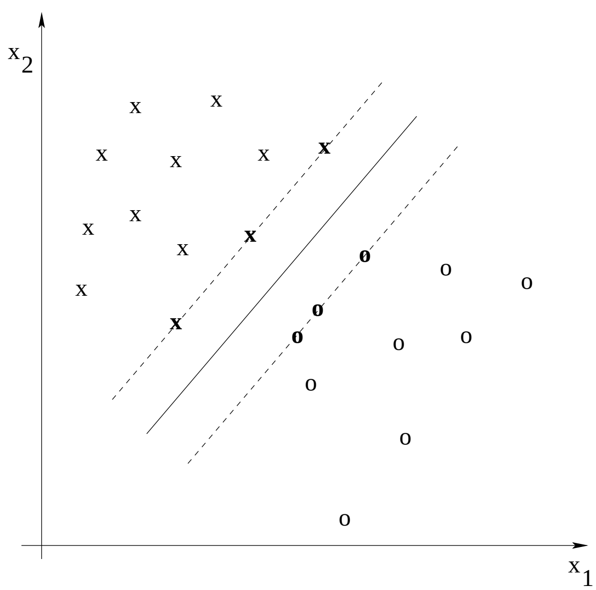

Figure 4: Training set, maximum margin linear separator, and the support vectors (in bold).

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)

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.

3. **(T/F – 2 points)** Structural risk minimization is guaranteed to find the model (among those considered) with the lowest 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?

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".

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.

Another set of figures

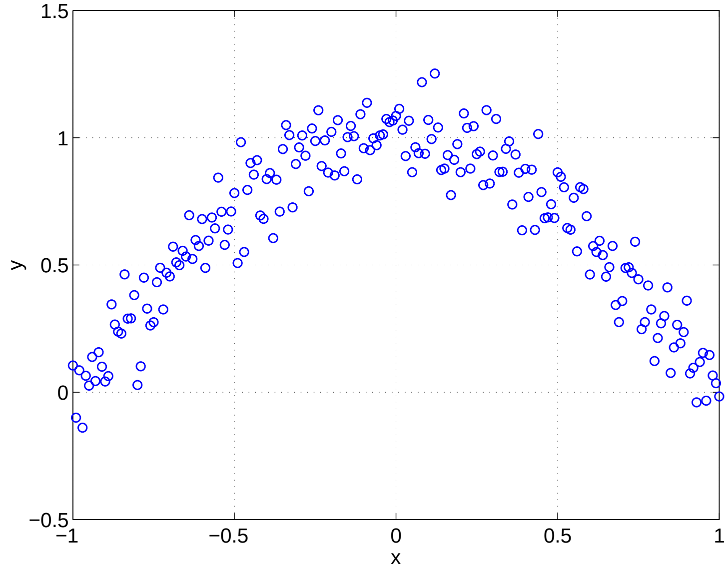

Figure 1. Samples from the underlying relation between the inputs x and outputs y. The outputs are not available to the learning algorithm unless specifically queried

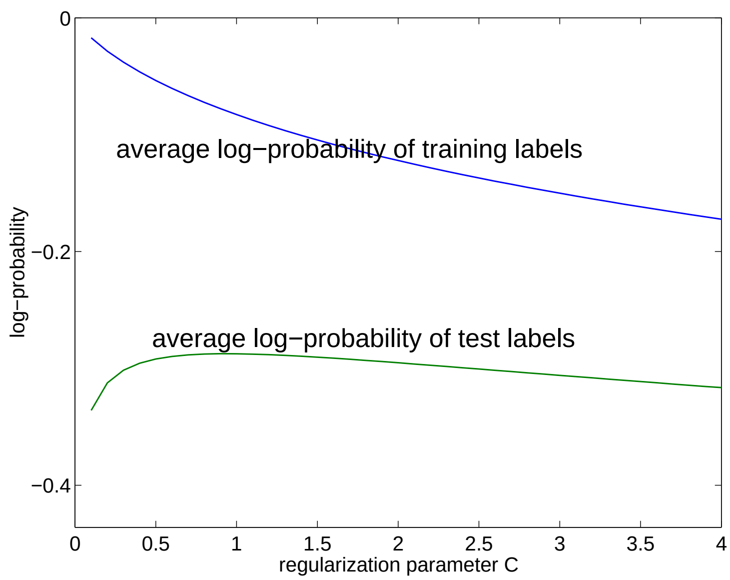

Figure 2. Log-probability of labels as a function of regularization parameter C

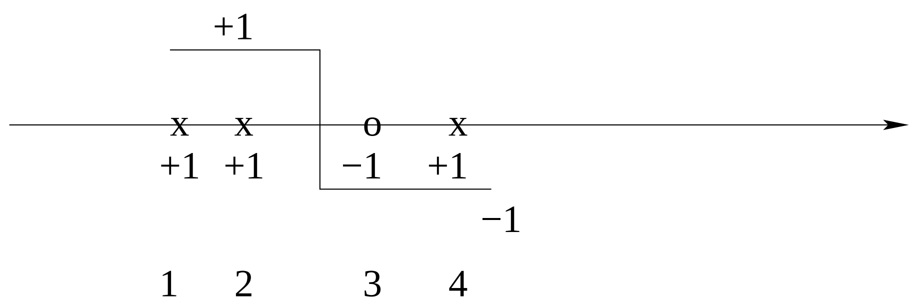

Figure 3. The first decision stump that the boosting algorithm finds.

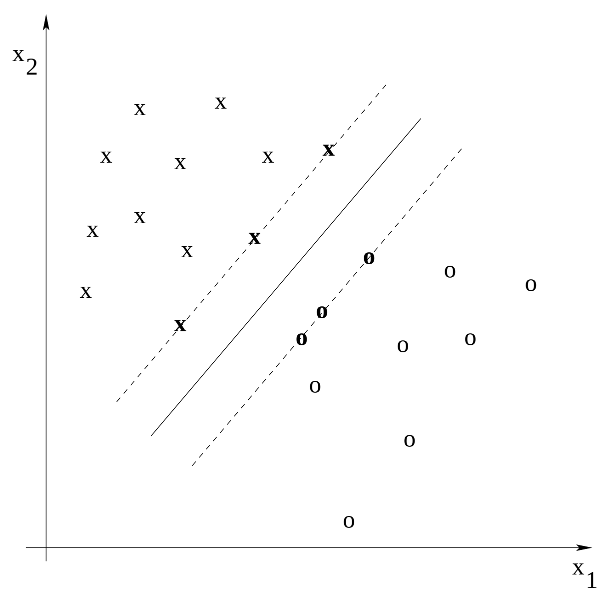

Figure 4. Training set, maximum margin linear separator, and the support vectors (in bold).