ML assignment 1

Saurabh Vaidya
Email- vaidya.saur@husky.neu.edu

February 2020

1 Problem 1

$$
\begin{aligned}
P(A) &= P(A \cap \Omega) && (1)\\
&= P(A \cap (B \cup B^c) && (2)\\
&= P((A \cap B) \cup (A \cap B^c)) && (3)\\
&= P(A \cap B) + P(A \cap B^c) - P((A \cap B) \cap (A \cap B^c)) && (4)\\
&= P(A|B).P(B) + P(A|B^c).P(B^c) - P(A \cap B \cap B^c) && (5)\\
&= P(A|B).P(B) + P(A|B^c).P(B^c) - P(A \cap \varnothing) && (6)\\
&= P(A|B).P(B) + P(A|B^c).P(B^c) - P(\varnothing) && (7)\\
P(A) &= P(A|B).P(B) + P(A|B^c).P(B^c) && (8)\\
& && (9)
\end{aligned}
$$

Lets take an example to prove that the equation in the question doesn't hold
$\Omega = 1, 2, 3, 4, 5, 6$ be outcomes of a roll of a die
Let A be an event of 1 coming up.
Let B be an event that a number ≤ 3 shows up So , B^c is an event that a number higher than 3 shows up

$$
\begin{aligned}
A &= 1 && (10)\\
B &= 1, 2, 3 && (11)\\
B^c &= 4, 5, 6 && (12)\\
P(A) = 1/6 &\text{As per equation in the problem} && (13)\\
P(A) &= P(A|B) + P(A|B^c) && (14)\\
&= 1/3 + 0 && (15)\\
&\neq 1/6 && (16)\\
&\text{However, As per equation derived above} && (17)\\
P(A) &= P(A|B).P(B) + P(A|B^c).P(B^c) && (18)\\
&= 1/3 * 3/6 + 0 && (19)\\
&= 1/6 && (20)
\end{aligned}
$$

2 Problem 2

2.1 a

$$
\begin{aligned}
E[f(X)] &= \sum_{x \in X} f(x)p(x) && (21)\\
&= 10 * 0.1 + 5 * 0.2 + 10/7 * 0.7 && (22)\\
&= 3 && (23)\\
& && (24)
\end{aligned}
$$

2.2 b

$$E[1/p(X)] = \sum_{x \in X} (1/p(X)).p(x) \tag{25}$$
$$= 0.1 * 1/0.1 + 0.2 * 1/0.2 + 0.7 * 1/0.7 \tag{26}$$
$$= 3 \tag{27}$$

2.3 c

For n elements with arbitrary p(x)

$$E[1/p(X)] = \sum_{x \in X} (1/p(X)).p(x) \tag{28}$$
$$= a_1 * 1/a_1 + a_2 * 1/a_2 ..... a_n * 1/a_n \tag{29}$$
$$= \sum_n 1 \tag{30}$$
$$= n \tag{31}$$

3 Problem 3

$$V[X+Y] = E[(X+Y) - E[X+Y])^2] \tag{32}$$
$$= E[(X+Y)^2 - 2(X+Y)E[X+Y] + E[X+Y]^2] \tag{33}$$
$$= E[X^2 + 2XY + Y^2 - 2XE[X+Y] - 2YE[X+Y] + (E[X] + E[Y])^2] \tag{34}$$
$$= E[X^2 - 2XE[X] + E[X]^2 + Y^2 - 2YE[Y] + E[Y]^2 - 2XE[Y] + 2XY - 2YE[X] + +2E[X]E[Y]] \tag{35}$$
$$= V[X] + V[Y] + 2E[XY - XE[Y] - YE[X] + E[X]E[Y]] \tag{36}$$
$$= V[X] + V[Y] + 2E[X(Y - E[Y]) - E[X](Y - E[Y])] \tag{37}$$
$$= V[X] + V[Y] + 2E[(Y - E[Y])(X - E[X])] \tag{38}$$
$$= V[X] + V[Y] + 2COV[X, Y] \tag{39}$$

4 Problem 4

4.1 a. Maximum Likelihood

$$\lambda_{ML} = \underset{\lambda \in (0,\infty)}{\operatorname{argmax}}\ p(D|\lambda) \tag{40}$$
$$= \prod_{i=1}^{n} p(x_i|\lambda) \tag{41}$$
$$= \frac{\lambda^{\sum_{i=1}^{n} x_i}.e^{-n\lambda}}{\prod_{i=1}^{n} x_i!} \tag{42}$$
$$ll(D, \lambda) = ln \sum_{i=1}^{n} x_i - n\lambda - \sum_{i=1}^{n} ln(x_i!) \tag{43}$$
$$\frac{\partial ll(D, \lambda)}{\partial \lambda} = 1/\lambda \sum_{i=1}^{n} (x_i - n) \tag{44}$$
$$\lambda_{ML} = 1/n \sum_{i=1}^{n} x_i \tag{45}$$
$$= 99/11 \tag{46}$$
$$= 9 \tag{47}$$
$$\tag{48}$$

4.2 b. Maximum a posteriori

$$\lambda_{ML} = \underset{\lambda \in (0,\infty)}{\operatorname{argmax}}\ p(D|\lambda)p(\lambda) \tag{49}$$

$$ll(\lambda_{MAP}) = ln(\frac{\lambda^{\sum_{i=1}^{n} x_i}.e^{-n\lambda}}{\prod_{i=1}^{n} x_i!}) + ln(\theta.e^{-\theta\lambda}) \tag{50}$$

$$= ln(\lambda^{\sum_{i=1}^{n} x_i}) - ln(\prod_{i=1}^{n} x_i!) + ln(\theta) + ln(e^{-\theta\lambda}) \tag{51}$$

$$\frac{\partial ll(\lambda_{MAP})}{\partial \lambda} = 1/\lambda.\sum_{i=1}^{n} x_i - n - 1/2 \tag{52}$$

$$0 = 1/\lambda.\sum_{i=1}^{n} x_i - n - 1/2 \tag{53}$$

$$1/\lambda.\sum_{i=1}^{n} x_i = n + 1/2 \tag{54}$$

$$99/\lambda = 23/2 \tag{55}$$

$$\lambda_{MAP} = 8.608 \tag{56}$$

4.3 c. Bayes Estimate

$$p(\lambda|D) = \frac{p(D|\lambda)p(\lambda)}{p(D)} \tag{57}$$

$$= \frac{p(D|\lambda)p(\lambda)}{\int_0^\infty p(D|\lambda)p(\lambda)} \tag{58}$$

$$p(D|\lambda) = \frac{\lambda^{\sum_{i=1}^{n} x_i}.e^{-n\lambda}}{\prod_{i=1}^{n} x_i!} \tag{59}$$

$$p(\lambda) = 1/2.e^{-\lambda/2} \tag{60}$$

$$p(\lambda|D) = \frac{\frac{\lambda^{\sum_{i=1}^{n} x_i}.e^{-n\lambda}}{\prod_{i=1}^{n} x_i!}.\frac{e^{-\lambda/2}}{2}.}{\int_0^\infty \frac{\lambda^{\sum_{i=1}^{n} x_i}.e^{-n\lambda}}{\prod_{i=1}^{n} x_i!}.\frac{e^{-\lambda/2}}{2}} \tag{61}$$

$$= \frac{\lambda^{99}.e^{-23/2\lambda}}{\int_0^\infty \lambda^{99}.e^{-23/2\lambda}} \tag{62}$$

$$\int_0^\infty x^{\alpha-1}.e^{-\beta.x}dx = \frac{\Gamma(\alpha)}{\beta^\alpha} \tag{63}$$

$$\therefore p(\lambda|D) = \frac{23^{100}.\lambda^{99}.e^{-23\lambda/2}}{99!.2^{99}} \tag{64}$$

$$E[\lambda|D] = \int_0^\infty \lambda.p(\lambda|D)d\lambda \tag{65}$$

$$= \frac{23^{100}}{99!.2^{99}} \int_0^\infty \lambda^{100}.e^{-23\lambda/2}d\lambda \tag{66}$$

$$= \frac{23^{100}.2.100!}{99!.2^{100}.(-23/2)^{101}} \tag{67}$$

$$= \frac{(23/2)^{100}.2.100 * 99!}{99!.(-23/2)^{100}.(23/2)} \tag{68}$$

$$= \frac{100 * 2}{23} \tag{69}$$

$$= 8.6956 \tag{70}$$

5 Problem 5

1. The graph plot of the function shows that its a decreasing function with respect to x.

2. Given the constraint where the function is defined only when $x \geq \theta_0$, the point on x-axis from where the function is defined falls at $x = \theta_0$.

3. And since the function is only decreasing from that point on for all values of $x \geq \theta_0$, the maximum likelihood of θ_0 where the function $p(x)$ becomes maximum is at the first value of x on $x - axis$ i.e. $\min_{i=1...n}(x_i)$.

4. Thus $\theta_{ML} = \min_{i=1...n}(x_i)$.

6 Problem 6

6.1 a

The log likelihood of the distribution is

$$ll(\alpha, \beta) = \sum_{i=1}^{n} log\left[\frac{e^{-\frac{x_i-\alpha}{\beta}}.e^{-e^{-\frac{x_i-\alpha}{\beta}}}}{\beta}\right] \tag{71}$$

$$= -\sum_{i=1}^{n}\left[-\frac{(x_i-\alpha)}{\beta} - e^{-\frac{x_i-\alpha}{\beta}}\right] - nln(\beta) \tag{72}$$

$$\frac{\partial ll}{\partial \beta} = \frac{\sum_{i=1}^{n}(x_i-\alpha)}{\beta^2} - \frac{\sum_{i=1}^{n} e^{-\frac{x_i-\alpha}{\beta}}(x_i-\alpha)}{\beta^2} - \frac{n}{\beta} \tag{73}$$

$$= \frac{-\sum_{i=1}^{n} e^{-\frac{x_i-\alpha}{\beta}}(x_i-\alpha) + \sum_{i=1}^{n} x_i - (\alpha+\beta)n}{\beta^2} \tag{74}$$

$$\tag{75}$$

Similarly, for α

$$\frac{\partial ll}{\partial \alpha} = \frac{n}{\beta} - \frac{e^{-\frac{x_i-\alpha}{\beta}}}{\beta} \tag{76}$$

Now to find the best parameters we will need to perform iterative estimation using Newton-Raphson process. So we need to calcula

$$\frac{\partial^2 ll}{\partial \alpha^2} = \frac{-\sum_{i=1}^{n} e^{-\frac{x_i-\alpha}{\beta}}}{\beta^2} \tag{77}$$

$$\frac{\partial^2 ll}{\partial \beta^2} = \frac{1}{\beta^2}\left[\frac{-\sum_{i=1}^{n} e^{-\frac{x_i-\alpha}{\beta}}(x_i-\alpha)^2}{\beta^2} - n\right] - \frac{2 * \left[(-\sum_{i=1}^{n} e^{-\frac{x_i-\alpha}{\beta}}(x_i-\alpha)) + \sum_{i=1}^{n} x_i - (\alpha+\beta)n\right]}{\beta^3} \tag{78}$$

$$= \frac{-\sum_{i=1}^{n} e^{-\frac{x_i-\alpha}{\beta}}(x_i-\alpha)^2 - n*\beta^2 - 2\beta\left(-\sum_{i=1}^{n} e^{-\frac{x_i-\alpha}{\beta}}(x_i-\alpha)\right) + \sum_{i=1}^{n} x_i - (\alpha+\beta)n\right]}{\beta^4}$$

$$= \frac{2n\alpha\beta + n\beta^2 - \sum_{i=1}^{n}\left[e^{-\frac{x_i-\alpha}{\beta}}(x_i-\alpha)(x_i-\alpha-2\beta) + 2x_i\beta\right]}{\beta^4} \tag{79}$$

$$\frac{\partial ll}{\partial \alpha \partial \beta} = \frac{-n}{\beta^2} - \left[\frac{\sum_{i=1}^{n} e^{-\frac{x_i-\alpha}{\beta}}(x_i-\alpha)}{\beta^3} - \frac{\sum_{i=1}^{n} e^{-\frac{x_i-\alpha}{\beta}}}{\beta^2}\right] \tag{80}$$

$$= \frac{-n\beta - \sum_{i=1}^{n} e^{-\frac{x_i-\alpha}{\beta}}(x_i-\alpha-\beta)}{\beta^3} \tag{81}$$

$$Hessian = \begin{bmatrix} \frac{-\sum_{i=1}^{n} e^{-\frac{x_i-\alpha}{\beta}}}{\beta^2} & \frac{-n\beta-\sum_{i=1}^{n} e^{-\frac{x_i-\alpha}{\beta}}(x_i-\alpha-\beta)}{\beta^3} \\ \frac{-n\beta-\sum_{i=1}^{n} e^{-\frac{x_i-\alpha}{\beta}}(x_i-\alpha-\beta)}{\beta^3} & \frac{2n\alpha\beta+n\beta^2-\sum_{i=1}^{n}\left[e^{-\frac{x_i-\alpha}{\beta}}(x_i-\alpha)(x_i-\alpha-2\beta)+2x_i\beta\right]}{\beta^4} \end{bmatrix}$$

$$\nabla = \begin{bmatrix} \frac{n}{\beta} - \frac{e^{-\frac{x_i-\alpha}{\beta}}}{\beta} \\ \frac{-\sum_{i=1}^{n} e^{-\frac{x_i-\alpha}{\beta}}(x_i-\alpha)+\sum_{i=1}^{n} x_i-(\alpha+\beta)n}{\beta^2} \end{bmatrix}$$

The update rules are

$$\alpha^{t+1} = \alpha^t - (H_{f(\alpha^t)})^{-1}.\nabla f(\alpha^t) \tag{82}$$

$$\beta^{t+1} = \beta^t - (H_{f(\beta^t)})^{-1}.\nabla f(\beta^t) \tag{83}$$

6.2 b

1. The data is generated using values of α, β as 2,3 respectively.
2. For $n = 100$ Mean $\alpha = 2.028878410078369$ and Mean $\beta = 2.9902547783873716$
   Standard deviation of $\alpha = 0.23552323398893274$ and $\beta = 0.25446178214290177$
3. For $n = 1000$ Mean $\alpha = 2.030697780288935$ and Mean $\beta = 2.9834408221753512$
   Standard deviation of $\alpha = 0.13653102774566644$ and $\beta = 0.07039535753676388$
4. For $n = 10000$ Mean $\alpha = 2.0081357055582023$ and Mean $\beta = 3.0092586993367547$
   Standard deviation of $\alpha = 0.022635367362198414$ and $\beta = 0.022373597735652163$

6.3 c

- The initial weights that lead to the best results were when the both parameters were initialized to the half of their actual values i.e. half of mean and variance for $\alpha\beta$ respectively. The iterative process continued until the difference between values of $\alpha\beta$ from $t+1$ iteration and t iteration was lesser than 0.0001 or max of 1000 iterations whichever was earlier. Choosing values other than half of mean and variance would lead to slow convergence with wrong values as well. Ideally algorithm can converge faster with accurate results when initial point is closer and in the curve of the minima/maxima. Since α, β as anologous to mean and spread of data, a safe bet is to start from some value closer to them but not very close so as to miss the minima/maxima point. Hence I took the above decisions of initial parameter value assginment.

7 Problem 7

7.1 a

For finding the parameters update formula for Gaussian, based on the equation from lectures notes

$$ll(\mu, \sigma) = \sum_n log(p(x_i|\mu^t, \sigma^t) * P_{yi}(j|x_i, \mu^t, \sigma^t) \tag{84}$$

$$\text{where j = 1 for Gaussian} \tag{85}$$

$$= \sum_n \left[log(e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}) - log(\sqrt{2\pi\sigma^2}) \right] P_{yi}(j|x_i, \mu^t, \sigma^t) \tag{86}$$

$$= \sum_n \left[-\frac{(x_i-\mu)^2}{2\sigma^2} - ln(\sigma) - ln(\sqrt{2\pi}) \right] * P_{yi}(j|x_i, \mu^t, \sigma^t) \tag{87}$$

$$\frac{\partial ll(\mu, \sigma)}{\partial \mu} = \sum_n P_{yi}(j|x_i, \mu^t, \sigma^t) \left[\frac{(x_i - \mu)}{\sigma^2} \right] \tag{88}$$

$$\frac{\partial ll(\mu, \sigma)}{\partial \mu} = 0 \tag{89}$$

$$\mu^{t+1} = \frac{\sum_{i=1}^n P_{yi}(j|x_i, \mu^t, \sigma^t) * x_i}{\sum_{i=1}^n P_{yi}(j|x_i, \mu^t, \sigma^t)} \tag{90}$$

$$\frac{\partial ll(\mu, \sigma)}{\partial \sigma} = \sum_n P_{yi}(j|x_i, \mu^t, \sigma^t) * \left[\frac{2 * (x_i - \mu)^2}{2\sigma^3} - \frac{1}{\sigma} \right] \tag{91}$$

$$\frac{\partial ll(\mu, \sigma)}{\partial \sigma} = 0 \tag{92}$$

$$\sum_n P_{yi}(j|x_i, \mu^t, \sigma^t) * \left[\frac{(x_i - \mu)^2 - \sigma^2}{\sigma^3} \right] = 0 \tag{93}$$

$$\sigma^{t+1} = \sqrt{\frac{\sum_{i=1}^n P_{yi}(j|x_i, \mu^t, \sigma^t) * (x_i - \mu^t)^2}{\sum_{i=1}^n P_{yi}(j|x_i, \mu^t, \sigma^t)}} \tag{94}$$

Taking the derivations for weight updates and probability of a point to each distribution from the lecture notes and Gumbel parameters update formulas from previous problem, We can state the EM algorithm for these mixtures as follows

1. Initialize $\mu^0, \sigma^0\alpha^0, \beta^0, w_1^0, w_2^0$
2. Set $t = 0$
3. Repeat until convergence
   (a) $P_{yi}(k|x_i, w_k^t, \mu^t\sigma^t, \alpha^t, \beta^t) = \frac{w_k^t p(x_i|\theta^t)}{\sum_{j=1}^m w_j^t p(x_i|\theta_j^t)}$ for $\forall(i,k)$ and $\theta^t = \mu^t, \sigma^t, \alpha^t, \beta^t$
   (b) $w_k^{t+1} = \frac{\sum_{i=1}^n P_{yi}(k|x_i, w_k^t, \mu^t\sigma^t, \alpha^t, \beta^t)}{n}$
   (c) $\mu^{t+1} = \frac{\sum_{i=1}^n P_{yi}(j|x_i, \mu^t, \sigma^t) * x_i}{\sum_{i=1}^n P_{yi}(j|x_i, \mu^t, \sigma^t)}$
   (d) $\sigma^{t+1} = \sqrt{\frac{\sum_{i=1}^n P_{yi}(j|x_i, \mu^t, \sigma^t) * (x_i - \mu^t)^2}{\sum_{i=1}^n P_{yi}(j|x_i, \mu^t, \sigma^t)}}$
   (e) $\alpha^{t+1} = \alpha^t - (H_{f(\alpha^t)})^{-1}.\nabla f(\alpha^t)$
   (f) $\beta^{t+1} = \beta^t - (H_{f(\beta^t)})^{-1}.\nabla f(\beta^t)$

where $Hessian =$
$$\begin{bmatrix} \frac{-\sum_n P_{yi} e^{-\frac{x_i-\alpha}{\beta}}}{\beta^2} & \frac{-\sum_n P_{yi}\beta - \sum_n P_{yi} e^{-\frac{x_i-\alpha}{\beta}}(x_i-\alpha-\beta)}{\beta^3} \\ \frac{-\sum_n P_{yi}\beta - \sum_n P_{yi} e^{-\frac{x_i-\alpha}{\beta}}(x_i-\alpha-\beta)}{\beta^3} & \frac{\sum_n P_{yi}\alpha\beta + \sum_n P_{yi}\beta^2 - \sum_n P_{yi}\left[(x_i-\alpha)(x_i-\alpha-2\beta)e^{-\frac{x_i-\alpha}{\beta}} + 2*x_i*\beta\right]}{\beta^4} \end{bmatrix}$$

and $\nabla f =$
$$\begin{bmatrix} \frac{1}{\beta}\left[\sum_n P_{yi} - \sum_n P_{yi} e^{-\frac{x_i-\alpha}{\beta}}\right] \\ \frac{-\sum_n P_{yi}(x_i-\alpha)e^{-\frac{x_i-\alpha}{\beta}} + \sum_n P_{yi} * x_i - (\alpha+\beta)\sum_n P_{yi}}{\beta^2} \end{bmatrix}$$
$P_{yi} = P_{yi}(k|x_i, w_k^t, \mu^t\sigma^t, \alpha^t, \beta^t)$

7.2 b

1. **The initial values of weights for Gaussian and Gumbel were 0.4 and 0.6 respectively. The initial $\mu\sigma$ were 6 and 1 respectively and initial $\alpha\beta$ were 2 and 1 respectively.**
2. **For n = 100 Mean** $w_{gauss} = 0.4586579803348747$ **and Mean** $w_{gumbel} = 0.5413420196651253$
   Standard deviation of $w_{gauss} = 0.09265872498378355$ **and** $w_{gumbel} = 0.09265872498378352$
   Mean $\mu = 5.277476836918109$ **and Mean** $\sigma = 0.9747107376271795$
   Standard deviation of $\mu = 1.4461008105236732$ **and** $\sigma = 0.19908331145851405$
   Mean $\alpha = 2.78431335437176$ **and Mean** $\beta = 0.8881174622028238$
   Standard deviation of $\alpha = 1.716585883932862$ **and** $\beta = 0.17791279032268786$
3. **For n = 1000 Mean** $w_{gauss} = 0.406285532547239$ **and Mean** $w_{gumbel} = 0.5937144674527609$
   Standard deviation of $w_{gauss} = 0.028496715065933494$ **and** $w_{gumbel} = 0.02849671506593322$
   Mean $\mu = 5.991479984574222$ **and Mean** $\sigma = 1.0026647859786773$
   Standard deviation of $\mu = 0.13868153163639632$ **and** $\sigma = 0.0743175467012543$
   Mean $\alpha = 1.9971755785655922$ **and Mean** $\beta = 0.9757037507710543$
   Standard deviation of $\alpha = 0.08143106991344701$ **and** $\beta = 0.0741274686385204$
4. **For n = 10000 Mean** $w_{gauss} = 0.402470004780078$ **and Mean** $w_{gumbel} = 0.5975299952199218$
   Standard deviation of $w_{gauss} = 0.00901042736438743$ **and** $w_{gumbel} = 0.009010427364388349$
   Mean $\mu = 5.994389664906983$ **and Mean** $\sigma = 1.0057948741616354$
   Standard deviation of $\mu = 0.027275172779177077$ **and** $\sigma = 0.022985159262992296$
   Mean $\alpha = 1.9959459813407872$ **and Mean** $\beta = 0.9991603588258154$
   Standard deviation of $\alpha = 0.03135767032748187$ **and** $\beta = 0.022141774781465364$

8 Problem 8

1. **Before carrying out the experiment, my expectation was the distance from points from each other would be a normal distribution and the dimensionality of data would not effect how far apart each point is from other.**

2. **However upon carrying out the observation, it is observed that as k increases every point seems to be equidistant from each other and thus no clear distinction or clustering can be made of them. When k is small the y axis has higher value but the r(k) diminished fast as k increases.**

3. **Increasing the data 10 fold also does not change this behaviour. The higher dimensions of data we have, more likely we would observe all those data points as the same thing.**