Machine Learning - Assignment 1

Due Thursday, September 19, 2002

Note: The text of the homework is rather long because I tried to provide lots of hints. The solution is likely going to be shorter.

1. [20 points] Mitchell, pg. 48, Problem 2.4.
2. [5 points] Mitchell, pg. 49, Problem 2.7.
3. [10 points] Mitchell, pg. 50, Problem 2.8.
4. [10 points] Suppose that a learning algorithm is trying to find a consistent hypothesis in a concept learning problem, when the classifications of the examples are actually begin generated randomly. The examples have n Boolean attributes. Each example is drawn uniformly randomly from the set of 2^n. How many examples have to be drawn, before the probability of reaching a contradiction is at least p?
5. [35 points] In this problem, we explore the notions of maximum likelihood hypothesis for estimating the parameters of Gaussian distributions.

   Recall that if we are given a univariate Gaussian (or normal) distribution with mean μ and variance σ^2, the probability of a random sample of value x is given by:

   $$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

   Suppose now that you have a sample $x_1, \ldots x_n$ of values drawn independently from a normal distribution with an (unknown) mean μ and variance σ. The *likelihood* of the sample is the probability of the data given the parameters:

   $$L(x_1, \ldots x_n, \mu\sigma) = P(x_, \ldots x_n | \mu, \sigma)$$

   We use the L notation in the homework just to emphasize the fact that this probability depends on the unknown parameters μ and σ.

(a) Write down $L(x_1, \ldots x_n, \mu\sigma)$, assuming that $x_1, \ldots x_n$ are drawn independently.

(b) The maximum likelihood estimator for the mean, $\hat{\mu}(x_1, \ldots x_n)$ is the value of μ that maximizes the likelihood (as in our class definitions):

$$\hat{\mu}(x_1 \ldots x_n) = \arg\max_{\mu}(\max_{\sigma} L(x_1, \ldots x_n, \mu, \sigma))$$

Instead of searching for the maximum of L we will search for the maximum of $\log L$. To find the maximum, we have to solve the following equation:

$$\frac{\partial \log L(x_1, \ldots x_n, \mu, \sigma)}{\partial \mu} = 0$$

Solve this equation and show that its solution $\hat{\mu}$ (luckily!) does not depends on σ.

(c) Note that $\hat{\mu}$ is a function of the sample values, $x_1 \ldots x_n$ and hence can itself be viewed as a random variable. Recall that the expectation of a random variable Y drawn from a probability density function p can be computed as:

$$E_p[Y] = \int yp(y)dy$$

Show that the expectation of $\hat{\mu}$ when the data is sampled from $\mathcal{N}(\mu, \sigma)$ is $E_{X_1, \ldots X_n \sim \mathcal{N}(\mu,\sigma)}[\hat{\mu}(X_1, \ldots X_n)] = \mu$. Such an estimator, for which the expected value is equal to the "true" expected value is called *unbiased.*

Hint: recall that the expectation of a sum is equal to the sum of the expectations.

(d) Now we proceed in the same way to compute the maximum likelihood estimator for σ:

$$\hat{\sigma}(x_1, \ldots x_n) = \arg\max_{\sigma} L(x_1, \ldots x_n, \hat{\mu}, \sigma)$$

Calculate $\hat{\sigma}$, using the ideas we had in the previous computation. Does it matter if you take the derivative with respect to σ or σ^2?

(e) Show that $\hat{\sigma}^2$ is not an unbiased estimate of σ^2. Hint: Recall that the expectation of a product of independent random variables is the product of the expectations of each variable.

(f) Suggest an unbiased estimator $\bar{\sigma}^2$ based on the maximum likelihood estimator derived above.

(g) Suppose we know the exact mean μ of the distribution and we just want to estimate σ^2. Derive a maximum likelihood estimate for this case. Is this estimate unbiased?

6. [20 points] Suppose we are trying to learn a conjunctive target concept over n Boolean variables, of the form discussed in Lecture 2. Suppose the target concept is uniformly drawn from the space of such concepts at the beginning of learning. After the concept is drawn, you get m training examples drawn uniformly randomly from the instance space. You use these examples to construct the version space.

   (a) Show that the expected error of the Gibbs algorithm is not more that twice that of the Bayes optimal classifier for this particular task. The expectation is taken with respect to the set of target concepts.

   (b) Suppose instead of just using one hypothesis from the version space, you draw two such hypotheses and you want to combine their predictions. What class label will you provide?

   (c) (10 points extra credit) Compute the expected error of the new classifier Can you generalize you result to k hypotheses?