Crib for 3F8: Inference

Exercise 1

a) *Maximum a posteriori* (MAP) estimation can be used to estimate the parameters $\boldsymbol{\theta}$ of a probabilistic model from data $\mathcal{D}$. The model specifies a likelihood function $p(\mathcal{D}|\boldsymbol{\theta})$ and a prior distribution $p(\boldsymbol{\theta})$. Bayes rule gives us the posterior distribution $p(\boldsymbol{\theta}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$. The MAP estimate $\hat{\boldsymbol{\theta}}_{\mathrm{MAP}}$ is the one that maximizes this posterior. In particular,

$$\hat{\boldsymbol{\theta}}_{\mathrm{MAP}} \quad = \quad \underset{\boldsymbol{\theta}}{\mathrm{argmax}} \, \log p(\mathcal{D}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) \,,$$

where the log is used to perform numerically stable optimization. The MAP estimate balances between maximizing the likelihood $p(\mathcal{D}|\boldsymbol{\theta})$ and the prior assumptions $p(\boldsymbol{\theta})$.

b)

(i) The likelihood function is $p(\mathrm{Data}|\rho) = \prod_{i=1}^{10} \rho^{x_i}(1-\rho)^{1-x_i} = \rho^3(1-\rho)^7$. To find the MLE we compute the log-likelihood and find the value of ρ that sets its gradient to zero. In particular,$\log p(\mathrm{Data}|\rho) = 3\log\rho + 7\log(1-\rho)$. Then

$$\frac{d \log p(\mathrm{Data}|\rho)}{d\rho} \quad = \quad \frac{3}{\rho} - \frac{7}{1-\rho} = 0 \Leftrightarrow \rho = 3/10 \;.$$

(ii) The MAP estimate is obtained as

$$\frac{d[\log p(\mathrm{Data}|\rho) + \log p(\rho)]}{d\rho} \quad = \quad \frac{4}{\rho} - \frac{7}{1-\rho} = 0 \Leftrightarrow = 4/11 \,.$$

(iii) $p(y|\rho) = \rho^y(1-\rho)$. The joint probability is given by

$$p(y_1, y_3, y_3, \rho) \quad = \quad p(y_1, y_2, y_3|\rho)p(\rho) = \rho^9(1-\rho)^3 2\rho \,.$$

We find the MAP estimate by taking the gradient of the log joint and setting it to zero:

$$\frac{\partial \log p(y_1, y_3, y_3, \rho)}{\partial \rho} \quad = \quad \frac{10}{\rho} - \frac{3}{1-\rho} = 0 \Leftrightarrow \rho = \frac{10}{13} \,.$$

c) $p(y_\star|x_\star, \mathcal{D}) = \int p(y_\star|x_\star, w)p(w|\mathcal{D})\, dw = \mathcal{N}(y_\star|x_\star m, x_\star^2 v + 1)$.

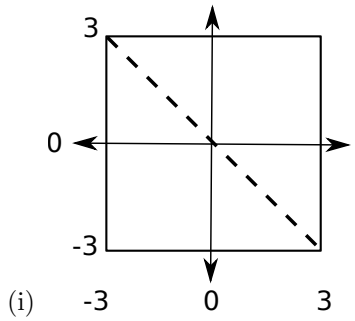
Assessor's comments: A very popular and straightforward question. Most candidates gave good answers. The main difficulty was coming up with the geometric distribution for a series of coin tosses until you get heads for the first time and computing the predictive distribution in a Bayesian linear regression model. The high average mark obtained by candidates in this question was compensated by the lower average values obtained in Q3 and Q4.

Exercise 2

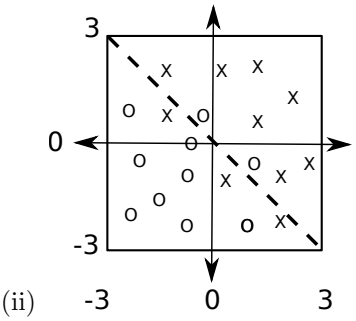
a) Bayesian decision theory selects the action that maximizes the expected reward under the posterior distribution. In particular,

$$a_\star \quad = \quad \underset{\boldsymbol{\theta}}{\text{argmax}} \int R(a, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{D}) \, .$$

b)



(i)



(ii)

(iii)

$$\begin{aligned} p(y_n|\mathbf{w},\epsilon,\mathbf{x}_n) &= \epsilon\frac{1}{2} + (1-\epsilon)p(y_n|\mathbf{w},\mathbf{x}_n) \\ & \epsilon\frac{1}{2} + (1-\epsilon)\left[\frac{1+y_n}{2}\sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n) + \frac{1-y_n}{2}(1-\sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n))\right]. \end{aligned}$$

(iv) We have that the joint distribution for y_n and z_n is given by

$$p(y_n, z_n|\mathbf{w},\epsilon,\mathbf{x}_n) = z_n\epsilon\frac{1}{2} + (1-z_n)(1-\epsilon)p(y_n|\mathbf{w},\mathbf{x}_n)\,.$$

The conditional is then

$$p(z_n = 1|y_n,\mathbf{w},\epsilon,\mathbf{x}_n) = \frac{\epsilon\frac{1}{2}}{\epsilon\frac{1}{2} + (1-\epsilon)p(y_n|\mathbf{w},\mathbf{x}_n)}\,.$$

We expect this probability to be high when $p(y_n|\mathbf{w},\mathbf{x}_n)$ is low. That is, when the model does not describe the data well.

Assessor's comments: The third most popular question. While most candidates gave good answers, many struggled with part b) ii where they did not take into account the randomness in the samples generated by the model when making predictions and with part b) iii and c) iv where they did not come up with the correct probabilities.

Exercise 3

a) Let $\mathbf{z} = (z_1,\ldots,z_N)$ and $\mathbf{x} = (x_1,\ldots,x_N)$ and $p(\mathbf{x},\mathbf{z}|\theta) = \prod_{i=1}^N p(x_i,z_i|\theta)$. The maximum likelihood estimate of θ is obtained by maximizing $\log p(\mathbf{x}|\theta) = \log\sum_{\mathbf{z}} p(\mathbf{x},\mathbf{z}|\theta)$, which is intractable. The EM algorithm optimizes the lower-bound

$$\log\sum_{\mathbf{z}} p(\mathbf{x},\mathbf{z}|\theta) \geq \sum_{i=1}^N\sum_{z_i} q(z_i)\log\frac{p(x_i,z_i|\theta)}{q(z_i)}\,.$$

where the $q(z_i)$ form a variational distribution. EM iteratively maximizes this lower bound with respect to the $q(z_i)$ (E-step) and with respect to θ (M-step). When the E-step guarantees that $q(z_i) = p(z_i|x_i,\theta)$ we have that the EM algorithm is guaranteed to find a maximum likelihood estimate.

b)

(i) We have that

$$p(x_i,z_i|\rho_A,\rho_B) = 0.5\left[\begin{pmatrix}10\\x_i\end{pmatrix}\rho_A^{x_i}(1-\rho_A)^{10-x_i}\right]^{z_i}\left[\begin{pmatrix}10\\x_i\end{pmatrix}\rho_B^{x_i}(1-\rho_B)^{10-x_i}\right]^{1-z_i}.$$

The free-energy is then

$$\begin{aligned}\mathcal{F}(\rho_A,\rho_B,q_1,\ldots,q_5) &= \sum_{i=1}^{5}\sum_{z_i=0}^{1} p_i^{z_i}(1-p_i)^{(1-z_i)} \\ &\quad \log\frac{0.5\left[\left(\begin{array}{c}10\\x_i\end{array}\right)\rho_A^{x_i}(1-\rho_A)^{10-x_i}\right]^{z_i}\left[\left(\begin{array}{c}10\\x_i\end{array}\right)\rho_B^{x_i}(1-\rho_B)^{10-x_i}\right]^{1-z_i}}{p_i^{z_i}(1-p_i)^{(1-z_i)}} \\ &= \sum_{i=1}^{5} p_i\left[x_i\log\rho_A+(10-x_i)\log(1-\rho_A)-\log p_i\right]+ \\ &\quad (1-p_i)\left[x_i\log\rho_B+(10-x_i)\log(1-\rho_B)-\log(1-p_i)\right]+\text{const}\,.\end{aligned}$$

(ii) We optimize $\mathcal{F}(\rho_A,\rho_B,q_1,\ldots,q_5)$ with respect to $\rho_A$and ρ_B. We obtain

$$\frac{\partial\mathcal{F}(\rho_A,\rho_B,q_1,\ldots,q_5)}{\partial\rho_A} = \sum_{i=1}^{5}\frac{p_i x_i}{\rho_A}-\frac{p_i(10-x_i)}{1-\rho_A}=0\Leftrightarrow\rho_A=\frac{\sum_{i=1}^{5}p_i x_i}{10\sum_{i=1}^{5}p_i}\,.$$

Similarly, we obtain

$$\frac{\partial\mathcal{F}(\rho_A,\rho_B,q_1,\ldots,q_5)}{\partial\rho_B} = \sum_{i=1}^{5}\frac{(1-p_i)x_i}{\rho_B}-\frac{(1-p_i)(10-x_i)}{1-\rho_B}=0\Leftrightarrow\rho_B=\frac{\sum_{i=1}^{5}(1-p_i)x_i}{10\sum_{i=1}^{5}(1-p_i)}\,.$$

(iii) We optimize $\mathcal{F}(\rho_A,\rho_B,q_1,\ldots,q_5)$ with respect to p_i. We obtain

$$\begin{aligned}\frac{\partial\mathcal{F}(\rho_A,\rho_B,q_1,\ldots,q_5)}{\partial p_i} &= x_i\log\rho_A+(10-x_i)\log(1-\rho_A)-x_i\log\rho_B- \\ &\quad (10-x_i)\log(1-\rho_B)-\log\frac{p_i}{1-p_i}\,,\end{aligned}$$

which is equal to zero if

$$p_i = \sigma\left(x_i\log\frac{\rho_A}{\rho_B}+(10-x_i)\log\frac{(1-\rho_A)}{(1-\rho_B)}\right)\,,$$

where $\sigma(x)=1/(1+\exp(-x))$.

Assessor's comments: One of the two hardest questions and the least popular one. Many students had problems deriving the update equations for the different steps (Estimation and Maximization) of the EM algorithm.

Exercise 4

a) Monte Carlo methods approximate expectations $\mathbb{E}_{p(x)}[f(x)]$ for an abitrary function $f(x)$ by drawing N samples $x_1,\ldots x_N$ from $p(x)$ and then computing an empirical average. In particular,

$$\mathbb{E}_{p(x)}[f(x)] \approx \frac{1}{N}\sum_{i=1}^{N}f(x_i)\,.$$

The main advantage of Monte Carlo methods is that they are unbiased, so the more samples we draw the more accurate that our approximation will be. However, the main disadvantage of these methods is that they can have a slow convergence and many samples will be necessary to obtain a small error in the approximation.

b)

(i) Rejection sampling works by sampling from a distribution $q(x)$ such that $Kq(x) \geq p(x)$ for $x \in [-1, 1]$. After drawing a sample x from $q(x)$, we sample a variable u uniformly between 0 and $Kq(x)$ and then accept the sample x if $u \leq p(x)$ and reject otherwise. To implement this, the first step is to obtain a form for $p(x)$. For this, have to normalize $\mathcal{N}(x|0, 1)$ in the interval $[-1, 1]$. The normalization constant is

$$Z = \int_{-1}^{1} \mathcal{N}(x|0, 1)\, dx = 1 - 2\int_{-\infty}^{-1} \mathcal{N}(x|0, 1)\, dx = 0.6826894\,.$$

Therefore, $p(x) = \mathcal{N}(x|0, 1)/0.6826894$. This density function attains its maximum at $x = 0$ with value $\mathcal{N}(0|0, 1)/0.6826894 = 0.5843686$. Since $q(x) = 0.5$ we have that $K = 0.5843686/0.5 = 1.168737$.

(ii) See previous point.

(iii) The acceptance probability is given by the ratio of the area of $p(x)$ and the area of $Kq(x)$ in $[-1, 1]$. In particular, it is qual to $1/K = 1/1.168737 = 0.855$.

c)

(i) The recursion equations are given by

$$\begin{aligned} p(Y_{1:T}) &= \sum_{X_T} p(Y_{1:T}, X_T) \\ p(Y_{1:T}, X_T) &= p(Y_T|X_T) \sum_{X_{T-1}} p(X_T|X_{T-1}) p(Y_{1:T-1}, X_{T-1})\,. \end{aligned}$$

(ii) Using the previous recurrsions, we must compute the following expressions to obtain $P(Y_1 =$

$A, Y_2 = B$). The solution only requires the last 6 lines from this list:

$$
\begin{aligned}
p(Y_1 = A, X_1) &= 0.7 \times 0.5(1 - X_1) + 0.3 \times 0.5X_1 \\
&= 0.35(1 - X_1) + 0.15X_1 \\
p(Y_1 = A, Y_2 = B, X_2) &= p(Y_2 = B|X_2) \sum_{X_1} p(X_2|X_1)p(Y_1 = A, X_1) \\
\sum_{X_1} p(X_2|X_1)p(Y_1 = A, X_1) &= \sum_{X_1} \left[X_1 0.8^{X_2} 0.2^{(1-X_2)} + (1 - X_1)0.2^{X_2} 0.8^{(1-X_2)} \right] \times \\
&\quad [0.35(1 - X_1) + 0.15X_1] \\
&= [X_2 0.19 + 0.31(1 - X_2)] \\
p(Y_1 = A, Y_2 = B, X_2) &= p(Y_2 = B|X_2)p(Y_1 = A, X_2) \\
&= p(Y_2 = B|X_2) \left[X_2 0.19 + 0.31(1 - X_2) \right] \\
&= \left[0.7X_2 + 0.3(1 - X_2) \right] \left[X_2 0.19 + 0.31(1 - X_2) \right] \\
&= 0.133X_2 + 0.093(1 - X_2) \\
P(Y_1 = A, Y_2 = B) &= \sum_{X_2} p(Y_1 = A, Y_2 = B, X_2) \\
&= 0.226 \, .
\end{aligned}
$$

Assessor's comments: One of the two most difficult questions in the exam. Many students had problems with the part on sequential models. The parts on Monte Carlo and sequential sampling were correctly addressed by most of them.