DD2447 - Assignment 1

Alexandra Hotti

25 November 2018

1 2.1 Basic Probability

1.1 Question 1

Pairwise independence does not imply mutual independence
We say that two random variables are pairwise independent if

$$p(X_2|X_1) = p(X_2) \tag{1}$$

and hence

$$p(X_2, X_1) = p(X_1)p(X_2|X_1) = p(X_1)p(X_2) \tag{2}$$

We say that n random variables are mutually independent if

$$p(X_i|X_S) = p(X_i) = \forall S \subseteq \{1, ..., n\} \setminus \{i\} \tag{3}$$

and hence

$$p(X_{1:n}) = \prod_{i=1}^{n} p(X_i) \tag{4}$$

Show that pairwise independence between all pairs of variables does not necessarily imply mutual independence. It suffices to give a counter example.
Solution:
A counter example is provided.
Assume that there is 2 independent random variables X and Y which outcomes are binary. Additionally there is a third binary random variable Z which is equal to 1 if one and only one of the outcomes of X and Y is equal to 1.
We can illustrate the joint probability distribution over (X,Y,Z) with the table below:

(X,Y,Z)	**Joint probability**
(X=0, Y=0, Z=0)	0.25
(X=1, Y=1, Z=0)	0.25
(X=1, Y=0, Z=1)	0.25
(X=0, Y=1, Z=1)	0.25

We see from the table that:

$$p_X(0) = p_X(1) = p_Y(0) = p_Y(1) = p_Z(0) = p_Z(1) = 0.5 \tag{5}$$

Now we want to consider the pairwise joint probabilities. We obtain these probabilities by marginalizing out the third variable. Which for instance means that: $p(X = 1, Y = 1) = p(X = 1, Y = 1, Z = 0) + p(X = 1, Y = 1, Z = 1) = 0.25 + 0$.
Doing this for all values for all pairs of random variables yields us:

$$p(X, Y) = p(X, Z) = p(Y, Z) \tag{6}$$

Where:

$$p(X = 0, Y = 0) = p(X = 1, Y = 0) = p(X = 0, Y = 1) = p(X = 1, Y = 1) = 0.25$$

Now we can show that pairwise independence exists since each pairwise joint distributions is equal to the multiplication of its corresponding marginal distributions (2).For instance

$$p(X = 0, Y = 0) = p_X(0)p_Y(0) = 0.5^2 = 0.25$$

This then holds for all pairwise combinations of X, Y and Z due to the equalities in (6) and (5). Meaning that:

$$p(X|Y) = p(X), p(Y|X) = p(Y), p(Z|X) = p(Z), p(X|Z) = p(X), p(Z|Y) = p(Z), p(Y|Z) = p(Y)$$

and thus they are pairwise independent.

Moving over to mutual Independence, from (4) mutual independence implies that:

$$p(X, Y, Z) = p(X)p(Y)p(Z) \tag{7}$$

for all outcomes of X, Y and Z.

We can prove that (7) does not hold in this example with an example outcome. First we look at the left hand side of (7).
Consider the outcome $(X = 0, Y = 0, Z = 0)$ which according to the table of joint probabilities has the probability $p(X = 0, Y = 0, Z = 0) = 0.25$

Now, looking at the right hand side, according to (5)

$$p(X = 0)p(Y = 0)p(Z = 0) = 0.5 \cdot 0.5 \cdot 0.5 = 0.125$$

Thus

$$p(X = 0, Y = 0, Z = 0) \neq p(X = 0)p(Y = 0)p(Z = 0)$$

Meaning that pairwise independence between all pairs of random variables does not imply mutual independence between the same random variables.

1.2 Question 2

In the text we said $X \perp Y|Z$ iff

$$p(x, y|z) = p(x|z)p(y|z) \tag{8}$$

for all x, y, z such that $p(z) > 0$.

Now prove the following alternate definition: $X \perp Y|Z$ iff there exist functions g and h such that

$$p(x, y|z) = g(x, z)h(y, z) \tag{9}$$

for all x, y, z such that $p(z) > 0$.

Proof:

By marginalizing over x and y we obtain the following

$$\int_y \int_x p(x, y|z)dxdy = \int_y \int_x g(x, z)h(y, z)dxdy = \int_x g(x, z)dx \int_y h(y, z)dy \tag{10}$$

By marginalizing out x and y and using the law of total probability we obtain the following

$$\int_y \int_x p(x, y|z)dxdy = 1 \tag{11}$$

Since the left hand side of expression (11) and (10) are equal we can combine the expressions into the following

$$\int_x g(x, z)dx \int_y h(y, z)dy = 1 \tag{12}$$

Marginalizing out x from $p(x, y|z)$ in (9) gives us an expression for $p(y|z)$

$$p(y|z) = \int_x p(x, y|z)dx = \int_x g(x, z)h(y, z)dx = h(y, z) \int_x g(x, z)dx \tag{13}$$

Which we can rewrite as the following

$$\frac{p(y|z)}{h(y, z)} = \int_x g(x, z)dx \tag{14}$$

Marginalizing out y from $p(x, y|z)$ in (9) gives us an expression for $p(x|z)$

$$p(x|z) = \int_y p(x, y|z)dy = \int_y g(x, z)h(y, z)dy = g(x, z) \int_y h(y, z)dy \tag{15}$$

Which we can rewrite as the following

$$\frac{p(x|z)}{g(x, z)} = \int_y h(y, z)dy \tag{16}$$

Multiplying (14) and (17) gives us

$$\frac{p(x|z)}{g(x,z)}\frac{p(y|z)}{h(y,z)} = \int_x g(x,z)dx \int_y h(y,z)dy \tag{17}$$

By using equation (1) and (2) we arrive at

$$\frac{p(x|z)}{g(x,z)}\frac{p(y|z)}{h(y,z)} = 1 \tag{18}$$

Which can be rewritten as

$$p(x|z)p(y|z) = h(y,z)g(x,z)$$

Thus we infer that $h(y,z) = p(y|z)$ as they both are functions in terms of y and z and that $g(x,z) = p(x|z)$ as they both are functions in terms of x and z. Finally combining this with (8) gives us

$$p(x,y|z) = p(x|z)p(y|z) = h(y,z)g(x,z) \Leftrightarrow X \perp Y|Z$$

Q.E.D.

1.3 Question 3

The poisson pmf is defined as $Poi(x|\lambda) = e^{-\lambda}\frac{\lambda^x}{x!}$ for x $\in \{0,1,2,..\} where \lambda > 0$ is the rate parameter. Derive the MLE.

Solution:

Assuming that $x_1, x_2, ..$ are iid random variables, the likelihood function for the Poisson pmf is given by the following expression:

$$L(\lambda) = \prod_{i=j}^{N} e^{-\lambda}\frac{\lambda^{x_j}}{x_j!} \tag{19}$$

However, instead of finding the λ which maximizes the likelihood of the data we compute the log likelihood of the data and maximize this expression instead with respect to λ. This is computationally easier and is valid since the logarithm is a monotonically increasing function. Thus we compute the logarithm of (19) below:

$$\mathrm{l}(\lambda) = \sum_{j=1}^{N} log(e^{-\lambda}\frac{\lambda^{x_j}}{x_j!}) = -\sum_{j=1}^{N}\lambda + log(\lambda)\sum_{j=1}^{N}x_j - \sum_{j=1}^{N} log(x_j!) =$$

$$= -N\lambda + log(\lambda)\sum_{j=1}^{N}x_j - \sum_{j=1}^{N} log(x_j!) \tag{20}$$

Next we want to find λ_{MLE}, i.e. the λ which maximizes the derivative of (20) with respect to λ

$$\frac{dl}{d\lambda} = -N + \frac{\sum_{j=1}^{N} x_j}{\lambda} \tag{21}$$

Setting $\frac{dl}{d\lambda} = 0$ and solving for λ finally gives us the maximum likelihood estimation:

$$\lambda_{MLE} = \frac{\sum_{j=1}^{N} x_j}{N} \tag{22}$$

1.4 Question 4

a. Derive the posterior $p(\lambda|D)$ assuming a conjugate prior $p(\lambda) = Ga(\lambda|a, b) \propto \lambda^{a-1}e^{-\lambda b}$. Hint: The posterior is also a Gamma distribution.

b.What does the posterior mean tend to as $a \to 0$ and $b \to 0$? (Recall that the mean of a Ga(a,b) distribution is $\frac{a}{b}$)
Solution to part a:
Using Bayes Theorem the posterior can be rewritten as:

$$p(\lambda|D) = \frac{p(D|\lambda)p(\lambda)}{p(D)} \tag{23}$$

Since the denominator in (23) is a constant factor we look at the following equivalent expression:

$$\frac{p(D|\lambda)p(\lambda)}{p(D)} \propto p(D|\lambda)p(\lambda) \tag{24}$$

Assuming that the data is iid $Poisson(\lambda)$, the likelihood is given by the following expression:

$$p(D|\lambda) = \prod_{j=1}^{N} e^{-\lambda}\frac{\lambda^{x_j}}{x_j!} \tag{25}$$

From the assignment description we know that the conjugate prior $p(\lambda)$ is equivalent to

$$p(\lambda) \propto \lambda^{a-1}e^{-\lambda b} \tag{26}$$

Thus we can compute the posterior from expression by using and

$$p(\lambda|D) \propto \prod_{j=1}^{N}(e^{-\lambda}\frac{\lambda^{x_j}}{x_j!})\lambda^{a-1}e^{-\lambda b} = e^{-N\lambda}\prod_{j=1}^{N}\frac{\lambda^{x_j}}{x_j!}\lambda^{a-1}e^{-\lambda b} =$$

$$e^{-N\lambda}\frac{\lambda^{\sum_{j=1}^{N} x_j}}{\prod_{j=1}^{N} x_j!}\lambda^{a-1}e^{-\lambda b} \tag{27}$$

Since $\prod_{j=1}^{N} x_j!$ is a constant factor we can use the equivalent expression:

$$e^{-N\lambda}\frac{\lambda^{\sum_{j=1}^{N} x_j}}{\prod_{j=1}^{N} x_j!}\lambda^{a-1}e^{-\lambda b} \propto e^{-N\lambda}\lambda^{\sum_{j=1}^{N} x_j}\lambda^{a-1}e^{-\lambda b} \tag{28}$$

Next we collect the terms

$$e^{-N\lambda}\lambda^{\sum_{j=1}^{N} x_j}\lambda^{a-1}e^{-\lambda b} = e^{-\lambda(N+b)}\lambda^{a-1+\sum_{j=1}^{N} x_j} \propto Gamma(\lambda|a+\sum_{j=1}^{N} x_j, b+N) \tag{29}$$

Thus, we have derived the posterior. Note that since we have used a conjugate prior the posterior is also a Gamma distribution.

Solution to part b:

The mean of $Ga(a,b) = \frac{a}{b}$ according to the assignment description. Thus, as our posterior distribution has the form $Ga(\lambda|a+\sum_{j=1}^{N} x_j, b+N)$, its posterior mean can be computed as

$$E[\lambda|D] = \frac{a+\sum_{j=1}^{N} x_j}{b+N} \tag{30}$$

Then we let a and b go towards zero:

$$E[\lambda|D]_{lim_{a,b\to 0}} = \frac{\sum_{j=1}^{N} x_j}{N} \tag{31}$$

Which is the the expression for the λ_{MLE}, see (22).

1.5 Question 5

Posterior predictive distribution for a batch of data with the dirichlet-multinomial model In Equation 3.51 (Murphy, p.83) , we gave the the posterior predictive distribution for a single multinomial trial using a dirichlet prior. Now consider predicting a batch of new data, $\widetilde{D} = (x_1, ..., x_m)$, consisting of m single multinomial trials (think of predicting the next m words in a sentence, assuming they are drawn iid). Derive an expression for

$$p(\widetilde{D}|D, \alpha) \tag{32}$$

Your answer should be a function of α, and the old and new counts (sufficient statistics), dened as

$$N_k^{old} = \sum_{i\in D} I(x_i = k) \tag{33}$$

$$N_k^{new} = \sum_{i\in \tilde{D}} I(x_i = k) \tag{34}$$

Hint: recall that, for a vector of counts, $N_{1:K}$, the marginal likelihood (evidence) is given by

$$p(D|\alpha) = \frac{\Gamma(\alpha)}{\Gamma(N+\alpha)} \prod_k \frac{\Gamma(N_k + \alpha_k)}{\Gamma(\alpha_k)} \tag{35}$$

where $\alpha = \sum_k \alpha_k$ and $N = \sum_k N_k$

Solution:

First, if we start by conceptually thinking about what a prior actually is. For instance think about a team going into a new football season. If a player, that has performed really good for 5 seasons, does not score during his first couple of games in a new season we will not predict that his average score will be equal to 0 for the entire season. Instead we also use a prior which represents his performance during the previous seasons. Therefore the prior represents our previous expectations regarding the players performance and becomes the only initial information that we use regarding how the player will score before the new season has begun. But as the season moves along we obtain more current data regarding the player, therefore we want to update our predictions about how the player will perform for the rest of the season. Therefore we must update the prior with this new data.

Therefore, the prior is our information going into the season and the posterior is our updated information according to new additional data. Thereby it makes sense that the shape of the prior corresponds to the shape of the posterior. Such as in the case of the conjugate prior.

Now if we use that the posterior of the original data can be expressed as:

$$p(\theta|D, \alpha) = Dir(\theta|\alpha + N^{old}) \tag{36}$$

Thus as it has the same form as the prior, we can simply update (35) with the new counts, i.e. using new hyper parameters, such that $\alpha_j^{new} = N_k^{old} + \alpha_k$ and with the additional data $\tilde{D}$. Note that old refers to the data set D and $\tilde{D}$ refers to the new batch of data. Which gives us:

$$p(\tilde{D}|D, \alpha) = p(\tilde{D}|\alpha) = \frac{\Gamma(\alpha^{new})}{\Gamma(N^{new} + \alpha^{new})} \prod_k \frac{\Gamma(N_k^{new} + \alpha_k^{new})}{\Gamma(\alpha_k^{new})} =$$

$$= \frac{\Gamma(\alpha + N^{old})}{\Gamma(N+\alpha)} \prod_k \frac{\Gamma(N_k + \alpha_k)}{\Gamma(\alpha_k + N_k^{old})} \tag{37}$$

1.6 Question 6

a. Suppose we compute the empirical distribution over letters of the Roman alphabet plus the space character (a distribution over 27 values) from 2000 samples. Suppose we see the letter "e" 260 times. What is $p(x_{2001} = e|D)$, if we

assume $\theta \sim Dir(\alpha_1, ..., \alpha_{27})$, where $\alpha_k = 10$ for all k?
b. Suppose, in the 2000 samples, we saw "e" 260 times, "a" 100 times, and "p" 87 times. What is $p(x_{2001} = p, x_{2002} = a|D)$, if we assume $\theta \sim Dir(\alpha_1, ..., \alpha_{27})$, where $\alpha_k = 10$ for all k?

Solution part a:
The posterior predictive for the dirichlet-multinomial model over one multinoulli trial can be computed from equation (3.49-3.51) (Murphy, p.83):

$$p(x = j|D) = E[\theta_k|D] = \frac{\alpha_j + N_j}{\alpha + N} \tag{38}$$

First from the assignment description we get:

$$\alpha_k = 10 \tag{39}$$

Furthermore, from the assignment description we get the prior $\theta \sim Dir(\alpha_1, ..., \alpha_{27})$. Thus we can compute α:

$$\alpha = \sum_{k}^{27} \alpha_k = 270 \tag{40}$$

Next, N_k is given by the number of times we see the letter e, i.e:

$$N_k = 260 \tag{41}$$

Lastly the total number of samples is also given in the assignment as:

$$N = 2000 \tag{42}$$

By plugging in values (39)-(42) into equation (38) we can compute $p(x_{2001} = e|D)$

$$p(x_{2001} = e|D) = \frac{\alpha_j + N_j}{\alpha + N} = \frac{10 + 260}{270 + 2000} = 0.11894 \approx 0.12$$

Solution part b:
When predicting multiple new samples we can use equation (37) from the previous question:

$$p(D_{new}|D_{old}, \alpha) = \frac{\Gamma(N + \alpha)}{\Gamma(M + N + \alpha)} \prod_k \frac{\Gamma(M_k + N_k + \alpha_k)}{\Gamma(N_k + \alpha_k)} =$$

$$\frac{(N + \alpha - 1)!}{(M + N + \alpha - 1)!} \prod_k \frac{(M_k + N_k + \alpha_k - 1)!}{(N_k + \alpha_k - 1)!}$$

Where M is the total number of new observations, N is the number of old observations, M_k is the count for each new type of sample and N_k is the count

for each old type of sample. Meaning that for a $M_a = 1$ and $N_a = 100$, and that for p $M_p = 1$ and $N_p = 87$. Then the product in (37) is computed over the new new samples.

Like in part a α can be computed as

$$\alpha = \sum_{k}^{27} \alpha_k = 270$$

(6.3)

M is given by the number of new samples, thus:

$$M = 2$$

Lastly N is given by the count of the old samples:

$$N = 2000$$

Lastly we put these numbers for M,N and α into equation (37) to obtain $p(x_{2001} = p, x_{2002} = a|D)$:

$$p(x_{2001} = p, x_{2002} = a|D) =$$

$$\frac{(2000 + 270 - 1)!}{(2 + 2000 + 270 - 1)!}(\frac{(1 + 100 + 10 - 1)!}{(100 + 10 - 1)!})(\frac{(1 + 87 + 1 - 1)!}{(87 + 10 - 1)!}) =$$

$$(\frac{2269!}{2271!})(\frac{110!}{109!})(\frac{97!}{96!}) = (\frac{2269!}{2269! \cdot 2270 \cdot 2271})(\frac{110 \cdot 109!}{109!})(\frac{97 \cdot 96!}{96!}) = \frac{110 \cdot 97}{2270 \cdot 2271} = 0.0020697 \approx 0.0021$$

1.7 Question 7 - Setting the beta hyper-parameters

Suppose $\theta \sim \beta(\alpha_1, \alpha_2)$ and we believe that $E[\theta] = m$ and $var[\theta] = v$. Using Equation 2.62, solve for α_1 and α_2 in terms of m and v. What values do you get if $m = 0.7$ and $v = 0.2^2$?
Solution:

By plugging in the notations from the assignment description in Equation 2.62 (Murphy) we get the following expressions for the mean and the variance

$$E[\theta] = m = \frac{\alpha_1}{\alpha_1 + \alpha_2} \tag{43}$$

$$var[\theta] = v = \frac{\alpha_1 \alpha_2}{(\alpha_1 + \alpha_2)^2(\alpha_1 + \alpha_2 + 1)} \tag{44}$$

Solving for α_2 from (43) yields:

$$m(\alpha_1 + \alpha_2) = \alpha_1$$
$$(m\alpha_1 + m\alpha_2) = \alpha_1$$

$$\alpha_2 = \alpha_1 \frac{(1-m)}{m} \tag{45}$$

Next we plug in the expression for α_2 into (44):

$$v = \frac{\alpha_1 \alpha_1 \frac{(1-m)}{m}}{(\alpha_1 + \alpha_1 \frac{(1-m)}{m})^2 (\alpha_1 + \alpha_1 \frac{(1-m)}{m} + 1)} =$$

$$\frac{\alpha_1 \alpha_1 \frac{(1-m)}{m}}{(\frac{\alpha_1 + m\alpha_1 - m\alpha_1}{m})^2 (\frac{\alpha_1 m + \alpha_1 - \alpha_1 m + m}{m})} =$$

$$\frac{\alpha_1^2 \frac{(1-m)}{m}}{\frac{\alpha_1^2}{m^2}(\frac{m+\alpha_1}{m})} = \frac{(1-m)}{\frac{1}{m^2}(m+\alpha_1)} = \frac{m^2(1-m)}{(m+\alpha_1)}$$

We solve for α_1:

$$v(m+\alpha_1) = m^2(1-m) \iff (vm+v\alpha_1) = m^2(1-m) \iff \alpha_1 = \frac{(1-m)m^2 - vm}{v} \iff$$

$$\alpha_1 = \frac{m(m-m^2-v)}{v} \tag{46}$$

Now we put this expression for α_1 into (45), which gives us:

$$\alpha_2 = (\frac{m(m-m^2-v)}{v})\frac{(1-m)}{m} = \frac{(1-m)(m-m^2-v)}{v} \tag{47}$$

Lastly we plug in the the given values for v and m, i.e: $m = 0.7$ and $v = 0.2^2$. Into (46) and (47):

$$\alpha_2 = \frac{(1-0.7)(0.7-0.7^2-0.2^2)}{0.2^2} = 1.275$$

$$\alpha_1 = \frac{0.7(0.7-0.7^2-0.2^2)}{0.2^2} = 2.975$$

1.8 Question 8 - Bayes factor for coin tossing

Suppose we toss a coin N = 10 times and observe $N_1 = 9$ heads. Let the null hypothesis be that the coin is fair, and the alternative be that the coin can have any bias, so $p(\theta) = Unif(0, 1)$. Derive the Bayes factor $BF_{1,0}$ in favor of the biased coin hypothesis. What if $N = 100$ and $N_1 = 90$? Hint: see Exercise 3.17.
Solution:

From Murphy (p.165) equation (5.39) we get the expression for the Bayes factor:

$$BF_{1,0} \triangleq \frac{p(D|M_1)}{p(D|M_0)} = \frac{p(M_1|D)}{p(M_0|D)} / \frac{p(M_1)}{p(M_0)} \tag{48}$$

Thus it computes the likelihood ratio. If $BF_{1,0} > 1$ then we prefer model 1, otherwise we prefer model 0. Additionally M_0 is the null hypothesis and M_1 is the alternative hypothesis.

First lets derive the marginal likelihood for the alternative hypothesis M_1.

Since we are only provided the counts for the coin tosses we are working with a binomial likelihood. Next, we know from the assignment description that the prior for the alternative distribution is $Unif(0, 1)$ which means that the prior is a $Beta(\theta|a, b)$ prior which gets the values: a=1 and b =1 (Murphy, p.34). Thus for M_1 we get the beta-binomial model where we have fixed, observed counts and $N_1 \sim Bin(N, \theta)$. The marginal likelihood for the Beta-binomial model was derived on lecture 5 as:

$$P(D) = \binom{N}{N_1} \frac{B(N_1 + a, N_0 + b)}{B(a, b)} \tag{49}$$

Where $N_0 = N - N_1$ and the Beta function is given by the following expression:

$$B(a, b) \triangleq \frac{\Gamma(a)\Gamma(b)}{\Gamma(a + b)} \tag{50}$$

Next the marginal likelihood for the null hypothesis M_0 is just the binomial distribution as it is a fair coin. Where N_1 = number of head and N= total number of coin tosses. The binomial was given on lecture 2 as:

$$Bin(k|n, \theta) = \binom{n}{k} \theta^k (1 - \theta)^{n-k} \tag{51}$$

Now finally we compute the Bayes factor for $N = 10$ and $N_1 = 9$ heads. Using (49) and (50) for M_1 and (51) for M_0

$$p(D|M_0) = \binom{10}{9} \frac{1}{2}^9 (1 - \frac{1}{2})^{10-9} = \binom{10}{9} (\frac{1}{2})^{10} = \frac{10}{1024} = 0.00976 \approx 0.0098$$

$$P(D|M_1) = \binom{10}{9} \frac{B(9 + 1, 1 + 1)}{B(1, 1)}$$

$$BF_{1,0} \triangleq \frac{p(D|M_1)}{p(D|M_0)} = \binom{10}{9} \frac{B(10, 2)}{B(1, 1)} \frac{1024}{10}$$

$$\frac{10!}{9!} \frac{\Gamma(10)\Gamma(2)}{\Gamma(12)} \frac{1024}{10} = \frac{9!}{11!} 1024 = 9.30909.. \approx 9.31$$

Thus, Jeffery's scale of evidence means that we have moderate evidence for preferring model M_1 over M_0.

The Bayes factor for $N = 100$ and $N_1 = 90$ heads. Using (49) and (50) for M_1 and (51) for M_0

$$p(D|M_0) = \binom{100}{90}\frac{1}{2}^{90}(1-\frac{1}{2})^{100-90} = \binom{100}{90}(\frac{1}{2})^{100}$$

$$P(D|M_1) = \binom{100}{90}\frac{B(90+1,10+1)}{B(1,1)}$$

$$BF_{1,0} \triangleq \frac{p(D|M_1)}{p(D|M_0)} = \binom{100}{90}\frac{B(91,11)}{B(1,1)}\frac{1}{\binom{100}{90}(\frac{1}{2})^{100}} = \frac{\Gamma(91)\Gamma(11)}{\Gamma(102)}\frac{1}{(\frac{1}{2})^{100}} =$$

$$\frac{90!10!}{101!}\frac{1}{(\frac{1}{2})^{100}} = 7.250590218..\cdot 10^{14} \approx 7.25\cdot 10^{14}$$

Thus, Jeffery's scale of evidence means that we have decisive evidence for preferring model M_1 over M_0.

1.9 Question 9 - Proof that a mixture of conjugate priors is indeed conjugate

Derive Equation 5.69.
Solution: Equation 5.69 in Murphy is given by:

$$p(\theta|D) = \sum_k p(z=k|D)p(\theta|D,z=k)$$

We rewrite the right hand side of the expression by using Baye's Theorem:

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)} \tag{52}$$

We can express a mixture of conjugate priors by introducing a latent variable z, where z=k means that θ comes from mixture component k. Which give the prior the following form (Murphy, p.171):

$$p(\theta) = \sum_k p(z=k)p(\theta|z=k) \tag{53}$$

where each $p(\theta|z=k)$ is conjugate and $p(z=k)$ is a mixing weight. Thus $p(\theta)$ becomes a weighted mix of conjugate priors.
We combine (52) and (53):

$$p(\theta|D) = \frac{1}{p(D)} p(D|\theta) \sum_k p(z=k) p(\theta|z=k) =$$

$$= \frac{1}{p(D)} \sum_k p(z=k) p(\theta|z=k) p(D|\theta) \tag{54}$$

Now we rewrite $p(D|\theta) = p(D|\theta, z=k)$. Where we use the fact that D is conditionally independent of z given θ, i.e. if we know θ, then z cannot give us any additional information about the data. Thus, we can use this in (54):

$$p(\theta|D) = \frac{1}{p(D)} \sum_k p(z=k) p(D|\theta, z=k) p(D|\theta) \tag{55}$$

Now we can rewrite the expression within the summation in (55) as a joint probability by applying the chain rule.

$$p(\theta|D) = \frac{1}{p(D)} \sum_k p(z=k, D, \theta) \tag{56}$$

We then apply the product rule twice to the previous expression:

$$p(\theta|D) = \frac{1}{p(D)} \sum_k p(D) p(z=k, \theta|D) = \frac{1}{p(D)} \sum_k p(D) p(\theta|D, z=k) p(z=k|D) =$$

$$\frac{1}{p(D)} p(D) \sum_k p(\theta|D, z=k) p(z=k|D) = \sum_k p(z=k|D) p(\theta|D, z=k)$$

Thus we have derived equation (5.69) from Murphy.

Showing that the posterior that has a mixture of conjugate priors, is in itself a combination of weighted conjugate distributions.

1.10 Question 10- MLE and model selection for a 2d discrete distribution

Let $x \in 0, 1$ denote the result of a coin toss (x =0 for tails, x =1 for heads). The coin is potentially biased, so that heads occurs with probability θ_1. Suppose that someone else observes the coin ip and reports to you the outcome, y. But this person is unreliable and only reports the result correctly with probability θ_2; i.e., $p(y|x, \theta_2)$ is given by

Assume that θ_2 is independent of x and θ_1.

a. Write down the joint probability distribution $p(x, y|\theta)$ as a 2x2 table, in terms of $\theta = (\theta_1, \theta_2)$.

Table 1: $p(y|x,\theta)$

	y=0	**y=1**
x=0	θ_2	1 - θ_2
x=1	1 -θ_2	θ_2

Solution part a:
From the product rule we can rewrite the joint probability as:

$$p(x, y|\theta) = p(y|x, \theta)p(x|\theta)$$

Which here gives us:

$$p(x, y|\theta) = p(y|x, \theta_2)p(x|\theta_1)$$

From the assignment description we then get:

$$p(x = 0|\theta_1) = (1 - \theta_1)$$

$$p(x = 1|\theta_1) = \theta_1$$

We use the information above to compute the table for $p(x, y|\theta)$

Table 2: $p(x, y|\theta)$

	y=0	**y=1**
x=0	(1 - θ_1)θ_2	(1 - θ_1)(1 − θ_2)
x=1	(1 -θ_2)θ_1	$\theta_1\theta_2$

b. Suppose have the following dataset: $\mathbf{x} = (1, 1, 0, 1, 1, 0, 0), \mathbf{y} = (1, 0, 0, 0, 1, 0, 1)$. What are the MLEs for θ_1 and θ_2? Justify your answer. Hint: note that the likelihood function factorizes,

$$p(x, y|\theta) = p(y|x, \theta_2)p(x|\theta_1) \tag{57}$$

What is $p(D|\hat{\theta}, M_2)$ where M_2 denotes this 2-parameter model? (You may leave your answer in fractional form if you wish.)
Solution to part b:

From (57) in the assignment description we are given an equation for the likelihood. However, like in Question 3 we instead compute the log likelihood of the data and maximize this expression, which is valid as the logarithm is monotonically increasing. We do this since this expression is computationally easier to maximize.

$$log(p(x, y|\theta)) = log(p(y|x, \theta_2)p(x|\theta_1)) = log(p(y|x, \theta_2)) + log(p(x|\theta_1)) \quad (58)$$

Since the entire sequence of data is observable and as the data takes on binary values we can use the Bernoulli distribution for the likelihood. Thus in general the likelihood becomes:

$$p(D|\theta) = \theta^{N_1}(1-\theta)^{N_0}$$

Thus the log likelihood for the Bernoulli becomes:

$$l(\theta) = log(p(D|\theta)) = N_1 log(\theta) + N_0 log(1-\theta)$$

Now we can find the MLE θ by differentiating the log likelihood with respect to θ, setting it equal to 0 and then solving for θ:

$$\frac{dl(\theta)}{d\theta} = \frac{N_1}{\theta} - \frac{N_0}{1-\theta} = 0$$
$$\frac{N_1}{\theta} = \frac{N_0}{1-\theta}$$
$$N_1(1-\theta) = N_0\theta$$
$$(N_1 - N_1\theta) = N_0\theta$$
$$N_1 = (N_0 + N_1)\theta$$
$$\theta_{MLE} = \frac{N_1}{(N_0 + N_1)}$$

Now looking at (58), we see that the right hand side consists of 2 terms. Maximizing the log likelihood on the right hand side is obviously equivalent to maximizing these two components individually since the two terms are summed together.

Thus, to get $\theta_{1,MLE}$ we start by looking at the term $log(p(x|\theta_1)$ and use the derived θ_{MLE} from above. Where N_1= # of tails (1) and N_0= # of heads (0). Remember that θ_1 is the probability for heads;

$$\theta_{1,MLE} = \frac{4}{(3+4)} = \frac{4}{7}$$

Now to get $\theta_{2,MLE}$ we maximize the second term of (58) i.e. $log(p(y|x, \theta_2))$. Remember that θ_2 is the probability of reporting the correct result.Thus N_1= # of times x and y are equal (1) and N_0= # of times x and y are not equal (0):

$$\theta_{2,MLE} = \frac{4}{7}$$

Now to the second part of the question: what is $p(\widetilde{D}|\hat{\theta}, M_2)$? Using the bernoulli, (57) and the computed $\theta_{MLE} : s$ we can compute the expression as:

$$p(D|\hat{\theta}, M_2) = p(y|x, \theta_2)p(x|\theta_1) = (\theta_{1,MLE}^{N_{1,1}}(1-\theta_{1,MLE})^{N_{1,0}})(\theta_{2,MLE}^{N_{2,1}}(1-\theta_{2,MLE})^{N_{2,0}})$$

Where we denote $N_{2,1}$= # of times x and y are equal (1) and $N_{2,0}$= # of times x and y are not equal $N_{1,1}$= # of tails in x (1) and $N_{1,0}$= # of heads in x Which gives us:

$$p(D|\hat{\theta}, M_2) = (\frac{4}{7}^4(1-\frac{4}{7})^3)(\frac{4}{7}^4(1-\frac{4}{7})^3) = \frac{4^8 \cdot 3^6}{7^{14}} = 0.00007044.. \approx 0.00007$$

c. Now consider a model with 4 parameters,$\theta = (\theta_{0,0}, \theta_{0,1}, \theta_{1,0}, \theta_{1,1})$, representing $p(x, y|\theta) = \theta_{x,y}$. (Only 3 of these parameters are free to vary, since they must sum to one.) What is the MLE of θ? What is $p(D|\hat{\theta}, M_4)$ where M4 denotes this 4-parameter model?

Solution:
Now the 4 different thetas correspond to:

$\theta_{0,0}$ = proportion of times when x=0 and y=0
$\theta_{0,1}$ =proportion of times when x=0 and y=1
$\theta_{1,1}$ =proportion of times when x=1 and y=1
$\theta_{1,0}$ =proportion of times when x=1 and y=0

We use the previous expression for $\theta_{MLE}$and the above definition for the counts to compute the four parameters MLE values from $\mathbf{x}$ and $\mathbf{y}$, and we get:

$$\theta_{0,0} = \frac{2}{7}, \theta_{0,1} = \frac{1}{7}, \theta_{1,1} = \frac{2}{7}, \theta_{1,0} = \frac{2}{7}$$

Now we can compute $p(D|\hat{\theta}, M_4)$ by using the above theta values. Since the 4 parameters sum to 1 and capture the 4 possible outcomes that can happen and as the data is iid its likelihood can be computed by simply multiply the θ values for each corresponding outcome:

$$p(D|\hat{\theta}, M_4) = (\frac{2}{7})^2(\frac{1}{7})(\frac{2}{7})^2(\frac{2}{7})^2 = 0.000077713... \approx 0.000078$$

d. Suppose we are not sure which model is correct. We compute the leave-one-out cross validated log likelihood of the 2-parameter model and the 4-parameter model as follows:

$$L(m) = \sum_{i=1}^{n} log(p(x_i, y_i|m, \hat{\theta}(D_{-i})$$

and $\hat{\theta}(D_{-i})$) denotes the MLE computed on D excluding row i. Which model will CV pick and why? Hint: notice how the table of counts changes when you omit each training case one at a time.
Solution:

$$\mathbf{x} = (1, 1, 0, 1, 1, 0, 0), \mathbf{y} = (1, 0, 0, 0, 1, 0, 1)$$

We start with the 2 parameter model and we leave out one row out sequentially starting from the first element in $\mathbf{x}$ both $\mathbf{y}$ to the last row in both of these vectors.
We start with $i = 0$:

$$D_{-0} : x' = (1, 0, 1, 1, 0, 0), y = (0, 0, 0, 1, 0, 1)$$

$$x_0 = 1, y_0 = 1$$

From fourth quadrant in the table in part 8.a we get

$$p(x = 1, y = 1|m_2, \theta) = \theta_1\theta_2$$

Using the equation for θ_{MLE} from part b on the data set D_{-0} : gives us:

$$\theta_1 = \frac{3}{6}, \theta_2 = \frac{3}{6}$$

Where θ_1 is the proportion of number of heads in $\mathbf{x}$ and θ_2 is the proportion of times that x and y are equal. Thus:

$$p(x = 1, y = 1|m_2, \theta) = \frac{3}{6} \cdot \frac{3}{6}$$

Thus

$$log(p(x = 1, y = 1|m_2, \theta)) = log(\frac{3}{6}) + log(\frac{3}{6})$$

Next $i = 1$:

$$D_{-1} : x' = (1, 0, 1, 1, 0, 0), y = (1, 0, 0, 1, 0, 1)$$

$$x_1 = 1, y_1 = 0$$

From the third quadrant in the table in part 8.a we get

$$p(x = 1, y = 0|m_2, \theta) = \theta_1(1 - \theta_2)$$

Using the equation for θ_{MLE} from part b on the data set D_{-1} : gives us:

$$\theta_1 = \frac{3}{6}, \theta_2 = \frac{4}{6}$$

Thus:

$$p(x=1,y=0|m_2,\theta)=\frac{3}{6}\cdot(1-\frac{4}{6})$$

Thus

$$log(p(x=1,y=0|m_2,\theta))=log(\frac{3}{6})+log(\frac{2}{6})$$

Next $i=2$:

$$D_{-2}:x'=(1,1,1,1,0,0),y=(1,0,0,1,0,1)$$

$$x_2=0,y_2=0$$

From the first quadrant in the table in part 8.a we get

$$p(x=0,y=0|m_2,\theta)=(1-\theta_1)\theta_2$$

Using the equation for θ_{MLE} from part b on the data set D_{-2} : gives us:

$$\theta_1=\frac{4}{6},\theta_2=\frac{3}{6}$$

Thus:

$$p(x=0,y=0|m_2,\theta)=(1-\frac{4}{6})\cdot\frac{3}{6}$$

Thus

$$log(p(x=0,y=0|m_2,\theta))=log(\frac{2}{6})+log(\frac{3}{6})$$

And so on for every element in $\mathbf{x}$ and $\mathbf{y}$. The summation of all these values finally gives us the leave-one-out cross validated log likelihood of the 2-parameter model as:

$$L(m_2)=(log(\frac{3}{6})+log(\frac{3}{6}))+(log(\frac{3}{6})+log(\frac{2}{6}))+(log(\frac{2}{6})+log(\frac{3}{6}))+(log(\frac{3}{6})+log(\frac{2}{6}))+$$

$$+(log(\frac{3}{6})+log(\frac{3}{6}))+(log(\frac{2}{6})+log(\frac{3}{6}))+(log(\frac{2}{6})+log(\frac{2}{6}))=-8log(2)-6log(3)=$$

$$=-12.1368511..\approx-12.14$$

If we instead look at the 4 parameter model, with the same disturbance of the data, we obtain:
We start with $i=0$:

$$D_{-0}:x'=(1,0,1,1,0,0),y=(0,0,0,1,0,1)$$

$$x_0 = 1, y_0 = 1$$

Thus we need the compute θ_0 with θ_{MLE} and D_{-0}. Which is defined as: $\theta_{0,0}$ = proportion of times when x=0 and y=0. Thus:

$$\theta_{0,0} = \frac{1}{6}$$

Next $i = 1$:

$$D_{-1} : x' = (1, 0, 1, 1, 0, 0), y = (1, 0, 0, 1, 0, 1)$$

$$x_1 = 1, y_1 = 0$$

$\theta_{1,0}$ = proportion of times when x=1 and y=0. Thus:

$$\theta_{1,0} = \frac{1}{6}$$

And so on for all 7 elements in $\mathbf{x}$ and $\mathbf{y}$. By multiplying all the likelihoods for all the 7 disturbed data sets and then taking the logarithm of this expression we finally obtain the leave one out cross validated log likelihood for the 4-parameter model as:

$$L(m_4) = log(1/6)+log(1/6)+log(1/6)+log(1/6)+log(1/6)+log(1/6)+log(0/6) \approx$$

$$\approx -10.75 + log(0/6) \to -\infty$$

Therefore the 2-parameter model will be chosen due to the 0 count for the last disturbed data point. However, it is somewhat unrealistic to give it a zero probability and say that the data could not occur. To avoid this issue we could have used some type of smoothing.

e. Recall that an alternative to CV is to use the BIC score, dened as

$$BIC(M, D) \triangleq logp(D|\hat{\theta}_{MLE}) - \frac{dof(M)}{2} log(N)$$

where dof(M) is the number of free parameters in the model, Compute the BIC scores for both models (use log base e). Which model does BIC prefer?
Solution:

Starting with the 2 parameter model:
dof(M)=2 since we have two hyper parameters θ , N=7 since the length of $\mathbf{x}$ and $\mathbf{y}$ is 7 and lastly $p(D|\hat{\theta}, M_2)$ was computed in part b as: $p(D|\hat{\theta}, M_2) = \frac{4^8 \cdot 3^6}{7^{14}}$.
Therefore we get the BIC score:

$$BIC(M_2, D) \triangleq logp(D|\hat{\theta}, M_2) - \frac{dof(M)}{2} log(N) = log(\frac{4^8 \cdot 3^6}{7^{14}}) - \frac{2}{2} log(7) = -11.5066... \approx -11.51$$

Next we look at the 4 parameter model:
dof(M)=3 since the 4 parameters must sum to 1 only 3 of the parameters are free to vary, N=7 since the length of $\mathbf{x}$ and $\mathbf{y}$ is 7 and lastly $p(D|\hat{\theta}, M_4)$ was computed in part c as: $p(D|\hat{\theta}, M_4) = (\frac{2}{7})^2(\frac{1}{7})(\frac{2}{7})^2(\frac{2}{7})$. Therefore we get the BIC score:

$$BIC(M_4, D) \triangleq logp(D|\hat{\theta}, M_4) - \frac{dof(M)}{2} log(N) = log((\frac{2}{7})^2(\frac{1}{7})(\frac{2}{7})^2(\frac{2}{7})) - \frac{3}{2} log(7)$$

$$= -12.381353... \approx -12.38$$

Thus also the BIC score prefers the first, 2-parameter model.