10-602: Statistical Approaches to Learning and Discovery

Problem Set 6
Date: April 17, 2001 **Due: start of class April 24, 2001**

1. This problem considers Bayesian methods for growing decision trees. At the top of the tree, the training data is divided into two groups by choosing an appropriate threshold. The same procedure is applied recursively to the groups until a stopping condition is reached, e.g. when a group only contains data from one class. What is the best threshold to use at each step? We can answer this question using statistical methods.

   Suppose we have a training set $D = \{(x_1, y_1), ..., (x_n, y_n)\}$ where x is univariate and $y = \pm 1$. The lecture notes describe a classification model where errors occur with uniform probability. Consider a generalization of that model to account for asymmetric error probability, i.e. errors above the threshold occur with rate e_1 and errors below the threshold occur with rate e_2:

$$p(y = 1|x, t, e_1, e_2) = (1 - e_1)\Theta(x - t) + e_2(1 - \Theta(x - t)) \quad (1)$$

$$\Theta(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

$$p(y = -1|x, t, e_1, e_2) = 1 - p(y = 1|x, t, e_1, e_2) \quad (3)$$

   We want to infer the threshold t. Let the error rates have a uniform prior distribution over their range $[0, 1]$.

   (a) What is the marginal likelihood for t, $p(D|t)$? (You should be able to obtain an analytical solution.) Maximizing this function is one way of choosing t.

   (b) What is the marginal likelihood for t for the symmetric model where $e_1 = e_2 = e$?

   (c) Is there an advantage to using the asymmetric model when growing a decision tree?

   Hint: Consider the dataset

x	0.1	0.2	0.3	0.4	0.5	0.6	0.7
y	-1	1	-1	1	1	1	1

2. This problem and the next explore numerical techniques for the threshold classifier with logistic noise, i.e.

$$p(y = \pm 1|x, t, a) = \sigma(ay(x - t)) \qquad \sigma(x) = \frac{1}{1 + e^{-x}} \quad (4)$$

$$p(t) \sim U(0, 1) \qquad p(a) \sim \mathcal{N}(0, 10) \quad (5)$$

   It is equivalent to the univariate logistic regression example in Tanner (p28), but parameterized differently. The parameter a controls the expected error rate of the classifier.

Given a data set $D = \{(x_1, y_1), ..., (x_n, y_n)\}$, the likelihood function is

$$p(D|t, a) = \prod_i p(y_i|x_i, t, a) \tag{6}$$

The parameter of interest is t, since it is all that is needed for classification. This involves integrals over a. Use the following data set D_1:

x	0.15	0.3	0.33	0.45	0.5	0.51	0.52	0.53	0.55	0.6
y	-1	-1	-1	-1	1	1	1	1	1	1

(a) Suppose t is known and we want to find the most probable a, i.e. $\text{argmax}_a\, p(D|t, a)p(a)$. Give a Newton step for updating a. Note that this is equivalent to a logistic regression problem on the modified data $z_i = y_i(x_i - t)$.

(b) The marginal likelihood for t, $p(D|t)$, summarizes what the data tells us about the threshold. A remarkably accurate approximation is possible via Laplace's method:

$$p(D|t) = \int_{-\infty}^{\infty} p(D|t, a)p(a)da \approx p(D|t, \hat{a})p(\hat{a})\sqrt{\frac{2\pi}{-h(\hat{a})}} \tag{7}$$

$$\hat{a} = \underset{a}{\text{argmax}}\, p(D|t, a)p(a) \qquad h(a) = \frac{d^2 \log(p(D|t, a)p(a))}{da^2} \tag{8}$$

The algorithm for $\hat{a}$ was found in part (a). Plot the resulting approximation to $p(D_1|t)$ by evaluating it over the range $t = 0, 0.01, 0.02, ..., 1$ (see e.g. Tanner p50). Please submit your code. Notice that the maximum-likelihood t is in an unusual place.

3. The optimal threshold t for the logistic classifier is approximately the median of $p(t|D)$. A suboptimal but simple alternative is the mean: $E[t|D]$. This value can be estimated by Monte Carlo using an appropriate importance distribution. There are several tradeoffs in choosing a Monte Carlo scheme, which are explored below. Answer as completely as you can.

(a) Given samples from the prior $p(t, a)$, how would you estimate $E[t|D]$? What if you had samples from the posterior $p(t, a|D)$? What are the tradeoffs involved in using the prior $p(t, a)$ for the importance distribution vs. using the posterior $p(t, a|D)$?

(b) Is there another distribution besides the above two which would be practical for importance sampling?

(c) To get independent samples from the posterior $p(t, a|D)$, we could use rejection sampling. What analytical work is required to obtain an efficient rejection sampler?

(d) Another approach to sampling from the posterior is to use a Markov chain, as in Gibbs sampling. What is the difficulty in applying Gibbs sampling to this problem?

(e) Markov chain sampling is also possible via the Metropolis algorithm with a symmetric proposal function ($Q(x|y) = Q(y|x)$). Choose a proposal function and detail the steps of a Metropolis sampler for the logistic regression posterior. What needs to be done to make sure the sampler is efficient for an arbitrary dataset?