DD2447 - Statistical Methods in Applied Computer Science, Fall 2020

Assignment 1

Adriano Mundo

November 20, 2020

1 Graphical Models

1.1 Question 1

Pairwise independence does not imply mutual independence

As we know, for more than two events a mutually independent set of events is pairwise independent but the opposite is not necessarily true. An example is to consider X, Y, Z as three binary random variables, hence $X, Y, Z \in \{0, 1\}$.

We can suppose X and Y as two tosses of a fair coin with value 1 for head and 0 for tail, while Z is equal to 1 if one of the tosses is an head, 0 otherwise.

The joint pmf is:

$$P(X = x, Y = y, Z = z) = f(x, y, z) = 1/4 \tag{1.1.1}$$

while the marginal pmf are identical:

$$P(X = x) = P(Y = y) = P(Z = z) = f(x) = f(y) = f(z) = 1/2 \tag{1.1.2}$$

We can conclude that, even if X, Y, Z are pairwise independent (i.e. $f(x, y) = f(x)f(y)$), they are not mutual independent because $f(x, y, z) \neq f(x)f(y)f(z)$. Indeed, $f(x, y, z) = 1/4$ while $f(x)f(y)f(z) = 1/8$.

1.2 Question 2

Conditional independence iff joint factorizes

The alternative definition of conditional independence says that it is possible to have two functions $g(x, z)$ and $h(y, z)$ in place of the probability functions iff there is a conditional independence.

First half of the proof is:

$$p(x,y|z) = p(x|z)p(y|z) \implies p(x,y|z) = g(x,z)h(y,z) \tag{1.2.1}$$

It can be denoted that $p(x|z)$ is a function of the two varibales x, z; hence $p(x|z) = g(x,z)$. The same holds for the case of y, x, so we have $p(y|z) = h(y,z)$. In this way, we proved the first part of the formula.

Second half of the proof is:

$$p(x,y|z) = p(x|z)p(y|z) \impliedby p(x,y|z) = g(x,z)h(y,z) \tag{1.2.2}$$

Hypothesis: for the sake of simplicity, I will assume Discrete RV's. First of all, a join pmf must satisfy the following condition:

$$\sum_x \sum_y p(x,y|z) = 1 \tag{1.2.3}$$

It means that:

$$\sum_x \sum_y p(x,y|z) = \sum_x \sum_y g(x,z)h(y,z) = \sum_x g(x,z) \sum_y h(y,z) \tag{1.2.4}$$

$$\sum_x g(x,z) \sum_y h(y,z) = 1 \tag{1.2.5}$$

By deriving the marginal distribution of $p(x,y|z)$:

$$p(y|z) = \sum_x p(x,y|z) = \sum_x g(x,z)h(y,z) = h(y,z) \sum_x g(x,z) \tag{1.2.6}$$

$$p(x|z) = \sum_y p(x,y|z) = \sum_y g(x,z)h(y,z) = g(x,z) \sum_y h(y,z) \tag{1.2.7}$$

By moving the summations on the other member:

$$\sum_x g(x,z) = \frac{p(y|z)}{h(y,z)} \tag{1.2.8}$$

$$\sum_y h(y,z) = \frac{p(x|z)}{g(x,z)} \tag{1.2.9}$$

By combining the previous results:

$$\sum_x \sum_y p(x,y|z) = \sum_x g(x,z) \sum_y h(y,z) = \frac{p(y|z)}{h(y,z)} \frac{p(x|z)}{g(x,z)} = 1 \tag{1.2.10}$$

At the end:

$$p(x|z)p(y|z) = g(x,z)h(y,z) \tag{1.2.11}$$

Now, we have shown that $p(x|z) = g(x,z)$ and $p(y|z) = h(y,z)$. In this way, we proved the second part of the formula.

2 Generative Models

2.1 Question 3

Bayesian analysis of the exponential distribution

In this question, it will be performed a complete analysis for the *Exp* distribution, given that a lifetime X of a machine is modelled by an exponential distribution with unknown parameter θ and the likelihood is $p(x|\theta) = \theta e^{-\theta x}$ for $x \geq 0, \theta > 0$.

(a) Since it is easier and the results coincide, it is possible to compute the log of the likelihood function, $l(\theta) = log(p(\mathcal{D}|\theta))$, where $\mathcal{D}$ represents data.

$$l(\theta) = \sum_{i=1}^{N} log(\theta e^{-\theta x_i}) = \sum_{i=1}^{N} \Big(log(\theta) - \theta x_i \Big) = N log(\theta) - \theta \sum_{i=1}^{N} x_i \quad (2.1.1)$$

In order to obtain the *MLE* we need to do the first derivative and equals it to 0.

$$\frac{\partial l(\theta)}{\partial x} = N \frac{1}{\theta} - \sum_{i=1}^{N} x_i = 0 \quad (2.1.2)$$

$$\hat{\theta}_{MLE} = \frac{N}{\sum_{i=1}^{N} x_i} = \frac{1}{\bar{x}} \quad (2.1.3)$$

(b) Since our data are: $X_1 = 5, X_2 = 6, X_3 = 4$, the *MLE* can be calculated with the previous formula knowing that $\bar{x} = \sum_{i=1}^{N} x_i$.

$$\bar{x} = \frac{5+6+4}{3} = 5 \quad (2.1.4)$$

$$\hat{\theta}_{MLE} = \frac{1}{\bar{x}} = \frac{1}{5} = 0.2 \quad (2.1.5)$$

(c) The *Expon* distribution is a particular case of the *Gamma* distribution, hence $Expon(\theta|\lambda) = Ga(\theta|1, \lambda)$.
Therefore, we can calculate the expected value by using the formula of the *Gamma* distribution.

$$\mathbb{E}[\theta]_{Expon(\theta|\lambda)} = \mathbb{E}[\theta]_{Gamma(\theta|1,\lambda)} = \frac{1}{\lambda} = \frac{1}{3} \quad (2.1.6)$$

As a result, by reverting the (2.1.6), we calculate the prior parameter $\hat{\lambda}$.

$$\hat{\lambda} = 3 \quad (2.1.7)$$

(d) The posterior $p(\theta|\mathcal{D}, \hat{\lambda})$ can be calculated by using the prior and the likelihood.

The likelihood is equal to:

$$p(\mathcal{D}|\theta) = \prod_{i=1}^{N} \theta e^{-\theta x_i} = \theta^N e^{-\theta \sum_{i=1}^{N} x_i} \tag{2.1.8}$$

The prior is equal to:

$$p(\theta|\hat{\lambda}) = Expon(\theta|\hat{\lambda}) \propto e^{-\hat{\lambda}\theta} \tag{2.1.9}$$

The posterior can be computed as:

$$p(\theta|\mathcal{D}, \hat{\lambda}) \propto p(\mathcal{D}|\theta) \times p(\theta|\hat{\lambda}) \tag{2.1.10}$$

$$p(\theta|\mathcal{D}, \hat{\lambda}) \propto \theta^N e^{-\theta(\hat{\lambda}+\sum_{i=1}^{N} x_i)} \tag{2.1.11}$$

$$p(\theta|\mathcal{D}, \hat{\lambda}) = Ga\Big(\theta|N+1, \hat{\lambda} + \sum_{i=1}^{N} x_i\Big) = Ga(\theta|4, 18) \tag{2.1.12}$$

As we can notice from the final result, the posterior is a *Gamma* distribution with the parameter α, β equal to the one in the formula (2.1.12). Therefore, the final result is calculated by using the values of point (c).

(e) Yes, since prior and likelihood have the general form of a *Gamma* distribution. Indeed, the prior has distribution $p(\theta|\hat{\lambda}) = Ga(\theta|1, \hat{\lambda})$ while the likelihood is proportional to a *Gamma* distribution $p(\mathcal{D}|\theta) \propto Ga(\theta|N + 1, \sum_{i=1}^{N} x_i)$.

(f) The posterior mean $\mathbb{E}[\theta|\mathcal{D}, \hat{\lambda}]$ can be computed by using the mean of the *Gamma* distribution for what we have shown in (2.1.12).

$$\mathbb{E}[\theta|\mathcal{D}, \hat{\lambda}] = \frac{N+1}{\hat{\lambda} + \sum_{i=1}^{N} x_i} = \frac{4}{18} = 0,\bar{2} \tag{2.1.13}$$

(g) To show the difference between MLE and the posterior mean $\mathbb{E}[\theta|\mathcal{D}, \hat{\lambda}]$, I start from rewriting the posterior mean as:

$$\mathbb{E}[\theta|\mathcal{D}, \hat{\lambda}] = \Big(\frac{\sum_{i=1}^{N} x_i}{N+1} + \frac{\hat{\lambda}}{N+1}\Big)^{-1} \tag{2.1.14}$$

In this way, we have two terms: the first is the information derived from our dataset, while the second is our prior information.
If $\hat{\lambda} = 0$, the prior is uninformative and the posterior mean is almost equal to the MLE:

$$\mathbb{E}[\theta|\mathcal{D}, 0] = \frac{N+1}{\sum_{i=1}^{N} x_i} = \frac{N}{\sum_{i=1}^{N} x_i} + \frac{1}{\sum_{i=1}^{N} x_i} = \hat{\theta}_{MLE} + \frac{1}{\sum_{i=1}^{N} x_i} \tag{2.1.15}$$

Moreover, if N is huge, the posterior converges to the MLE since the second term relative to the prior goes to zero. Now that the difference has been shown, I would argue that in our case is better to use the posterior mean since from the point (c) we have an expert that gave as a informative prior, $\hat{\lambda} = 3$ and N is small, $|N| = 3$.

2.2 Question 4

Posterior predictive distribution for a batch of data with the dirichlet-multinomial model

The posterior predictive distribution for a single multinomial trial is:

$$p(X = j|D, \alpha) = \frac{\alpha_j + N_j}{\alpha + N} \tag{2.2.1}$$

To derive the final result, we have to consider the batch of data as a series of single trials:

$$p(\tilde{D}|D, \alpha) = p(\tilde{x_1}|D)p(\tilde{x_2}|\{D, \tilde{x_2}\})p(\tilde{x_3}|\{D, \tilde{x_1}\tilde{x_2}\})... \tag{2.2.2}$$

Now, it possible to use the (2.2.1) in (2.2.2), and by updating the number of empirical counts for the total amount and of the trial for each instance, we can:

$$\begin{aligned} p(\tilde{D}|D, \alpha) &= \frac{1}{\prod_{i=0}^{N-1}(N^{old} + \alpha + i)} \prod_{j=1}^{K} \prod_{i=o}^{N_i^{new}-1} \left(N_j^{old} + \alpha_j + i\right) = \\ &= \frac{\Gamma(N^{old} + \alpha)}{\Gamma(N + \alpha)} \prod_{j=1}^{K} \frac{\Gamma(N_j + \alpha_j)}{\Gamma(N_j^{old} + \alpha_j)} \end{aligned} \tag{2.2.3}$$

3 Bayesian Inference

3.1 Question 5

Bayesian Inference for the Univariate Normal

In this question, we start from the fact that in the standard form, the likelihood has two parameters, the mean μ and the variance σ^2, and we want to find conjugate priors distributions for these parameters.

$$p(x_1, x_2, ..., x_n|\mu, \sigma^2) \propto \frac{1}{\sigma^n} exp\Bigg(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\Bigg) \tag{3.1.1}$$

(a) 1^{st} case - *fixed variance* σ^2: in this case, by keeping the variance σ^2 fixed, the conjugate prior for μ is a *Gaussian*:

$$p(\mu|\mu_0, \sigma_0^2) \propto \frac{1}{\sigma_o} exp\Big(-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\Big) \tag{3.1.2}$$

To get the posterior $(\mu|x)$, we have to put together the prior (3.1.2) and the likelihood (3.1.1). To do that, we can notice that the joint distribution for x and μ is a *Gaussian* distribution and it is:

$$p(x, \mu|\sigma^2, \mu_0, \sigma_0^2) = p(x|\mu, \sigma^2)p(\mu|\mu_0, \sigma_0^2) \tag{3.1.3}$$

where, we previously assumed that $p(x|\mu, \sigma^2) = \mathcal{N}(x|\mu, \sigma^2)$ and $p(\mu|\mu_0, \sigma_0^2) = \mathcal{N}(\mu|\mu_0, \sigma_0^2)$. The prior has *Gaussian* distribution, and the important property of this prior is that it is conjugate to the *Gaussian* distribution used to model the probability. Thus, even the conditional $(\mu|x)$ is a *Gaussian* with parameters.

$$\begin{aligned}\mathbb{E}[\mu|x] &= \mathbb{E}[\mu] + \frac{Cov(\mu, x)}{Var(x)}(x-\mathbb{E}[x]) = \mu_0 + \frac{\sigma_0^2}{\sigma^2+\sigma_0^2}(x-\mu_0) = \\ &= \frac{\sigma_0^2}{\sigma^2+\sigma_0^2}x + \frac{\sigma^2}{\sigma^2+\sigma_0^2}\mu_0\end{aligned} \tag{3.1.4}$$

$$Var[\mu|x] = Var(\mu) - \frac{Cov^2(\mu, x)}{Var(x)} = \frac{\sigma^2\sigma_0^2}{\sigma^2+\sigma_0^2} = \frac{1}{\frac{1}{\sigma_0^2}+\frac{1}{\sigma^2}} \tag{3.1.5}$$

To summarise, assumed that $x|\mu \sim \mathcal{N}(\mu, \sigma^2)$ and $\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$, then:

$$\mu|x \sim \mathcal{N}\Bigg(\frac{\sigma_0^2}{\sigma^2+\sigma_0^2}x + \frac{\sigma^2}{\sigma^2+\sigma_0^2}\mu_0, \Big(\frac{1}{\sigma_0^2}+\frac{1}{\sigma^2}\Big)^{-1}\Bigg) \tag{3.1.6}$$

This is not concluded since the exercise shows the posterior for multiple measurements ($n \geq 1$). The best way to obtain the final result is to reduce the problem to the univariate case by using the empirical mean as the new variable, $\bar{x} = \frac{\sum x_i}{n}$. The likelihood becames:

$$x_i|\mu \sim \mathcal{N}(\mu, \sigma^2), iid \implies \bar{x}|\mu \sim \mathcal{N}\Bigg(\mu, \frac{\sigma^2}{n}\Bigg) \tag{3.1.7}$$

$$\begin{aligned} p(x_1, x_2, ..., x_n|\mu) &\propto \frac{1}{\sigma} exp\Bigg(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\Bigg) \\ &\propto exp\Bigg(-\frac{1}{2\sigma^2}\Bigg(\sum_{i=1}^{n} x_i^2 - 2\mu\sum_{i=1}^{n} x_i + n\mu^2\Bigg)\Bigg) \\ &\propto exp\Bigg(-\frac{n}{2\sigma^2}\big(-2\mu\bar{x}+\mu^2\big)\Bigg) \\ &\propto exp\Bigg(-\frac{n}{2\sigma^2}\big(\bar{x}-\mu\big)^2\Bigg) \propto p(\bar{x}|\mu) \end{aligned} \tag{3.1.8}$$

In the case of the posterior:

$$\begin{aligned} p(\mu|x_1, x_2, ..., x_n) &\propto p(x_1, x_2, ..., x_n|\mu)p(\mu) \propto p(\bar{x}|\mu)p(\mu) \\ &\propto p(\mu|\bar{x}) \end{aligned} \tag{3.1.9}$$

By inserting $\bar{x}$ in the previous result, and assuming $x_i|\mu \sim \mathcal{N}(\mu, \sigma^2), iid$ and $\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$, we obtain:

$$\mu|x \sim \mathcal{N}\Bigg(\frac{\sigma_0^2 n}{\sigma^2 + n\sigma_0^2}\bar{x} + \frac{\sigma^2}{\sigma^2+n\sigma_0^2}\mu_0, \Bigg(\frac{1}{\sigma_0^2}+\frac{1}{\sigma^2}\Bigg)^{-1}\Bigg) \tag{3.1.10}$$

(b) 2^{nd} case - *fixed mean* μ: in this case, by keeping the mean μ fixed, the conjugate prior for σ^2 is an *Inverse Gamma* distribution. Given that $z|\alpha,\beta \sim \mathcal{IG}(\alpha,\beta)$, we can say that:

$$p(z|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} z^{\alpha-1} exp\Bigg(-\frac{\beta}{z}\Bigg) \tag{3.1.11}$$

The posterior is an *Inverse Gamma* distribution too:

$$\begin{aligned} p(\sigma^2|\alpha,\beta) &\propto (\sigma^2)^{-\left(\alpha+\frac{n}{2}\right)^{-1}} exp\left(-\frac{\beta+\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^2}\right) \\ &\propto (\sigma^2)^{-\alpha_{post}-1} exp\left(-\frac{\beta_{post}}{\sigma^2}\right) \end{aligned} \tag{3.1.12}$$

As we have shown in (3.1.12), by having an *Inverse Gamma* prior (3.1.11), also the posterior is an *Inverse Gamma*, hence if we assume $x_i|\mu,\sigma^2 \sim \mathcal{N}(\mu,\sigma^2), iid$ and $\sigma^2 \sim \mathcal{IG}(\alpha,\beta)$:

$$\sigma^2|x_1, x_2, ..., x_n \sim \mathcal{IG}\left(\alpha+\frac{n}{2}, \beta+\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right) \tag{3.1.13}$$

(c) 3^{rd} case - *mean μ and variance σ^2 unknown*: in this case we need a prior on both parameters together. If we use the results from (a) and (b) we do not obtain a conjugate prior. The assumption for obtaining the conjugate prior is to believe that, conditioned on x, the two parameters μ and σ^2 are dependent and this should be expressed by a conjugate prior.
By using the prior distribution provided from the text, the result is conjugate to the *Gaussian* likelihood.

The first term to consider is $\mu|x, \tau$. Since we know that $x_i|\mu, \tau \sim \mathcal{N}(\mu, \tau^{-1}), iid$ and $\tau \sim Ga(\alpha, \beta)$, then:

$$\mu|x, \tau \sim \mathcal{N}\left(\frac{n\tau}{n\tau + n_0\tau}\bar{x} + \frac{n_0\tau}{n\tau + n_0\tau}\mu_0, (n\tau + n_0\tau)^{-1}\right) \tag{3.1.14}$$

The second term is $\tau|x$ and it can be expressed as:

$$\begin{aligned}
&p(\tau, \mu|x) \propto p(\tau) \cdot p(\mu|\tau) \cdot p(x|\tau, \mu) \\
&\propto \tau^{\alpha-1}e^{-\beta\tau}\tau^{\frac{1}{2}}exp\left(-\frac{n_0\tau}{2}(\mu - \mu_0)^2\right)\tau^{\frac{n}{2}}exp\left(-\frac{\tau}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right) \\
&\text{by doing: } x_i - \bar{x} + \bar{x} - \mu, \text{ we obtain:} \\
&\propto \tau^{\alpha+\frac{n}{2}-1}exp\left(-\tau\left(\beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)\right)\tau^{\frac{1}{2}} \times \\
&exp\left(-\frac{\tau}{2}(n_o(\mu - \mu_0)^2 + n(\bar{x} - \mu)^2)\right)
\end{aligned} \tag{3.1.15}$$

By integrating on μ we can obtain the normalization constant:

$$\tau^{-\frac{1}{2}}exp\left(\frac{nn_0\tau}{2(n + n_0)}(\bar{x} - \mu_0)^2\right) \tag{3.1.16}$$

By leveraging on the (3.1.16), we obtain the *Gamma* posterior for τ:

$$p(\tau|x) \propto \tau^{\alpha+\frac{n}{2}-1}exp\left(-\tau\left(\beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{nn_0}{2(n + n_0)}(\bar{x} - \mu_0)^2\right)\right) \tag{3.1.17}$$

At the end, if we assume $x_i|\mu, \tau \sim \mathcal{N}(\mu, \tau^{-1}), iid$, and $\mu|\tau \sim \mathcal{N}(\mu_0, (n_0\tau)^{-1})$, and $\tau \sim Ga(\alpha, \beta)$, we have as posterior result the (3.1.14) shown before and:

$$\tau|x \sim Ga\left(\alpha + \frac{n}{2}, \beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{nn_0}{2(n + n_0)}(\bar{x} - \mu_0)^2\right) \tag{3.1.18}$$

3.2 Question 6

Bayer factor for coin tossing

In this question it will be performed hypothesis testing for a coin tossing problem using summary statistics N, N_1. As we know, in Bayesian statistics to test an hypothesis we need the Bayes Factor (BF), that can be calculated as:

$$BF_{1,0} = \frac{p(\mathcal{D}|M_1)}{p(\mathcal{D}|M_0)} \tag{3.2.1}$$

with $\mathcal{D}$ data and M_0, M_1 null and alternative hypothesis.

The marginal likelihood for the alternative hypothesis is:

$$p(\mathcal{D}|M_1) = p(N_1|N) = \frac{Bin(N_1|N,\theta)Beta(\theta|1,1)}{Beta(\theta|N_1+1, N-N_1+1)} \tag{3.2.2}$$

It is a marginalization, hence it does not depend on θ.

$$\begin{aligned} p(\mathcal{D}|M_1) = p(N_1|N) = \frac{\binom{N}{N1}B(N_1+1, N-N_1+1)}{B(1,1)} = \\ = \frac{N!}{(N-N_1)!N_1!}\frac{\Gamma(N_1+1)\Gamma(N-N_1+1)}{\Gamma(N+2)} = \frac{1}{N+1} \end{aligned} \tag{3.2.3}$$

The marginal likelihood for the null hypothesis with the fair coin assumption:

$$\begin{aligned} p(\mathcal{D}|M_0) = \int_0^1 p(\mathcal{D}|\theta)p(\theta|M_0)d\theta = p(\mathcal{D}|\theta = 0.5) = \\ \binom{N}{N_1} = 0.5^{N_1}0.5^{N-N_1} = \binom{N}{N_1}0.5^N \end{aligned} \tag{3.2.4}$$

Now, we can calculate the Bayes Factor as:

$$BF_{1,0} = \frac{1}{N+1}\frac{1}{\binom{N}{N_1}0.5^N} = \frac{2^N}{(N+1)\binom{N}{N_1}} \tag{3.2.5}$$

By substituting the letters with the values provided from the questions, we have:

1. $N = 10, N_1 = 9$

$$BF_{1,0} = \frac{2^{10}}{(10+1)\binom{10}{9}} = 9.31 \tag{3.2.6}$$

2. $N = 100, N_1 = 90$

$$BF_{1,0} = \frac{2^{100}}{(100+1)\binom{100}{90}} = 7.2 \times 10^{14} \tag{3.2.7}$$

In both case 1 and case 2 we prefer M_1 over M_0. As conclusion, we can state that the high proportion of heads favoured the results for the alternative hypothesis over the null hypothesis; or, that the coin can have any bias. Finally, we can notice that the second result has a stronger preference, which means that having more samples help at identifying the best model.

4 Mixture and MLE

4.1 Question 7

Proof that a mixture of conjugate priors is indeed conjugate

We want to derive the formula:

$$p(\theta|\mathcal{D}) = \sum_k p(z=k|\mathcal{D})p(\theta|\mathcal{D}, z=k) \tag{4.1.1}$$

that can be expressed as since it is the posterior:

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} \tag{4.1.2}$$

Now, we derive the prior $p(\theta)$ as a mixture of models:

$$p(\theta) = \sum_k p(z=k)p(\theta|z=k) \tag{4.1.3}$$

and the $p(\mathcal{D}|\theta)$ by the fact that if we know θ there's no advantage in knowing the variable z, hence:

$$p(\mathcal{D}|\theta) = p(\mathcal{D}|\theta, z=k) \tag{4.1.4}$$

By substituting the result from (4.1.3) and (4.1.4) in (4.1.2):

$$p(\theta|\mathcal{D}) = \sum_k \frac{p(z=k)}{p(\mathcal{D})}p(\theta|z=k)p(\mathcal{D}|\theta, z=k) \tag{4.1.5}$$

To reach the final result, we have to take into account that:

$$p(\mathcal{D}|\theta, z=k) = \frac{p(\theta|\mathcal{D}, z=k)p(\mathcal{D}|z=k)}{p(\theta|z=k)} \tag{4.1.6}$$

By inserting the (4.1.6) in (4.1.5), we reach the expected result (4.1.1):

$$p(\theta|\mathcal{D}) = \sum_k \frac{p(z=k)p(\mathcal{D}|z=k)}{p(\mathcal{D})}p(\theta|\mathcal{D}, z=k) = \sum_k p(z=k|\mathcal{D})p(\theta|\mathcal{D}, z=k) \tag{4.1.7}$$

We derived the posterior with a mixture of priors and we end by saying that the posterior for a mixture of priors is a mixture of those priors posteriors.

4.2 Question 8

MLE and model selection for a 2d discrete distribution

(a) We have to calculate the joint probability distribution $p(x, y|\boldsymbol{\theta})$, the results are summarised by the following table:

$p(x, y\|\boldsymbol{\theta})$	$y = 0$	$y = 1$
$x = 0$	$(1 - \theta_1)\theta_2$	$(1 - \theta_1)(1 - \theta_2)$
$x = 1$	$\theta_1(1 - \theta_2)$	$\theta_1\theta_2$

(b) The MLE can be calculated as:

$$\hat{\theta}_{MLE} = argmax\left(Nlog\left(\frac{1-\theta_1}{1-\theta_2}\right) + N_x log\left(\frac{\theta_1}{1-\theta_1}\right) + N_{\mathbb{I}(x=y)} log\left(\frac{\theta_2}{1-\theta_2}\right)\right) \tag{4.2.1}$$

Thus, the calculation of the MLE for the two parameters gives:

$$\hat{\theta}_{1,MLE} = \frac{4}{7}, \hat{\theta}_{2,MLE} = \frac{4}{7} \tag{4.2.2}$$

The probability $p(\mathcal{D}|\hat{\theta}, M_2)$ with 2-parameters model can be calculated as:

$$\begin{aligned} p(\mathcal{D}|\hat{\theta}, M_2) &= \left(\frac{4}{7}\right)^4 \left(1 - \frac{4}{7}\right)^{7-4} \left(\frac{4}{7}\right)^4 \left(1 - \frac{4}{7}\right)^{7-4} = \left(\frac{4}{7}\right)^8 \left(\frac{3}{7}\right)^6 = \\ &= \frac{16}{49} \cdot \frac{16}{49} \cdot \frac{16}{49} \cdot \frac{16}{49} \cdot \frac{9}{49} \cdot \frac{9}{49} \cdot \frac{9}{49} \approx 0.0000704 \end{aligned} \tag{4.2.3}$$

(c) In this case we have a model with 4-parameters, $\boldsymbol{\theta} = (\theta_{0,0}, \theta_{0,1}, \theta_{1,0}, \theta_{1,1})$ with 3 parameters free to vary.

The MLE of $\boldsymbol{\theta}$ calculation has the following results:

$$\hat{\theta}_{0,0,MLE} = \frac{2}{7}, \hat{\theta}_{0,1,MLE} = \frac{1}{7}, \hat{\theta}_{1,0,MLE} = \frac{2}{7}, \hat{\theta}_{1,1,MLE} = \frac{2}{7} \tag{4.2.4}$$

The probability $p(\mathcal{D}|\hat{\theta}, M_4)$ with 4-parameters model can be calculated as:

$$\begin{aligned} p(\mathcal{D}|\hat{\theta}, M_4) &= \left(\frac{2}{7}\right)^2 \left(\frac{1}{7}\right)^1 \left(\frac{2}{7}\right)^2 \left(\frac{2}{7}\right)^2 = \left(\frac{2}{7}\right)^6 \frac{1}{7} = \\ &= \frac{2}{7} \cdot \frac{2}{7} \cdot \frac{2}{7} \cdot \frac{2}{7} \cdot \frac{2}{7} \cdot \frac{2}{7} \cdot \frac{1}{7} \approx 0.0000777 \end{aligned} \tag{4.2.5}$$

(d) By using the formula provided for leave-one-out cross validation, we can compute the two cases.

$$L(m) = \sum_{i=1}^{n} logp(x_i y_i|m, \hat{\theta}(\mathcal{D}_{-i})) \tag{4.2.6}$$

For the 2-parameter model, we have:

$$L(M_2) = log\left(\frac{3}{6}\cdot\frac{3}{6}\cdot\frac{3}{6}\cdot\frac{2}{6}\cdot\frac{2}{6}\cdot\frac{3}{6}\cdot\frac{3}{6}\cdot\frac{2}{6}\cdot\frac{3}{6}\cdot\frac{3}{6}\cdot\frac{2}{6}\cdot\frac{2}{6}\cdot\frac{2}{6}\cdot\frac{2}{6}\right) \approx -12.5423 \tag{4.2.7}$$

Instead, for the 4-parameter model:

$$L(M_4) = -\infty \text{ since } logp(x_6, y_6|m, \hat{\theta}(\mathcal{D}_{-6})) = log0 \tag{4.2.8}$$

As we can see from the previous results, the CV will pick the model M_2. We do not have to use training data for model selection otherwise the more complex model always wins, which it is not a surprise.

(e) In this case, we again use the formula provided in the exercise for computing the BIC score.

$$BIC(M, \mathcal{D}) = logp(\mathcal{D}|\hat{\boldsymbol{\theta}}_{MLE}) - \frac{dof(M)}{2}logN \tag{4.2.9}$$

For the 2-parameter model, we have:

$$BIC(M_2, \mathcal{D}) \approx -9.561 - log7 \approx -11.507 \tag{4.2.10}$$

Instead, for the 4-parameter model:

$$BIC(M_4, \mathcal{D}) \approx -12.381 \tag{4.2.11}$$

As result, the first model is preferred by using the BIC score.

5 Numerical Problems

5.1 Question 9

Doctor and patient disease probability

We can start by summarising all the information provided from the exercise.

First of all, we have equal probability for each disease:

$$p(d_1) = p(d_2) = p(d_3) = p \tag{5.1.1}$$

Instead, the results from the test. I will indicate with the letter i the positive result of the test, hence:

$$p(i|d_1) = 0.8 \tag{5.1.2}$$

$$p(i|d_2) = 0.6 \tag{5.1.3}$$

$$p(i|d_3) = 0.4 \tag{5.1.4}$$

The question asks to calculate the three probabilities given that the outcome was positive, $p(d_1|i)$, $p(d_2|i)$, $p(d_3|i)$.

To compute the previous probabilities, we need $p(i)$:

$$p(i) = p(i|d_1)p(d_1) + p(i|d_2)p(i) + p(i|d_3)p(i) = 0.8p + 0.6p + 0.4p = 1.8p \tag{5.1.5}$$

Now, we are able to compute the requested values by using the conditional probability formula:

$$p(d_1|i) = \frac{p(d_1 \cap i)}{p(i)} = \frac{0.8p}{1.8p} = 0.\bar{4} \tag{5.1.6}$$

$$p(d_2|i) = \frac{p(d_2 \cap i)}{p(i)} = \frac{0.6p}{1.8p} = 0.\bar{3} \tag{5.1.7}$$

$$p(d_3|i) = \frac{p(d_3 \cap i)}{p(i)} = \frac{0.4p}{1.8p} = 0.\bar{2} \tag{5.1.8}$$

5.2 Question 10

Posterior predictive for Dirichlet-multinomial

In this question, by leveraging on the posterior predictive of the *Dir* in the multinomial case, we try to predict the next character in a sequence, where the distribution is over 27 values from 2000 samples.

(a) We can compute the requested $p(x_{2001} = e|\mathcal{D})$ given that e has been seen 260 times:

$$p(x_{2001} = e|\mathcal{D}) = \frac{\alpha_e + N_e}{\alpha + N} = \frac{10 + 260}{270 + 2000} = \frac{270}{2270} \approx 0.1189 = 11.9\% \tag{5.2.1}$$

(b) Now, it is requested to compute the $p(x_{2001} = p, x_{2002} = a|\mathcal{D})$, knowing that we saw e 260 times, a 100 times, and p 87 times.

$$p(x_{2001} = p, x_{2002} = a|\mathcal{D}) = \frac{10 + 87}{270 + 2000}\frac{10 + 100}{270 + 2001} = \frac{97 \times 110}{2270 \times 2271} \approx 0.00207 = 0.207\% \tag{5.2.2}$$