3F8 Exam Crib

2020-2021

Question 1

a)

$$
\begin{aligned}
p(x_n|\sigma^2) &= \mathcal{N}(x_n; 0, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} x_n^2\right) \\
p(\{x_n\}_{n=1}^N|\sigma^2) &= \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{n=1}^N x_n^2\right) \\
p(\sigma^2|\{x_n\}_{n=1}^N) &\propto p(\sigma^2|\alpha,\beta) p(\{x_n\}_{n=1}^N|\sigma^2) \\
&= p(\sigma^2|\alpha,\beta) \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{n=1}^N x_n^2\right) \\
&\propto (\sigma^2)^{-\alpha/2} (\sigma^2)^{-N/2} \exp\left(-\frac{1}{2\sigma^2}\left[\sum_{n=1}^N x_n^2 + \beta\right]\right) \\
&= (\sigma^2)^{-(\alpha+N)/2} \exp\left(-\frac{1}{2\sigma^2}\left[\sum_{n=1}^N x_n^2 + \beta\right]\right)
\end{aligned}
$$

Recognising that $\alpha' = \alpha + N$ and $\beta' = \sum_{n=1}^N x_n^2 + \beta$, shows the posterior has the same form as the prior with normalising constant $Z(\alpha', \beta')$. Here α is the number of pseudo-observations in the prior. β/α is the sample second moment of the pseudo-observations, i.e., the empirical value of σ^2 for the pseudo-observations.

Note that in the exam, a fairly large number of people included the prior N times (i.e. raised to the power of N) rather than once for the entire data set.

b) (i) Given a posterior distribution $p(\sigma^2|\{x_n\}_{n=1}^N)$, the MAP estimate is the value of σ^2 that maximises the posterior. The ML estimate is the value that maximises the likelihood $p(\{x_n\}_{n=1}^N|\sigma^2)$. They are both point estimates. MAP estimate also depends on the prior.

(ii) Take logarithms and set the derivative to zero, ignoring constant terms. We get:

$$\left(-\frac{\alpha'}{2}\right)\frac{1}{\sigma^2}+\frac{\beta'}{2(\sigma^2)^2}=0$$

$$\sigma^2=\frac{\beta'}{\alpha'}$$

(iii) Need

$$\frac{\beta'}{\alpha'}=\frac{\sum_{n=1}^{N}x_n^2}{N}$$

$$\frac{\sum_{n=1}^{N}x_n^2+\beta}{\alpha+N}=\frac{\sum_{n=1}^{N}x_n^2}{N}$$

Which is true if $\beta=0$ and $\alpha=0$, or in the limit of infinite data $N\to\infty$ where the data overwhelm the prior.

(iv) You could computer some measure of posterior uncertainty, such as the posterior variance or the second moment around the MAP estimate. However, the posterior will be highly asymmetric in general, and so a better solution to reporting the variance would be so-called credible intervals (e.g. report the median of the posterior and where the 25% and 75% quantiles lie – the underlying value should be between these two values half the time).

Question 2

a) Compute the log posterior and recognise the quadratic form, dropping terms that don't depend on m. We get:

$$-\frac{1}{2}\left[m^2\left(\sum_{n=1}^{N}\frac{x_n^2}{1+x_n^4}+1\right)-m\left(\sum_{n=1}^{N}\frac{x_n y_n}{1+x_n^4}\right)\right]+\text{constant}$$

Completing the square, we get that the posterior is Gaussian with mean and variance given by:

$$\mu=\frac{\sum_{n=1}^{N}\frac{x_n y_n}{1+x_n^4}}{\sum_{n=1}^{N}\frac{x_n^2}{1+x_n^4}+1} \tag{1}$$

$$\sigma^2=\frac{1}{\sum_{n=1}^{N}\frac{x_n^2}{1+x_n^4}+1} \tag{2}$$

Note that in the exam, many people did not read the question carefully and started to compute the distribution over the output variable y with m marginalised out i.e. $p(y|x)$. In addition, many people incorrectly thought that the following identity holds when simplifying the solution: $\frac{\sum_n f_n/g_n}{\sum_n h_n/g_n}=\sum_n \frac{f_n}{h_n}$ which is incorrect.

b) (i) $\mu=0$, $\sigma^2=1$ (data points at $x=0$ provide no information about the slope as the model predictions at this location do not depend on the outputs y)

(ii) $\mu=-3/2$ and $\sigma^2=1/2$ (maximum likelihood with constant observation noise would have slope $\mu=-3$, but observation noise is relatively large at $x=\pm1$ being equal to 2 so estimate is reduced downwards as the points could have arisen from a smaller slope that had noise added to it; similarly the posterior uncertainty is still high)

c) We seek the value of x that is a maximum of $x^2/(1+x^4)$, since that leads to a minimum of the posterior variance (minimum parameter uncertainty after the observation). Taking derivatives with respect to x^2, we get $x^2=1$ or $x=\pm1$.

In the exam, many got the intuition that there was a trade-off (observation noise increases with x *and* signal increases) but few realised that they could minimise the posterior variance on m for one data point wrt the input x to find an analytic solution.

Question 3

a) Set the approximate posterior $q(s_n)$ to $p(s_n|x_n)$, and compute $\mathcal{F}(\theta, \{p(s_n|x_n)\}_{n=1}^N)$.

$$\begin{aligned} p(s_n = 1|x_n) &= p(x_n|s_n = 1)p(s_n = 1)/p(x_n) \\ &= \frac{\frac{\rho}{\lambda_k}\exp(-x_n/\lambda_k)}{\frac{(1-\rho)}{\lambda_0}\exp(-x_n/\lambda_0) + \frac{\rho}{\lambda_1}\exp(-x_n/\lambda_1)} \end{aligned}$$

or using the logistic rather than softmax form

$$p(s_n = 1|x_n) = \frac{1}{1 + \frac{1-\rho}{\rho}\frac{\lambda_1}{\lambda_0}\exp(-x_n(1/\lambda_0 - 1/\lambda_1))}$$

b) Holding $\{q(s_n)\}_{n=1}^N$ fixed, maximise $\mathcal{F}(\theta, \{q(s_n)\}_{n=1}^N)$ with respect to θ.

$$\begin{aligned} \mathcal{F}(\theta, \{q(s_n)\}_{n=1}^N) &= \sum_{n=1}^{N}\sum_{k=0}^{1} q(s_n = k)\log p(s_n = k, x_n) + \text{constant} \\ &= \sum_{n=1}^{N}\sum_{k=0}^{1} q(s_n = k)\left(-\log\lambda_k - x_n/\lambda_k + x_n\log\rho + (1 - x_n)\log(1-\rho)\right) + \text{constant} \\ \frac{\partial\mathcal{F}}{\partial\lambda_k} &= \sum_{n=1}^{N} q(s_n = k)\left(-\frac{1}{\lambda_k} + \frac{x_n}{\lambda_k^2}\right) = 0 \\ \lambda_k &= \frac{\sum_{n=1}^N q(s_n = k)x_n}{\sum_{n=1}^N q(s_n = k)}. \end{aligned}$$

Similarly, derivatives for ρ – which requires a Lagrange multiplier – yield $\rho = \frac{1}{N}\sum_{n=1}^N q(s_n = 1)$

In the exam, lots of people substituted in the optimal form for q from question 2a into the expression for the free-energy and then took derivatives of the resulting expression. This actually recovers the true likelihood. Instead, in EM, q is treated as fixed i.e. it does not depend on the parameters, and so $\frac{\mathrm{d}q}{\mathrm{d}\theta} = 0$.

(c) Probability decay events given parameters is:

$$p(x_n|\lambda_0, \lambda_1) = (1-\rho)\frac{1}{\lambda_0}\exp(-x_n/\lambda_0) + \rho\frac{1}{\lambda_1}\exp(-x_n/\lambda_1)$$

This expression is also the likelihood of the parameters. After the E-Step the EM algorithm's free energy is equal to the likelihood of the parameters $\mathcal{F}(\theta, \{q(s_n)\}_{n=1}^N) = \sum_{n=1}^N \log p(x_n|\lambda_0, \lambda_1)$. Also, at convergence the EM algorithm finds parameters which are a (local) optimum of the log-likelihood of the parameters. Finally notice that this expression is the normalising constant of the posterior computed in part (a).

Question 4

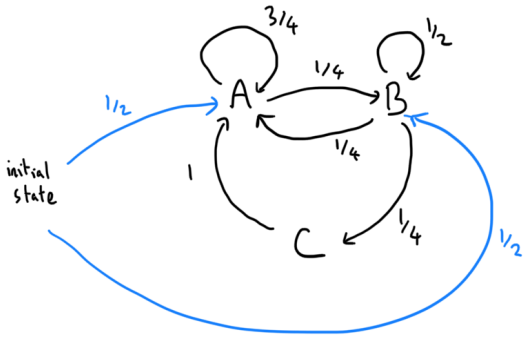
(a) (i) Let $p(y_1 = k) = \pi_k$ and $p(y_{t+1} = j | y_t = k) = T_{jk}$.

$$\log p(y_{1:T}^{(1)}, y_{1:T}^{(2)}) = \log p(y_{1:T}^{(1)}) + \log p(y_{1:T}^{(2)})$$

Use standard trick to write the log probability as a sum of logs with indicator functions. Optimise π_k with a Lagrange multiplier to obtain:

$$\pi_k \propto \mathbb{1}[y_1^{(1)} = k] + \mathbb{1}[y_1^{(2)} = k]$$

Hence $\pi_A = 0.5, \pi_B = 0.5, \pi_C = 0$. Do the same with T_{jk} to get the following state-transition probabilities.



In the exam, a large number of people missed out the maximum likelihood setting of the initial state distribution from their solutions.

(ii) The probability of the sequence under this model is 0 since $A \to C$ never occurs in the training data. To improve, could put a prior on π and T and do MAP estimation or go the whole-hog and perform Bayesian inference instead of using point estimates.

(b) (i) Kalman filter. This is a hidden Markov model which has linear Gaussian observation likelihoods and a linear Gaussian hidden state transition probability.

(ii) Idea is to modify λ and σ^2 in the AR model so that a single transition is equal in distribution to two transitions under the original AR model. Let $\epsilon \sim \mathcal{N}(0, 1)$

$$\begin{aligned} x_{t+2} &= \lambda x_{t+1} + \sigma \epsilon_{t+1} \\ &= \lambda(\lambda x_t + \sigma \epsilon_t) + \sigma \epsilon_{t+1} \\ &= \lambda^2 x_t + \lambda \sigma \epsilon_t + \sigma \epsilon_{t+1}. \end{aligned}$$

We can achieve this behaviour by setting $\lambda' = \lambda^2$ and $\sigma'^2 = \sigma^2(\lambda^2 + 1)$. In the exam, a lot of people made a mistake in this last step.
The likelihood is unchanged.