CSE 5522 HOMEWORK 2

BRENDAN WHITAKER

1. (a) *What is the* log-*likelihood* L *of observing* $x_1, ..., x_N$ *? Write the formula.*

Let L denote the log-likelihood. By definition of the Gaussian, we have:

$$\begin{aligned}
L = logP(x_1, ..., x_N) &= log\left(\prod_{i=1}^{B} x_i\right) = \sum_{i=1}^{N} logP(x_i) \\
&= \sum_{i=1}^{N} log\left(\frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}\right) \\
&= \sum_{i=1}^{N} log\left(\frac{1}{\sqrt{2\pi}\sigma}\right) + log\left(e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}\right) \\
&= \sum_{i=1}^{N} log\left(\frac{1}{\sqrt{2\pi}\sigma}\right) + \sum_{i=1}^{N} log\left(e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}\right) \\
&= Nlog\left(\frac{1}{\sqrt{2\pi}\sigma}\right) + \sum_{i=1}^{N} \frac{-(x_i-\mu)^2}{2\sigma^2} \\
&= Nlog1 - Nlog\sqrt{2\pi} - Nlog\sigma + \sum_{i=1}^{N} \frac{-(x_i-\mu)^2}{2\sigma^2} \\
&= -Nlog\sqrt{2\pi} - Nlog\sigma + \sum_{i=1}^{N} \frac{-(x_i-\mu)^2}{2\sigma^2} \\
&= -Nlog\sqrt{2\pi} - Nlog\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2.
\end{aligned} \tag{1}$$

(b) *What is* $\frac{\partial L}{\partial \sigma}$*?*

Observe:

$$\begin{aligned}
\frac{\partial L}{\partial \sigma} &= \frac{\partial}{\partial \sigma}\left(-Nlog\sqrt{2\pi} - Nlog\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2\right) \\
&= -N\frac{\partial}{\partial \sigma}log\sigma - \frac{1}{2}\sum_{i=1}^{N}(x_i-\mu)^2\frac{\partial}{\partial \sigma}\frac{1}{\sigma^2} \\
&= -N\frac{1}{\sigma} + \sum_{i=1}^{N}(x_i-\mu)^2\frac{1}{\sigma^3}.
\end{aligned} \tag{2}$$

Date: SP18.

(c) *What is the MLE of σ?*

We set $\frac{\partial L}{\partial \sigma} = 0$ and solve for σ:

$$\begin{aligned} 0 &= -N\frac{1}{\sigma} + \sum_{i=1}^{N}(x_i - \mu)^2\frac{1}{\sigma^3} \\ N\frac{1}{\sigma} &= \sum_{i=1}^{N}(x_i - \mu)^2\frac{1}{\sigma^3} \\ N\sigma^2 &= \sum_{i=1}^{N}(x_i - \mu)^2 \\ \hat{\sigma} &= \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}. \end{aligned} \tag{3}$$

2. *Bayesian estimation for coin tosses.*

(a) *Prove that $P(x_{m+1} = H|x_1, ..., x_m) = (\alpha + \#H)/(\alpha + \#H + \beta + \#T)$.*

Proof. Define Θ to be a random variable representing the probability of heads, our prior. Also define the beta function:

$$B(\alpha, \beta) = \frac{(\alpha - 1)!(\beta - 1)!}{(\alpha + \beta - 1)!}.$$

Recall the definition of the beta probability density function on $[0, 1]$:

$$P_{Beta(\alpha,\beta)}(\Theta) = \frac{1}{B(\alpha, \beta)}\Theta^{\alpha-1}(1 - \Theta)^{\beta-1}.$$

And since this is a probability density function, we know it integrates to 1 on this interval, which is a fact we will use later in the derivation.
We use the equality given in the problem statement:

$$P(x_{m+1} = H, \Theta|x_1, ..., x_m) = P(x_{m+1} = H|\Theta)P(\Theta|x_1, ..., x_n). \tag{4}$$

Now we marginalize to find $P(x_{m+1} = H|x_1, ..., x_m)$:

$$
\begin{aligned}
&P(x_{m+1} = H|x_1, ..., x_m) \\
&= \int_0^1 P(x_{m+1} = H|\Theta)P(\Theta|x_1, ..., x_n)d\Theta \\
&= \int_0^1 \Theta \frac{1}{B(\#H + \alpha, \#T + \beta)} \Theta^{\#H+\alpha-1}(1-\Theta)^{\#T+\beta-1}d\Theta \\
&= \int_0^1 \frac{1}{B(\#H + \alpha, \#T + \beta)} \Theta^{(\#H+\alpha+1)-1}(1-\Theta)^{\#T+\beta-1}d\Theta \\
&= \int_0^1 \frac{B(\#H + \alpha + 1, \#T + \beta)}{B(\#H + \alpha, \#T + \beta)} \frac{1}{B(\#H + \alpha + 1, \#T + \beta)} \Theta^{(\#H+\alpha+1)-1}(1-\Theta)^{\#T+\beta-1}d\Theta \\
&= \frac{B(\#H + \alpha + 1, \#T + \beta)}{B(\#H + \alpha, \#T + \beta)} \int_0^1 \frac{1}{B(\#H + \alpha + 1, \#T + \beta)} \Theta^{(\#H+\alpha+1)-1}(1-\Theta)^{\#T+\beta-1}d\Theta \\
&= \frac{B(\#H + \alpha + 1, \#T + \beta)}{B(\#H + \alpha, \#T + \beta)} \int_0^1 P_{Beta(\#H+\alpha+1, \#T+\beta)}(\Theta)d\Theta \\
&= \frac{B(\#H + \alpha + 1, \#T + \beta)}{B(\#H + \alpha, \#T + \beta)} \\
&= \frac{(\#H + \alpha + 1 - 1)!(\#T + \beta - 1)!}{(\#H + \alpha + 1 + \#T + \beta - 1)!} \frac{(\#H + \alpha + \#T + \beta - 1)!}{(\#H + \alpha - 1)!(\#T + \beta - 1)!} \\
&= \frac{(\#H + \alpha)!(\#T + \beta - 1)!}{(\#H + \alpha + \#T + \beta)!} \frac{(\#H + \alpha + \#T + \beta - 1)!}{(\#H + \alpha - 1)!(\#T + \beta - 1)!} \\
&= \frac{(\#H + \alpha)!}{(\#H + \alpha + \#T + \beta)!} \frac{(\#H + \alpha + \#T + \beta - 1)!}{(\#H + \alpha - 1)!} \\
&= \frac{(\#H + \alpha)}{(\#H + \alpha + \#T + \beta)!} (\#H + \alpha + \#T + \beta - 1)! \\
&= \frac{\#H + \alpha}{\#H + \alpha + \#T + \beta}.
\end{aligned} \tag{5}
$$

□

3. (a) *Define the variance* $var(X) = E((X - E(X))^2)$. *Prove* $E(X^2) = var(X) + E(X)^2$.

Proof. Observe:

$$
\begin{aligned}
var(X) &= E((X - E(X))^2) \\
&= E(X^2 - 2XE(X) + E(X)^2) \\
&= E(X^2) - 2E(X)E(X) + E(X)^2 \\
&= E(X^2) - E(X)^2.
\end{aligned} \tag{6}
$$

Hence:

$$E(X^2) = var(X) + E(X)^2.$$

□

(b) *Define the conditional variance* $var(X|Y) = E((X - E(X|Y))^2|Y)$. *Prove that* $var(X) = E(var(X|Y)) + var(E(X|Y))$.
Proof. Recall the law of total expectation:

$$E(X) = E(E(X|Y)).$$

Behold:

$$\begin{aligned} var(X) &= E(X^2) - E(X)^2 \\ &= E(E(X^2|Y)) - E(E(X|Y))^2 \\ &= E(var(X|Y) + E(X|Y)^2) - E(E(X|Y))^2 \\ &= E(var(X|Y)) + E(E(X|Y)^2)) - E(E(X|Y))^2 \\ &= E(var(X|Y)) + var(E(X|Y)). \end{aligned} \tag{7}$$

Note the second equality is by the law of total expectation, the third is by part (a), the fourth equality is by linearity of expectation, and the fifth is by the alternate definition of variance derived in part (a). □

4. *Recall that KL-divergence is defined as:*

$$KL(q||p) = E_p\left[\log \frac{p(Z)}{q(Z)}\right].$$

(a) *Let* $N(-1, 1)$ *and* $N(1, 1)$ *be two univariate Gaussian distributions with mean* -1 *and* -1 *and unit variance. What is* $KL(N(-1, 1), N(1, 1))$*? What is* $KL(N(1, 1), N(-1, 1))$*? Are they the same?*

Observe:

$$
\begin{aligned}
& KL(N(\mu_1,1),N(\mu_2,1)) \\
=& E_1\left[\log \frac{\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(X-\mu_1)^2}}{\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(X-\mu_2)^2}}\right] \\
=& E_1\left[\log\left(\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(X-\mu_1)^2}\right)-\log\left(\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(X-\mu_2)^2}\right)\right] \\
=& E_1\left[\log\frac{1}{\sqrt{2\pi}}+\log e^{-\frac{1}{2}(X-\mu_1)^2}-\left[\log\frac{1}{\sqrt{2\pi}}+\log e^{-\frac{1}{2}(X-\mu_2)^2}\right]\right] \\
=& E_1\left[\log\frac{1}{\sqrt{2\pi}}-\frac{1}{2}(X-\mu_1)^2-\left[\log\frac{1}{\sqrt{2\pi}}-\frac{1}{2}(X-\mu_2)^2\right]\right] \\
=& E_1\left[-\frac{1}{2}log(2\pi)-\frac{1}{2}(X-\mu_1)^2-\left[-\frac{1}{2}log(2\pi)-\frac{1}{2}(X-\mu_2)^2\right]\right] \\
=& E_1\left[-\frac{1}{2}log(2\pi)-\frac{1}{2}(X-\mu_1)^2+\frac{1}{2}log(2\pi)+\frac{1}{2}(X-\mu_2)^2\right] \\
=& E_1\left[-\frac{1}{2}(X-\mu_1)^2+\frac{1}{2}(X-\mu_2)^2\right] \\
=& E_1\left[\frac{1}{2}\left[(X-\mu_2)^2-(X-\mu_1)^2\right]\right] \\
=& \frac{1}{2}\left[E_1\left[(X-\mu_2)^2-(X-\mu_1)^2\right]\right] \\
=& \frac{1}{2}\left[E_1[(X-\mu_2)^2]-E_1[(X-\mu_1)^2]\right] \\
=& \frac{1}{2}\left[E_1[(X-\mu_2)^2]-E_1[X^2-2X\mu_1+\mu_1^2]\right] \\
=& \frac{1}{2}\left[E_1[(X-\mu_2)^2]-E_1[X^2]+2\mu_1^2-\mu_1^2\right] \\
=& \frac{1}{2}\left[E_1[(X-\mu_2)^2]-(\sigma_1^2+\mu_1^2)+\mu_1^2\right] \\
=& \frac{1}{2}\left[E_1[(X-\mu_2)^2]-\sigma_1^2\right] \\
=& \frac{1}{2}\left[E_1[(X-\mu_1+\mu_1-\mu_2)^2]-\sigma_1^2\right] \\
=& \frac{1}{2}\left[E_1[(X-\mu_1)^2+2(X-\mu_1)(\mu_1-\mu_2)+(\mu_1-\mu_2)^2]-\sigma_1^2\right] \\
=& \frac{1}{2}\left[\sigma_1^2+(\mu_1-\mu_2)^2-\sigma_1^2\right] \\
=& \frac{1}{2}\left[(\mu_1-\mu_2)^2\right].
\end{aligned}
\tag{8}
$$

So plugging in $\mu_1=-1$, $\mu_2=1$, we have:

$$KL(N(\mu_1,1),N(\mu_2,1))=\frac{1}{2}\left[(-1-1)^2\right]=2.$$

And plugging in $\mu_1 = 1$, $\mu_2 = -1$, we have:

$$KL(N(\mu_1, 1), N(\mu_2, 1)) = \frac{1}{2}\left[(1+1)^2\right] = 2.$$

So they are the same.

(b) *Let $P(x)$ be the probability distribution of a fair die ($P(x = 1) = \cdots = P(x = 6) = 1/6$) and $Q(x)$ be the distribution of a biased die, with $Q(x = 1) = \cdots = Q(x = 4) = \frac{1}{8}$ and $Q(x = 5) = Q(x = 6) = \frac{1}{4}$. What is $KL(P||Q)$? What is $KL(Q||P)$? You will need a calculator for this.* Observe:

$$\begin{aligned}
KL(P||Q) &= E_P\left[\log_2 \frac{P(X)}{Q(X)}\right] \\
&= E_P\left[\log_2 \frac{1}{6Q(X)}\right] \\
&= E_P\left[-\log_2 6Q(X)\right] \\
&= \sum_{i=1}^{6}\left[-\log_2(6Q(X=i))P(X=i)\right] \\
&= 4\left[-\log_2(6\frac{1}{8})\frac{1}{6}\right] + 2\left[-\log_2(6\frac{1}{4})\frac{1}{6}\right] \\
&\approx 0.277 + -.195 \\
&\approx 0.082.
\end{aligned} \tag{9}$$

$$\begin{aligned}
KL(Q||P) &= E_Q\left[\log_2 \frac{Q(X)}{P(X)}\right] \\
&= E_Q\left[\log_2(6Q(X))\right] \\
&= \sum_{i=1}^{6}\left[\log_2(6Q(X=i))Q(X=i)\right] \\
&= 4\left[\log_2(6\frac{1}{8})\frac{1}{8}\right] + 2\left[\log_2(6\frac{1}{4})\frac{1}{4}\right] \\
&\approx -0.207 + 0.292 \\
&\approx 0.085.
\end{aligned} \tag{10}$$

5. *Suppose there is a bag containing two biased coins A and B with probabilities of coming up heads of* 0.4, *and* 0.6, *respectively. You will draw a coin randomly from the bag, denoted Y, with an equal chance of $\frac{1}{2}$, and then flip that coin three times, to get the outcomes X_1, X_2, X_3. Note that variable Y can be either A or B, and that X_i can be either H(eads) or T(ails).*

6. (a) *E step.*
Behold:

$$
\begin{aligned}
&P(\text{unobserved variables}|\text{observed variables}, \pi_A, \Theta_A, \Theta_B)\\
=&P(Y = A|X_1 = T, X_2 = H, X_3 = T, \pi_A, \Theta_A, \Theta_B)\\
=&\frac{P(Y, X_1 = T, X_2 = H, X_3 = T|\pi_A, \Theta_A, \Theta_B)}{\sum_y P(Y = y, X_1 = T, X_2 = H, X_3 = T|\pi_A, \Theta_A, \Theta_B)}\\
=&\frac{P(X_1 = T, X_2 = H, X_3 = T|A) \cdot P(Y = A)}{P(X_1 = T, X_2 = H, X_3 = T|A)P(Y = A) + P(X_1 = T, X_2 = H, X_3 = T|B)P(Y = B)}\\
=&\frac{(1 - \Theta_A)^2\Theta_A \cdot \pi_A}{(1 - \Theta_A)^2\Theta_A \cdot \pi_A + (1 - \Theta_B)^2\Theta_B \cdot (1 - \pi_A)}.
\end{aligned}
\tag{11}
$$

(b) *M step.* We compute:

$$
\begin{aligned}
&\text{argmax}_{\pi_A,\Theta_A,\Theta_B} \sum_y P(Y = y|X_1 = T, X_2 = H, X_3 = T, \pi_A^{old}, \Theta_A^{old}\Theta_B^{old}) \log P(X_1 = T, X_2 = H, X_3 = T, Y = y|\pi_A, \Theta_A, \Theta_B)\\
=&\text{argmax}_{\pi_A,\Theta_A,\Theta_B} \frac{(1 - \Theta_A^{old})^2\Theta_A^{old} \cdot \pi_A^{old}}{(1 - \Theta_A^{old})^2\Theta_A^{old} \cdot \pi_A^{old} + (1 - \Theta_B^{old})^2\Theta_B^{old} \cdot (1 - \pi_A^{old})} \log((1 - \Theta_A)^2\Theta_A \cdot \pi_A)\\
+&\frac{(1 - \Theta_B^{old})^2\Theta_B^{old} \cdot (1 - \pi_A^{old})}{(1 - \Theta_A^{old})^2\Theta_A^{old} \cdot \pi_A^{old} + (1 - \Theta_B^{old})^2\Theta_B^{old} \cdot (1 - \pi_A^{old})} \log((1 - \Theta_B^{old})^2\Theta_B^{old} \cdot (1 - \pi_A^{old})).
\end{aligned}
\tag{12}
$$

7. *We perform two iterations of k-means, using the Lloyd's algorithm implementation. We define each of the points:*

$$
\begin{aligned}
A &= (0.43, 0.36)\\
B &= (0.79, 0.27)\\
C &= (0.13, 0.82)\\
D &= (0.34, 0.36)\\
E &= (0.55, 0.89)\\
F &= (0.31, 0.05).
\end{aligned}
\tag{13}
$$

And our centroids are initialized to:

$$
\begin{aligned}
C_1 &= (0.0, 0.05)\\
C_2 &= (1.0, 0.5)
\end{aligned}
\tag{14}
$$

(a) *Iteration 1:*

$$
\begin{aligned}
d(C_1, A) &= 0.452\\
d(C_1, B) &= 0.823\\
d(C_1, C) &= 0.345\\
d(C_1, D) &= 0.368\\
d(C_1, E) &= 0.674\\
d(C_1, F) &= 0.546.
\end{aligned}
\tag{15}
$$

$$\begin{aligned} d(C_2, A) &= 0.587 \\ d(C_2, B) &= 0.311 \\ d(C_2, C) &= 0.927 \\ d(C_2, D) &= 0.675 \\ d(C_2, E) &= 0.595 \\ d(C_2, F) &= 0.824. \end{aligned} \tag{16}$$

So Define the clusters U_1, U_2 corresponding to the centroids C_1, C_2, respectively. So we have:

$$\begin{aligned} U_1 &= \{ A, C, D, F \} \\ U_2 &= \{ B, E \}. \end{aligned} \tag{17}$$

So our new centroids are the midpoints of the points in these sets:

$$\begin{aligned} C_1' &= (0.30, 0.40) \\ C_2' &= (0.31, 0.58). \end{aligned} \tag{18}$$

(b) *Iteration 2:*

$$\begin{aligned} d(C_1', A) &= 0.136 \\ d(C_1', B) &= 0.507 \\ d(C_1', C) &= 0.453 \\ d(C_1', D) &= 0.057 \\ d(C_1', E) &= 0.550 \\ d(C_1', F) &= 0.350. \end{aligned} \tag{19}$$

$$\begin{aligned} d(C_2', A) &= 0.251 \\ d(C_2', B) &= 0.571 \\ d(C_2', C) &= 0.300 \\ d(C_2', D) &= 0.222 \\ d(C_2', E) &= 0.392 \\ d(C_2', F) &= 0.530. \end{aligned} \tag{20}$$

So Define the clusters U_1, U_2 corresponding to the centroids C_1, C_2, respectively. So we have:

$$\begin{aligned} U_1 &= \{ A, B, D, F \} \\ U_2 &= \{ C, E \}. \end{aligned} \tag{21}$$

So our new centroids are the midpoints of the points in these sets:

$$\begin{aligned} C_1' &= (0.47, 0.26) \\ C_2' &= (0.34, 0.86). \end{aligned} \tag{22}$$