CSE 569 Homework #2 Solutions

Problem 2 Sample Answer:

Looking at the following expression for $p(\mathbf{x}|D)$:

$$p(\mathbf{x}|D) = \int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}\,|D)d\boldsymbol{\theta} = \int p(\mathbf{x}|\boldsymbol{\theta})\frac{p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(D)}d\boldsymbol{\theta}$$

It appears that, if we have a uniform prior, i.e., $p(\boldsymbol{\theta})$=constant, then it will disappear from the equation (i.e., replacing $p(\boldsymbol{\theta})$ in the equation by a constant). This appears to suggest that a uniform prior means the prior information has no impact on the final density. But it is not the case, and even if the final expression does not seem to contain $p(\boldsymbol{\theta})$ as a function of $\boldsymbol{\theta}$, the prior information is reflected in the integration interval over which the prior is a constant.

You can read "Example 1. Recursive Bayesian Learning" and "Section 3.5.2 Noninformative Priors and Invariances" to gain more insight into this discussion.

Problem 3 Sample Answer:

One extreme example is a 2-D dataset with samples evenly distributed inside a circle. In such a case, if we use PCA to identify a direction to project the samples (for reducing the dimension from 2 to 1), it cannot be effective or useful (since the direction given by PCA is simply as good (or as bad) as any random directions including any of the original features).

Problem 4 Sample Answer:

(1) The E-step finds the expectation of the likelihood.
(2) The expectation in (1) is found based on the conditional density of the missing features given the good features and the parameters estimated previously.
(3) Yes, the distribution is a conditional density, with the training data in the "given" part, and that is how the training data contribute to the estimation of a better parameter.

Problem 5 Sample Answer:

When the state sequences are also given, the problem is easily solved. The MLE estimates can be found by computing the relative frequencies. See Slide 72 of the lecture notes for t & e (and the initial probabilities can be similarly estimated). We will just need to go through the training set to find all those counts (the number of certain state, the number of transitions from certain state i to state j, etc.).

Problem 6 Sample Answer:

With such a special emission probability matrix, the observations become non-informative and the HMM is essentially reduced to a discrete Marko process (e.g., like the one illustrated on Slide 60). Revisiting the Viterbi Algorithm (Slide 68 in Notes 04), we can see that such a special emission probability matrix is equivalent to having the same constant scaler 1/M in place of all e(o|S) and thus the observations have no impact on the final solution.

Problem 1 Solution:

The given P is called the Poisson distribution. From "Exercises on MLE", you would know the MLE solution for this problem is simply X_1. Now let's find the Bayesian estimate and compare with the MLE.

We need to first find $f(\lambda|D) = P(D|\lambda)f(\lambda) / \int P(D|\lambda)f(\lambda)d\lambda$, under the given $f(\lambda)$, before we can find its mean. The actual computation is quite involving although it is conceptually easy. See next two pages for details.

Problem 1 Solution:

This needs some extra efforts on getting the results right, although it is easy to write down the initial steps.

Let's first compute $f(\lambda|D)$,

$$f(\lambda|D) = \frac{P(D|\lambda) f(\lambda)}{\int P(D|\lambda) f(\lambda) d\lambda}, \leftarrow P(D|\lambda) = \frac{\lambda^x}{x!} e^{-\lambda}$$

So if we can get $W \triangleq \int P(D|\lambda) f(\lambda) d\lambda$, we can simply plug W, $P(D|\lambda)$, and $f(\lambda)$ into the above expression to get $f(\lambda|D)$.

With that, finding the mean of $f(\lambda|D)$ can be solved by $\int \lambda f(\lambda|D) d\lambda$.

It is conceptually easy & clear to this point.

Now, let's do the actually computation of W and $\int \lambda f(\lambda|D) d\lambda \triangleq E[\lambda|D]$

$$W = \int P(D|\lambda) f(\lambda) d\lambda = \int_0^\infty \frac{\lambda^x}{x!} e^{-\lambda} \cdot e^{-\lambda} d\lambda = \int_0^\infty \frac{\lambda^x e^{-2\lambda}}{x!} d\lambda$$

$$= \int_0^\infty \frac{(2\lambda)^x e^{-2\lambda}}{2^x \, x!} d\lambda = \frac{1}{2^{x+1}} \int_0^\infty \frac{t^x e^{-t}}{x!} dt$$

Change of variable: $t = 2\lambda$

Now, we'll need to know the integral is actually for a Gamma distribution with parameters $\{K = x+1, \theta = 1\}$, and thus the integral $= 1$.

e.g. see wikipedia.

$$\therefore W = \frac{1}{2^{x+1}}$$

$$\therefore f(\lambda|D) = 2^{x+1} \frac{\lambda^x}{x!} e^{-2\lambda}, \quad \lambda > 0.$$

Now, we need to compute $E[\lambda|D] \triangleq \int_0^\infty \lambda f(\lambda|D) d\lambda$

$$E(\lambda|D)=\int_0^\infty \lambda f(\lambda|D)\,d\lambda$$

$$=\int_0^\infty 2^{x+1}\frac{\lambda^{x+1}}{x!}e^{-2\lambda}\,d\lambda$$

This appears again hard to compute, but we will again use the same trick of Gamma distribution to help,

$$=\int_0^\infty \left(\tfrac{1}{2}\right)(x+1)\frac{\lambda^{x+1}e^{-2\lambda}}{(x+1)!\left(\frac{1}{2}\right)^{x+2}}\,d\lambda$$

Looking at this part, it is a Gamma distribution with parameter $\{K=x+2$ & $\theta=\frac{1}{2}\}$, and thus

$$=\frac{x+1}{2},$$ (since the rest integrates to 1).

As you can see, now, even if we have only one data point x which happens to be 0, we will get a meaningful estimate of $\frac{0+1}{2}=\frac{1}{2}$.

In contrast, if we use MLE with a single sample and that sample happen to be 0, then the MLE gives 0, which is not a meaningful estimate since the parameter needs to be >0.