10-602: Statistical Approaches to Learning and Discovery

Problem Set 6 Solutions

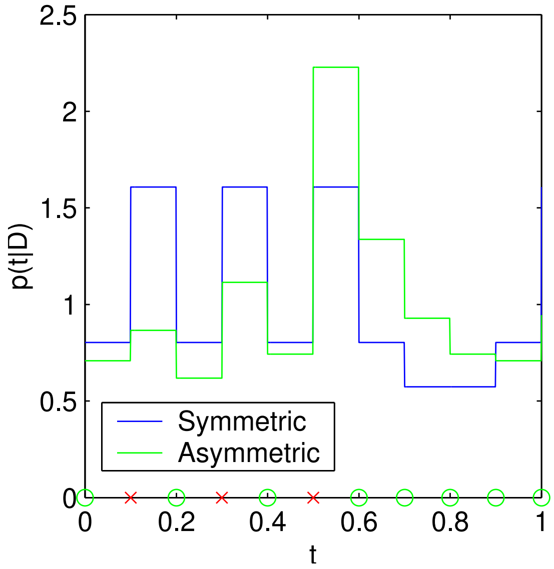
---

1. This problem comes from W. Buntine, "Learning Classification Trees" (`http://citeseer.nj.nec.com/buntine91learning.html`). Let $r_1(t)$ be the number of errors out of the n_1 points below t. Let $r_2(t)$ be the number of errors out of the n_2 points above t.

$$\begin{aligned} p(D|t,e_1,e_2) &= e_1^{r_1(t)}(1-e_1)^{n_1-r_1(t)}e_2^{r_2(t)}(1-e_2)^{n_2-r_2(t)} && (1)\\ p(D|t) &= \int_0^1\int_0^1 p(D|t,e_1,e_2)p(e_1)p(e_2)de_1de_2 && (2)\\ &= \frac{\Gamma(r_1(t)+1)\Gamma(n_1-r_1(t)+1)}{\Gamma(n_1+2)}\frac{\Gamma(r_2(t)+1)\Gamma(n_2-r_2(t)+1)}{\Gamma(n_2+2)} && (3)\end{aligned}$$

(b) With a uniform prior on e:

$$p(D|t) = \frac{\Gamma(r_1(t)+r_2(t)+1)\Gamma(n_1-r_1(t)+n_2-r_2(t)+1)}{\Gamma(n_1+n_2+2)} \qquad (4)$$



(c) The symmetric model is ambivalent between splits which have the same total number of errors. Of the splits with minimal error, the asymmetric model prefers splits that balance the size of the groups. This is advantageous for tree growing because it makes the tree shorter and ensures an adequate amount of data in each group. Another way to see it is that the asymmetric model incorporates lookahead by estimating how complex the trees will be for the groups (via the homogeniety of the groups).

By invoking the entropy approximation to the beta function, we can show that the asymmetric likelihood is a smoothed version of the "information gain" criterion typically used in decision trees. This parallels the statistical derivation we gave for mutual information feature selection.

A more subtle difference is that, because maximizing likelihood is suboptimal for the symmetric model, the maximum likelihood solution for the asymmetric model is a better approximation of the Bayes-optimal threshold, even when the symmetric model holds. For details see "Bayesian analysis of a threshold classifier."

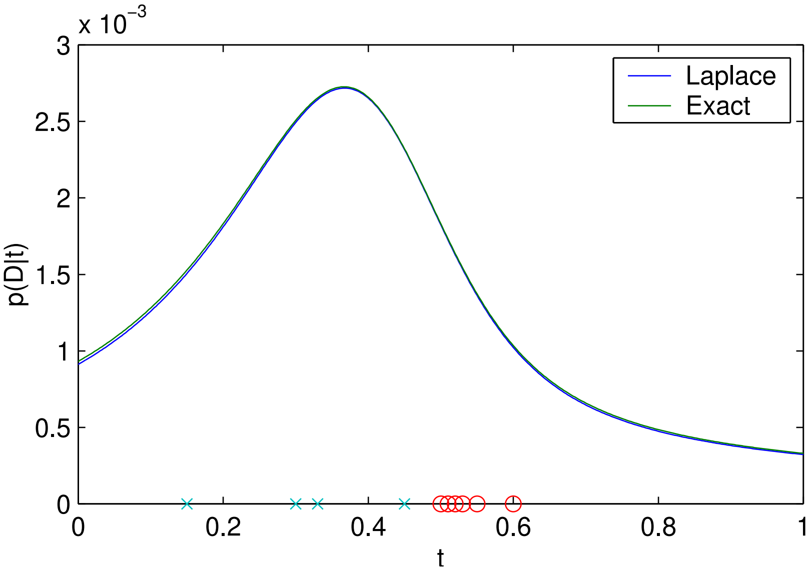
2. (a) Defining $v = 10$, $z_i = y_i(x_i - t)$ gives

$$\log p(D|t,a)p(a) = -\sum_{i=1}^{n} \log(1 + e^{-az_i}) - \frac{a^2}{2v} - \frac{1}{2}\log(2\pi v) \tag{5}$$

$$\frac{d \log p(D|t,a)p(a)}{da} = \sum_i (1 - \sigma(az_i))z_i - \frac{a}{v} \tag{6}$$

$$\frac{d^2 \log p(D|t,a)p(a)}{da^2} = -\sum_i \sigma(az_i)(1 - \sigma(az_i))z_i^2 - \frac{1}{v} \tag{7}$$

$$a^{new} = a^{old} + \left(\frac{1}{v} + \sum_i \sigma(az_i)(1 - \sigma(az_i))z_i^2\right)^{-1} \left(-\frac{a}{v} + \sum_i (1 - \sigma(az_i))z_i\right) \tag{8}$$



(b) The Laplace approximation to $p(D|t)$ is remarkably close to the exact curve, which is plotted underneath. Notice that the maximum-likelihood value, $\hat{t} = 0.37$, does not minimize training set error. The crowd of points labeled 1 suggests that the boundary is significantly left of 0.5, because otherwise there would be more ambiguity in that region. A less restrictive prior on a, i.e. bigger variance than 10, moves the MLE closer to 0.475, the maximum-margin solution.

3. (a) Using samples (t_k, a_k) from the prior:

$$E[t|D] = \frac{\int\int tp(D|t,a)p(t,a)dtda}{\int\int p(D|t,a)p(t,a)dtda} \quad (9)$$

$$p(D) \approx \frac{1}{K}\sum_{i=1}^{K} p(D|t_k, a_k) \quad (10)$$

$$E[t|D] \approx \frac{1}{Kp(D)}\sum_{i=1}^{K} t_k p(D|t_k, a_k) \quad (11)$$

Using samples (t_k, a_k) from the posterior:

$$E[t|D] = \int\int tp(t,a|D)dtda \quad (12)$$

$$\approx \frac{1}{K}\sum_{i=1}^{K} t_k \quad (13)$$

Sampling from the prior is easier than sampling from the posterior, but due to the sparsity of $p(D|t,a)$ most of those samples will be wasted. For a large data set, $p(D|t,a)$ will be a narrow peak around the MLE, so it is much more efficient to use samples from the posterior, even though they are more difficult to obtain. Notice that sampling from the prior still requires the computation of $p(D|t,a)$ for each sample, just as in a Metropolis scheme (below).

(b) We could make a normal approximation to the posterior and use that. The tails of the posterior must fall like e^{-x^2} because the prior for a was Gaussian. However, the variance of the normal approximation will probably be too small or too large. It can be tuned using adaptation—set the variance to the posterior variance estimated from previous samples. As long as each sample is normalized by the distribution it was sampled from, this approach is valid.

(c) Two steps are needed: finding a simple density to approximate the posterior and finding the minimal scale factor for it to be a bound. A simple choice for the proposal is the prior $p(t,a)$. Since the posterior is $\frac{p(D|t,a)}{p(D)}p(t,a)$, the minimal scale factor is the maximal value of $\frac{p(D|t,a)}{p(D)}$. The maximum of $p(D|t,a)$ is computed from the MLE, however $p(D)$ must estimated somehow. Monte Carlo is out, since if we had a Monte Carlo scheme for $p(D)$ we might as well use it for $E[t|D]$ and forget about rejection sampling. Instead we'll have to settle for a lower bound on $p(D)$, which might be obtained by variational methods. Another problem is that the posterior is always more peaked than the prior which means a large scale factor and a lot of wasted samples. We could use a normal approximation to the posterior instead, which would require bounding $\frac{p(D|t,a)p(t,a)}{p(D)g(t,a)}$ if $g(t,a)$ is the approximation. This requires bounding $p(D)$ and finding the maximum of $p(D|t,a)p(t,a)/g(t,a)$ using optimization on (t,a). The bottom line is that there is a fair amount of work involved in getting a decent rejection sampler.

(d) Applying Gibbs sampling directly to $p(t, a|D)$ is difficult since the conditionals $p(t|D, a)$ and $p(a|D, t)$ cannot be sampled from. Since the conditionals are one-dimensional, we could approximate them with a grid as described in Tanner. Another approach would be to introduce a hidden variable h such that $p(t, a|D) = \int p(t, a, h|D)dh$ and the conditionals are simple given h. This technique works for probit regression but not so well for logistic regression, where such an h is not available. This method is a special case of Metropolis-Hastings, which we discuss below.

(e) From the current sample for (t, a), new values (t^{new}, a^{new}) are proposed according to some distribution $Q(t^{new}, a^{new}|t, a)$ and accepted with probability

$$A = \min(1, \frac{p(t^{new}, a^{new}|D)}{p(t, a|D)}) = \min(1, \frac{p(D|t^{new}, a^{new})p(t^{new})p(a^{new})}{p(D|t, a)p(t)p(a)}) \tag{14}$$

Notice that $p(D)$ is not required. If the new values are not accepted, the old values are used again. A decent proposal to use in this context is the random walk:

$$Q(t^{new}, a^{new}|t, a) = \mathcal{N}(t^{new}; t, v_t)\mathcal{N}(a^{new}; a, v_a) \tag{15}$$

This proposal is symmetric, as required. To make this efficient for an arbitrary dataset, v_t and v_a need to be tuned to minimize the correlation coefficient of the chain, i.e. the correlation coefficient of (t_k, t_{k+1}). Another approach is to use a full multivariate Gaussian, whose covariance matrix comes from a second-order expansion of the posterior. This requires solving the maximum-likelihood problem first.

Another approach, not mentioned in the problem set, is to use an asymmetric Hastings rule. This is probably the best approach since the proposal could be tuned to approximate the posterior locally.