Name:

Massachusetts Institute of Technology

6.893 Machine Learning

Final Examination   Date: Dec 14, 1998

Problem 1: Perceptrons

Given the following input/output relationship:

X_1	X_2	T
0	0	1
0	1	0
1	0	0
1	1	1

Is there a simple perceptron that can learn this relationship?

Problem 2: Gaussians

Given the data:

X
1
2
3
4
5

Part: (A) (5 Points)

Compute the parameters of the Gaussian distribution which are most likely to have generated this data.

Part: (B) (5 Points)

Compute the mixture of Gaussian estimator which is most likely to have generated this data. I.e. find σ_i, and c_i for the Parzen estimator:

$$\hat{P}(x) = \sum_{i=1}^{k} G(x, \mu = c_i, \sigma = \sigma_i)$$

where $G(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ is the Gaussian density. You must also choose k, the number of mixtures components. (If you determine that there are multiple solutions, please give only one.)

Problem 3: Two Classes

There are two classes of data with the following distributions:

$$P(x|class1) = G(x, \mu = 0, \sigma = 1)$$

and

$$P(x|class2) = G(x, \mu = 5, \sigma = 1)$$

Where $G()$ is the Gaussian density. Data is drawn randomly from the two classes with equal probability.

Part: (A) (5 Points)

You are asked to design a machine learning approach which can learn to classify examples of each class. In order to find a good classifier for this problem, you are are given as much training data as you would like and you are free to pick any machine learning approach. What is the best error on test data that any learning algorithm can attain? (*Error* is the probability of making a mistake on a test example.)

- **(a)** 0% error
- **(b)** More that 0% but less than 10%.
- **(d)** More than 20%.
- **(e)** Can't tell from this information.

Part: (B) (5 Points)

Will the perceptron learning algorithm presented in the book always converge when trained on a set of training data for this problem?

Problem 4: Entropy

Part: (A) (5 Points)

Compute the entropy of a fair 6 sided die. (If you don't have a calculator, you need not compute the logarithm.)

Part: (B) (5 Points)

How many bits are required to transmit 1 role of this die?

Part: (C) (5 Points)

How many bits are required to send 1 sample from a Gaussian distribution with mean 0 and standard deviation 1 (i.e. $G(x, 0, 1)$).

Problem 5: High Dimensionality

Part: (A) (5 Points)

You are given 1000 two-dimensional data points distributed randomly and uniformly inside a circle of unit radius. Assume we wish to find a point that is near the center of the circle. Give a distance r such that 50% of the time at least 1 point will be within r of the center.

Part: (B) (5 Points)

You are given 1000 k-dimensional data points distributed randomly and uniformly inside a k-dimensional hyper-sphere of unit radius. Assume we wish to find a point that is near the center of the hyper-sphere.

(The volume of a hyper-sphere in k dimensions is some constant times the radius raised to the dimensionality: $V_k(r) = c_k r^k$.)

For $k = 500$, give a distance r such that 50% of the time at least 1 point will be within r of the center.

Part: (C) (5 Points)

What does this imply for problems involving 500-dimensional data? Explain. Give an example of a learning algorithm which is *unlikely* to work well on such data?

Part: (D) (5 Points)

I have experimented with a number of classification problems with a few thousand input dimensions, and a few thousand examples. Some of these are not difficult to solve, while others are very difficult. What property of this real data might have made it easy to build an accurate classifier.

Problem 6: Multi-layer Perceptrons

In this problem we will analyze the training of large multi-layer perceptrons (MLP's) which use sigmoid units. Assume that we have setup a network with several hidden layers and will be training it with back-propagation. Before training, weights will be initialized with small random values. The training data is divided into a training set (80%) and a validation set (20%) – the network is trained using back-prop to minimize the error on the training set only.

During training the following behavior is often observed: the error on training set and the validation set decreases as training progresses. For a while, the two errors track each other closely. At some point in the training process the error on the training set continues to decrease, while the error on the validation set begins to increase. This behavior does not seem to depend on the particular data points placed into the training and validation sets (a different subset of points could be selected for the validation set and a similar behavior would result).

Part: (A) (5 Points)

Give one reason why the error on the validation set initially decreases? (Try to keep your answer to two sentences or fewer.)

Give one reason why the error on the validation set might eventually increase. (Try to keep your answer to two sentences or fewer.)

Part: (B) (5 Points)

The classic approach is to stop training when error on the validation set is minimized. This is called "early stopping."

An alternative approach is to reduce the size of the network before training begins (either the number of layers, or the number of hidden units is decreased). Sometimes for the smaller network the training and validation errors never diverge. They both decrease together until training error achieves a minimum. Let's call this approach "network reduction."

Suggest a method to decide whether "early stopping" or "network reduction" is a better approach for a particular problem. (Try to keep your answer to a few sentences.)

Problem 7: Regression

Recall that regression is the name given to the problem of predicting a continuous output value for a set of continuous input values.

You are given the following data:

X	T
0.1	1
1.5	1.1
1.9	1.8
3.2	2.2
4.4	2.7

Part: (A) (5 Points)

How many continuous functions are consistent with the training data? (A function consistent if it predicts the training data perfectly.)

Part: (B) (5 Points)

How many polynomials of 4th order are consistent with this training data? How many 5th order polynomials are consistent with this data?

Part: (C) (5 Points)

One scheme for reducing the number of possible solutions is to use prior knowledge about valid solutions in order to select one solution as better than any other. This prior knowledge is expressed as a regularizer. Propose a reasonable regularizer that encodes your intuitions about regression problems. Describe what types of solutions this regularizer enforces.

Part: (D) (5 Points)

Consider the regularizer:

$$\frac{\nu}{2}\int (y^2+1)\left|\frac{dy}{dx}\right|^2 dx$$

Why might someone specify a regularizer like this one? In other words, what properties of the problem might be captured by this regularizer?

Recall that for some regularizers, all solutions will be of the form of a radial basis function network with a fixed radial basis function determined solely by the form of the regularizer. Will this regularizer yield an RBF solution?

Problem 8: SVM's

Part: (A) (5 Points)

You are given 10^4 examples of 3-dimensional data. The data is linearly separable. Would you expect a simple support vector machine to classify test data much better on this problem than a simple perceptron? (Neither the simple SVM nor the simple perceptron use non-linear features of the data.)

Part: (B) (5 Points)

You are given 1000 examples of 10^5-dimensional data. The data is linearly separable. Would you expect a simple support vector machine to classify test data much better on this problem than a simple perceptron?

(Some extra paper.)

(Some extra paper.)

(Some extra paper.)