CS340 Fall 2006: Homework 3

Out Mon 25 Sep, back Mon 2 Oct

1 Maximum likelihood estimation of multinomials

Suppose $X \in \{1, 2\}$ and $Y \in \{1, 2, 3\}$. Define the joint distribution $P(X = j, Y = k) = \theta_{j,k}$. Consider the training data $\mathcal{D}$ below

X	Y
1	1
2	2
1	3
1	1
2	2
2	3

Find the maximum likelihood estimates

$$\hat{\theta}_{jk} = \arg\max \prod_{i=1}^{n} p(x_i, y_i|\theta) \tag{1}$$

where there are $n = 6$ training points. Hint: just normalize the counts! (The answer should be a 2×3 table of numbers that sum to one.)

2 Presidential debate

[Source: *Bayesian Data Analysis* 2nd edition (2004) p95, Gelman, Carlin, Stern and Rubin.]
On September 25, 1988, the evening of a presidential campaign debate in the USA, ABC News conducted two surveys of voting intentions, one before and after the debate, with these results:

Survey	Bush	Dukakis	Other
Pre	294	307	38
Post	288	332	19

Let us ignore the "other" responses. Let π_j represent the fraction of voters who prefer Bush in survey j ($j = 1$ is pre debate survey, $j = 2$ is post debate survey). Assume that the two surveys are independent samples from the population of registered voters. Let π_j have a $beta(1, 1)$ prior before the survey. Hence we have $p(\pi_1|S_1) = Beta(295, 308)$ and $p(\pi_2|S_2) = Beta(289, 333)$, where S_j is the j'th survey. What is the probability that there was a shift towards Bush as a result of the debate?
Answer: The probability there was a shift towards Bush is given by

$$\begin{aligned} p(\pi_2 > \pi_1|S_1, S_2) &= EI(\pi^2 > \pi^1) && (2) \\ &= \int_0^1 \int_0^1 p(\pi_1|S_1)p(\pi_2|S_2)I(\pi^2 > \pi^1)d\pi_1 d\pi_2 && (3) \end{aligned}$$

where we have used the trick that the probability of a binary event, $X = \pi_2 > \pi_1$, is the expectation of its indicator, $EI(X)$, where $I(X) = 1$ if X is true and $I(X) = 0$ otherwise. We can further simplify this integral thus

$$p(\pi_2 > \pi_1|S_1, S_2) = \int_0^1 \int_0^{\pi_2} p(\pi_1|S_1)p(\pi_2|S_2)d\pi_1 d\pi_2 \tag{4}$$

which shows that we are just computing the posterior probability mass above the diagonal line $\pi_1 = \pi_2$.
We can approximately solve this integral using **Monte Carlo integration**

$$E[f(\theta)|D] = \int f(\theta)p(\theta|D) \approx= \frac{1}{N}\sum_{i=1}^{N} f(\theta^i) \tag{5}$$

where $\theta^i \sim p(\theta|D)$ is a sample from the appropriate posterior and N is the number of samples (say, 1000).
In this case, we can use

$$\begin{aligned} p(\pi_2 > \pi_1|S_1, S_2) &= EI(\pi_2 > \pi_1) &(6)\\ &\approx \frac{1}{N}\sum_{i=1}^{N} I(\pi_2^i > \pi_1^i) &(7)\end{aligned}$$

where $\pi_1^i \sim Beta(295, 308)$ and $\pi_2^i \sim Beta(289, 333)$.
Question: implement Equation 7 in matlab. Turn in your code and numerical answer.
Hint: in the matlab statistics toolbox, you can use `betarnd` to draw samples from a beta distribution. If you don't have the statistics toolbox, you can use `randbeta` from the lightspeed toolbox, which is freely available (google Tom Minka's web page).
Bonus: plot the exact (factored) posterior $p(\pi_1, \pi_2|S_1, S_2)$ on a grid of points, superimpose the line $\pi_1 = \pi_2$ and your sampled points. The fraction of points lying above the line is your estimate. Use numerical integration to compute the exact answer.

3 Bayesian concept learning

In this question, you will implement the Bayesian concept learning framework for the "number game" we discussed in class. You are provided the following functions

- `hypSpace = mkHypSpace()` which creates the hypothesis space (a structure). The only field you should need is called 'hyps', which is a cell array. To extract the set of integers defined by the h'th hypothesis (this is called the support or extension of the hypothesis), use the following:

  ```
  hypSpace.hyps{h}
  ```

  There are `hypSpace.Hmath=23` mathematical hypotheses, and `hypSpace.Nint =5050` interval hypotheses, stored in order in order of increasing size. Thus

  ```
  hypSpace.hyps{24} = 1
  hypSpace.hyps{25} = 2,
  hypSpace.hyps{124} = [1,2]
  ```

  etc.

- `prior = mkPrior(hypSpace)`, which creates a (row) vector, in which $prior(h) = p(h)$, for h=1:5073.

Use these to answer the following questions

1. Write a function `lik = likelihood(hypSpace, X)` which computes

   $$lik(h) = p(X|h) = \begin{cases} \frac{1}{|size(h)|^n} & \text{if all } x_1, \ldots, x_n \in h \\ 0 & \text{if any } x_i \notin h \end{cases}$$

   where lik is a vector, which one element for each possible hypothesis. Turn in your code.

2. Write a function `post = mkPost(hypSpace, X)` which computes

$$post(h) = \frac{p(X|h)p(h)}{\sum_{h'} p(X|h')p(h')}$$

where $post$ is a vector. Turn in your code.

3. Suppose $X = [32]$. Compute the posterior $post(h) = p(h|X)$. Plot the posterior over the mathematical hypotheses
$post(1:32)$ Turn in your plot.

4. What is the maximum a posterior (MAP) hypothesis

$$h_{MAP} = \arg\max_h p(h|X)$$

Print out the extension of hMAP (i.e. its list of integers). From the extension, you should be able to infer the name of the "rule" that defines it. e.g., if hMAP is $[2, 4, \ldots, 96, 98, 100]$, then the rule is "even enumbers"; if hMAP is $[33, 34, 35]$, then the rule is "interval 33:35". (You can also look at `mkHypSpace.m` to figure out the rule from the index h.) What is the rule corresponding to h_{MAP}?

5. Sort the hypotheses into decreasing order of posterior probability. What are the top 5 most probable hypotheses? (Return their names/rules in addition to their numeric ids.) Hint: you may find the function `celldisp` helpful.

6. Draw a sample of 5 hypotheses from the posterior. Return their names/rules in addition to their numeric ids. Please set the random number seed as shown below, to ensure everyone's results are the same

```
seed = 0;
rand('state', seed);
randn('state', seed);
```

Hint: you can use the provided function `data = sample_discrete(prob, 1, n)` to sample n points from a discrete probability distribution.

7. Write a function to compute the posterior predictive distribution

$$pred(x) = p(y(x) = 1|X) = \sum_{h \in \mathcal{H}} p(y(x) = 1|h)p(h|X)$$

where $pred(x)$ is a vector, with one element for each $x = 1 : 100$, X is the training data, and $y(x) = 1$ if x is in the concept, and $y(x) = 0$ otherwise. (Obviously $y(x) = 1$ for all $x \in X$; the goal is to generalize beyond the training set, i.e., to predict which other numbers are in the concept class.) Plot pred(x) as a histogram. Turn in your code and plot.

8. What is $p(y(6) = 1|X)$? What is $p(y(7) = 1|X)$? What is $p(y(8) = 1|X)$?

9. Write a function to compute the maximum likelihood estimate

$$\hat{h}_{ML} = \arg\max_h p(X|h)$$

Turn in your code. What is $\hat{h}_{ML}$? (Give its number and name/rule.)

10. Write a function to compute the plug-in estimate

$$predML(x) = p(y(x) = 1|\hat{h}_{ML}(X))$$

Plot predML(x) as a histogram. Turn in your code and plot. Why is $predML$ worse than $pred$?

11. Now repeat steps 3-10 using $X = [32, 2, 44, 64, 88, 2, 10]$. Turn in your new plots and numbers. What are the main qualitative differences when using this larger, less ambiguous sample?