STA414H1 - Assignment 1

1.4: For $p = p(\text{LLM}|\text{Submission})$, the expected loss for each action is

$$\begin{aligned} E(\text{Accept}) &= 10p \\ E(\text{Suspect}) &= 2p + 2(1-p) = 2 \\ E(\text{TA review}) &= 5p + 2(1-p) = 3p + 2 \\ E(\text{Minor penalty}) &= 10p + 20(1-p) = 20 - 10p \\ E(\text{Report to professor}) &= 100(1-p) = 100 - 100p. \end{aligned}$$

The probability thresholds at which we switch from accept to any other action are

$$\begin{aligned} &E(\text{Accept}) = E(\text{Suspect}) \implies 10p = 2 \implies p = 0.2 \\ &E(\text{Accept}) = E(\text{TA review}) \implies 10p = 3p + 2 \implies p = \frac{2}{7} \approx 0.2857 \\ &E(\text{Accept}) = E(\text{Minor penalty}) \implies 10p = 20 - 10p \implies p = 1 \\ &E(\text{Accept}) = E(\text{Report to professor}) \implies 10p = 100 - 100p \implies p = \frac{10}{11} \approx 0.9091. \end{aligned}$$

In order for each threshold to be < 0.01, we increase the loss of incorrectly accepting an LLM-generated submission from 10 to some n:

$$\begin{aligned} &E(\text{Accept}) = E(\text{Suspect}) \implies np = 2 \implies p = \frac{2}{n} < 0.01 \implies n > 200 \\ &E(\text{Accept}) = E(\text{TA review}) \implies np = 3p + 2 \implies p = \frac{2}{n-3} < 0.01 \implies n > 203 \\ &E(\text{Accept}) = E(\text{Minor penalty}) \implies np = 20 - 10p \implies p = \frac{20}{n+10} < 0.01 \implies n > 1990 \\ &E(\text{Accept}) = E(\text{Report to professor}) \implies np = 100 - 100p \implies p = \frac{100}{n+100} < 0.01 \implies n > 9900. \end{aligned}$$

Thus, a loss value > 9900 will make all probability thresholds < 0.01.

2.1: Let n be the number of images and note that $p(c^{(i)}|\pi) = \prod_{c=1}^{28} \pi_c^{1\{c^{(i)}=c\}}$. The log-likelihood function is

$$\begin{aligned} \ell(\theta, \pi | x, c) &= \log p(x, c|\theta, \pi) = \log \prod_{i=1}^{n} p(x^{(i)}, c^{(i)}|\theta, \pi) = \log \prod_{i=1}^{n} p(c^{(i)}|\pi) \prod_{j=1}^{D} p(x_j^{(i)}|c^{(i)}, \theta) \\ &= \sum_{i=1}^{n} \log p(c^{(i)}|\pi) + \sum_{i=1}^{n}\sum_{j=1}^{D} \log p(x_j^{(i)}|c^{(i)}, \theta) \\ &= \sum_{i=1}^{n}\sum_{c=1}^{28} 1\{c^{(i)} = c\} \log \pi_c + \sum_{i=1}^{n}\sum_{j=1}^{D} [x_j^{(i)} \log \theta_{jc^{(i)}} + (1 - x_j^{(i)}) \log(1 - \theta_{jc^{(i)}})]. \end{aligned}$$

We can maximize the terms separately and use the fact that $\sum_{c=1}^{28} \pi_c = 1$:

$$\begin{aligned} \frac{\partial}{\partial \pi_d} \sum_{i=1}^{n}\sum_{c=1}^{28} 1\{c^{(i)} = c\} \log \pi_c &= \sum_{i=1}^{n} \frac{\partial}{\partial \pi_d} [\sum_{c=1}^{28} 1\{c^{(i)} = c\} \log \pi_c + 1 - \sum_{c=1}^{28} \pi_c] \\ &= \sum_{i=1}^{n} [\frac{1\{c^{(i)} = d\}}{\pi_d} - 1] \overset{\text{set}}{=} 0 \\ \implies \hat{\pi}_d &= \frac{\sum_{i=1}^{n} 1\{c^{(i)} = d\}}{n} \end{aligned}$$

$$\frac{\partial}{\partial\theta_{kc}}\sum_{i=1}^{n}\sum_{j=1}^{D}[x_j^{(i)}\log\theta_{jc^{(i)}}+(1-x_j^{(i)})\log(1-\theta_{jc^{(i)}})]$$

$$=\sum_{i=1}^{n}(\frac{x_k^{(i)}}{\theta_{kc}}-\frac{1-x_k^{(i)}}{1-\theta_{kc}})1\{c^{(i)}=c\}=\sum_{i=1}^{n}\frac{x_k^{(i)}-\theta_{kc}}{\theta_{kc}-\theta_{kc}^2}1\{c^{(i)}=c\}\overset{set}{=}0$$

$$\implies\sum_{i=1}^{n}x_k^{(i)}1\{c^{(i)}=c\}=\hat{\theta}_{kc}\sum_{i=1}^{n}1\{c^{(i)}=c\}$$

$$\implies\hat{\theta}_{kc}=\frac{\sum_{i=1}^{n}x_k^{(i)}1\{c^{(i)}=c\}}{\sum_{i=1}^{n}1\{c^{(i)}=c\}}.$$

2.2: For each entry θ_{kc} in θ, the prior density is $f(\theta_{kc})=\frac{\Gamma(2\alpha)}{\Gamma(\alpha)\Gamma(\alpha)}\theta_{kc}^{\alpha-1}(1-\theta_{kc})^{\alpha-1}$. Since $p(\theta_{kc}|x,c,\pi)\propto p(x,c|\theta,\pi)f(\theta_{kc})$, we can maximize $p(x,c|\theta,\pi)f(\theta_{kc})$ to obtain the estimate (using ℓ from previously):

$$\log[p(x,c|\theta,\pi)f(\theta_{kc})]=\ell(\theta,\pi|x,c)+(\alpha-1)\log(\theta_{kc}-\theta_{kc}^2)+\log\Gamma(2\alpha)-2\log\Gamma(\alpha)$$

$$\frac{\partial}{\partial\theta_{kc}}\log[p(x,c|\theta,\pi)f(\theta_{kc})]=\frac{\partial}{\partial\theta_{kc}}\ell(\theta,\pi|x,c)+\frac{(\alpha-1)(1-2\theta_{kc})}{\theta_{kc}-\theta_{kc}^2}$$

$$=\sum_{i=1}^{n}(\frac{x_k^{(i)}}{\theta_{kc}}-\frac{1-x_k^{(i)}}{1-\theta_{kc}})1\{c^{(i)}=c\}+\frac{(\alpha-1)(1-2\theta_{kc})}{\theta_{kc}-\theta_{kc}^2}\overset{set}{=}0$$

$$\implies\sum_{i=1}^{n}\frac{x_k^{(i)}-\theta_{kc}}{\theta_{kc}-\theta_{kc}^2}1\{c^{(i)}=c\}=\frac{(\alpha-1)(2\theta_{kc}-1)}{\theta_{kc}-\theta_{kc}^2}$$

$$\implies\sum_{i=1}^{n}(x_k^{(i)}-\theta_{kc})1\{c^{(i)}=c\}=(\alpha-1)(2\theta_{kc}-1)=2\alpha\theta_{kc}-2\theta_{kc}-\alpha+1$$

$$\implies\sum_{i=1}^{n}x_k^{(i)}1\{c^{(i)}=c\}+\alpha-1=\theta_{kc}(2\alpha-2+\sum_{i=1}^{n}1\{c^{(i)}=c\})$$

$$\implies\hat{\theta}_{kc}=\frac{\sum_{i=1}^{n}x_k^{(i)}1\{c^{(i)}=c\}+\alpha-1}{\sum_{i=1}^{n}1\{c^{(i)}=c\}+2\alpha-2}.$$

2.3: Using Bayes' theorem,

$$\log p(c|x^{(i)},\theta,\pi)=\log\frac{p(x^{(i)},c|\theta,\pi)}{p(x^{(i)}|\theta,\pi)}=\log\frac{p(c|\pi)p(x^{(i)}|c,\theta)}{\sum_{c=1}^{28}p(x^{(i)},c|\theta,\pi)}=\log\frac{p(c|\pi)p(x^{(i)}|c,\theta)}{\sum_{c=1}^{28}p(c|\pi)p(x^{(i)}|c,\theta)}$$

$$=\log p(c|\pi)+\log p(x^{(i)}|c,\theta)-\log\sum_{c=1}^{28}\exp[\log p(c|\pi)+\log p(x^{(i)}|c,\theta)]$$

where $p(c|\pi)=\pi_c$ and $\log p(x^{(i)}|c,\theta)=\sum_{j=1}^{D}[x_j^{(i)}\log\theta_{jc}+(1-x_j^{(i)})\log(1-\theta_{jc})]$. The average log-likelihood on the training test is -0.253, while the training and test errors are 0.989 and 0.990 respectively. Note that it was important to use the MAP estimators since using $\alpha=1$ (which corresponds to using the MLE estimators) resulted in entries of θ estimated as 0, which led to errors when calculating $\log(\theta)$.

2.4: Since Naive Bayes assumes conditional independence, x_i and x_j are independent when conditioned on c, i.e. $p(x_i,x_j|c)=p(x_i|c)p(x_j|c)$. However, they are not independent after marginalizing over c since

$$p(x_i,x_j)=\sum_c p(x_i,x_j,c)=\sum_c p(x_i,x_j|c)p(c)=\sum_c p(x_i|c)p(x_j|c)p(c)$$

$$\neq\sum_c p(x_i|c)p(c)\sum_c p(x_j|c)p(c)=\sum_c p(x_i,c)\sum_c p(x_j,c)=p(x_i)p(x_j).$$

This also implies that they are not independent when unconditioned on c.

2.5: The Bernoulli Naive Bayes model needs $784 * 28 + 27 = 21979$ parameters: there is a θ_{kc} parameter for each pixel x_k and each class c, and there is a π_c for each class c except one since $\sum_c \pi_c = 1$. The model with the assumption removed needs $28 \sum_{i=1}^{784} 2^{i-1} + 27$ parameters: $p(x_i|x_1, \ldots, x_{i-1}, c)$ yields a different parameter for each permutation of $\{x_1, \ldots, x_{i-1}, c\}$, and as before there is a π_c for each class c except one.

2.7: Marginalizing over c and using the Naive Bayes assumption, we have

$$\begin{aligned} p(x_j|x_E, \theta, \pi) &= \frac{p(x_j, x_E|\theta, \pi)}{p(x_E|\theta, \pi)} = \frac{\sum_c p(x_j, x_E, c|\theta, \pi)}{\sum_c p(x_E, c|\theta, \pi)} = \frac{\sum_c p(c|\pi)p(x_j, x_E|c, \theta)}{\sum_c p(c|\pi)p(x_E|c, \theta)} \\ &= \frac{\sum_c p(c|\pi)p(x_j|c, \theta)p(x_E|c, \theta)}{\sum_c p(c|\pi)p(x_E|c, \theta)} = \frac{\sum_c \pi_c \theta_{jc}^{x_j}(1 - \theta_{jc})^{1-x_j} p(x_E|c, \theta)}{\sum_c \pi_c p(x_E|c, \theta)}. \end{aligned}$$

2.8: I find that the images in part (a) is better than the ones in part (b). Both sets of images are mostly similar, but the images in part (b) has noticeable grid patterns which make them look "artificial" and distinct from the ground truth images, whereas the images in part (a) have comparatively "natural"-looking random patterns and are thus closer to the ground truth images.

2.9: Due to the assumption of conditional independence, Σ_c is diagonal, symmetric, and positive semi-definite and takes the form

$$\begin{bmatrix} \sigma_1^2 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \sigma_D^2 \end{bmatrix}.$$

The log-likelihood function is

$$\begin{aligned} \ell(\mu, \Sigma, \pi|x, c) &= \log p(x, c|\mu, \Sigma, \pi) = \log \prod_{i=1}^{n} p(x^{(i)}, c^{(i)}|\mu, \Sigma, \pi) = \log \prod_{i=1}^{n} p(c^{(i)}|\pi)p(x^{(i)}|c^{(i)}, \mu, \Sigma) \\ &= \sum_{i=1}^{n} \log p(c^{(i)}|\pi) + \sum_{i=1}^{n} \log p(x^{(i)}|c^{(i)}, \mu, \Sigma) \\ &= \sum_{i=1}^{n} \sum_{c=1}^{28} 1\{c^{(i)} = c\} \log \pi_c \\ &\quad + \sum_{i=1}^{n} [-\frac{D}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma_{c^{(i)}}| - \frac{1}{2}(x^{(i)} - \mu_{c^{(i)}})^T \Sigma_{c^{(i)}}^{-1} (x^{(i)} - \mu_{c^{(i)}})] \\ &= \sum_{i=1}^{n} \sum_{c=1}^{28} 1\{c^{(i)} = c\} \log \pi_c - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{D} [\log(2\pi) + \log \Sigma_{c^{(i)},jj} + \frac{(x_j^{(i)} - \mu_{c^{(i)},j})^2}{\Sigma_{c^{(i)},jj}}] \end{aligned}$$

since Σ_c is diagonal. As with Bernoulli Naive Bayes, we maximize the first term to yield $\hat{\pi}_c = \frac{1}{n} \sum_{i=1}^{n} 1\{c^{(i)} =$

$c\}$. Since the second term is negative, we minimize it (the procedure is the same):

$$\begin{aligned}
&\nabla_{\mu_{ck}} \sum_{i=1}^{n}\sum_{j=1}^{D}[\log(2\pi) + \log \Sigma_{c^{(i)},jj} + \frac{(x_j^{(i)} - \mu_{c^{(i)},j})^2}{\Sigma_{c^{(i)},jj}}] \\
=&\nabla_{\mu_{ck}} \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{(x_j^{(i)} - \mu_{c^{(i)},j})^2}{\Sigma_{c^{(i)},jj}} = -\frac{2}{\Sigma_{c,jj}} \sum_{i=1}^{n} 1\{c^{(i)} = c\}(x_j^{(i)} - \mu_{cj}) \overset{set}{=} 0 \\
\implies &\hat{\mu}_{ck} = \frac{\sum_{i=1}^{n} 1\{c^{(i)} = c\}x_k^{(i)}}{\sum_{i=1}^{n} 1\{c^{(i)} = c\}}
\end{aligned}$$

$$\begin{aligned}
&\nabla_{\Sigma^{-1}_{c,kk}} \sum_{i=1}^{n}\sum_{j=1}^{D}[\log(2\pi) + \log \Sigma_{c^{(i)},jj} + \frac{(x_j^{(i)} - \mu_{c^{(i)},j})^2}{\Sigma_{c^{(i)},jj}}] \\
=&\nabla_{\Sigma^{-1}_{c,kk}} \sum_{i=1}^{n}\sum_{j=1}^{D}[-\log \Sigma^{-1}_{c^{(i)},jj} + \frac{(x_j^{(i)} - \mu_{c^{(i)},j})^2}{\Sigma_{c^{(i)},jj}}] \\
=&-\sum_{i=1}^{n} 1\{c^{(i)} = c\}[\Sigma_{c,kk} - (x_k^{(i)} - \mu_{ck})^2] \overset{set}{=} 0 \\
\implies &\hat{\Sigma}_{c,kk} = \frac{\sum_{i=1}^{n} 1\{c^{(i)} = c\}(x_k^{(i)} - \hat{\mu}_{ck})^2}{\sum_{i=1}^{n} 1\{c^{(i)} = c\}}.
\end{aligned}$$

2.10: Using Bayes' theorem,

$$\begin{aligned}
\log p(c|x^{(i)}, \mu, \Sigma, \pi) &= \log \frac{p(x^{(i)}, c|\mu, \Sigma, \pi)}{p(x^{(i)}|\mu, \Sigma, \pi)} = \log \frac{p(c|\pi)p(x^{(i)}|c, \mu, \Sigma)}{\sum_{c=1}^{28} p(x^{(i)}, c|\mu, \Sigma, \pi)} = \log \frac{p(c|\pi)p(x^{(i)}|c, \mu, \Sigma)}{\sum_{c=1}^{28} p(c|\pi)p(x^{(i)}|c, \mu, \Sigma)} \\
&= \log p(c|\pi) + \log p(x^{(i)}|c, \mu, \Sigma) - \log \sum_{c=1}^{28} \exp[\log p(c|\pi) + \log p(x^{(i)}|c, \mu, \Sigma)]
\end{aligned}$$

where $p(c|\pi) = \pi_c$ and $\log p(x^{(i)}|c, \mu, \Sigma) = -\frac{1}{2}\sum_{j=1}^{D}[\log(2\pi) + \log \Sigma_{c^{(i)},jj} + \frac{(x_j^{(i)} - \mu_{c^{(i)},j})^2}{\Sigma_{c^{(i)},jj}}]$.

3.1: Given $\{G, A\}$, there are no nodes that are independent of H. To see why, we first identify some triplets in the graph:

1. LCG is a causal chain triplet with G observed, meaning it is an active path.
2. CGE is a common effect triplet with G observed, meaning it is an active path.
3. GED is a common cause triplet with G observed, meaning it is an active path.
4. BED is a causal chain triplet, meaning it is an active path.
5. EDF is a causal chain triplet, meaning it is an active path.
6. DFH is a causal chain triplet, meaning it is an active path.

Hence, we can reach H from each node using the following paths (the numbers denote all triplets in each path according to the previous list):

- F to H using the path FH (6)
- D to H using the path DFH (6)

- E to H using the path $EDFH$ (5, 6)
- B to H using the path $BEDFH$ (4, 5, 6)
- C to H using the path $CGEDFH$ (2, 3, 5, 6)
- L to H using the path $LCGEDFH$ (1, 2, 3, 5, 6).

Given $\{G, F\}$, the nodes that are independent of H are $\{A, B, C, D, E, L\}$. To see why, we first identify some triplets in the graph:

1. CGH is a causal chain triplet with G observed, meaning it is an inactive path.
2. EGH is a causal chain triplet with G observed, meaning it is an inactive path.
3. DFH is a causal chain triplet with F observed, meaning it is an inactive path.

Hence, all possible paths from each node to H are inactive (the numbers denote an inactive triplet in each path according to the previous list):

- A to H must use the path $ACGH$ (1), $ABEGH$ (2), or $ABEDFH$ (3).
- B to H must use the path $BACGH$ (1), $BEGH$ (2), or $BEDFH$ (3).
- L to H must use the path $LCGH$ (1), $LCABEGH$ (2), or $LCABEDFH$ (3).
- C to H must use the path CGH (1), $CABEGH$ (2), or $CABEDFH$ (3).
- E to H must use the path $EBACGH$ (1), EGH (2), or $EDFH$ (3).
- D to H must use the path $DEBACGH$ (1), $DEGH$ (2), or DFH (3).

3.2: We use the shorthand $\sum_S$ to denote $\sum_{S \in \mathbf{S}}$ for any variable S and its domain $\mathbf{S}$.

$$
\begin{aligned}
p(A,G,C) &= \sum_B \sum_D \sum_E \sum_F \sum_H \sum_L p(A)p(B|A)p(L)p(C|A,L)p(E|B)p(G|C,E)p(D|E)p(F|D)p(H|FG) \\
&= p(A) \sum_B p(B|A) \sum_L p(L)p(C|A,L) \sum_E p(E|B)p(G|C,E) \sum_D p(D|E) \sum_F p(F|D) \sum_H p(H|FG) \\
&= p(A) \sum_B p(B|A) \sum_L p(L)p(C|A,L) \sum_E p(E|B)p(G|C,E) \sum_D p(D|E) \sum_F p(F|D) \\
&= p(A) \sum_B p(B|A) \sum_L p(L)p(C|A,L) \sum_E p(E|B)p(G|C,E) \sum_D p(D|E) \\
&= p(A) \sum_B p(B|A) \sum_L p(L)p(C|A,L) \sum_E p(E|B)p(G|C,E) \\
&= p(A) \sum_B p(B|A) \sum_L p(L)p(C|A,L)p(G|B,C) \\
&= p(A) \sum_L p(L)p(C|A,L) \sum_B p(B|A)p(G|B,C) \\
&= p(A) \sum_L p(L)p(C|A,L)p(G|A,C) \\
&= p(A)p(G|A,C) \sum_L p(L)p(C|A,L) \\
&= p(A)p(G|A,C)p(C|A) \\
p(G,C) &= \sum_A p(A,G,C) = \sum_A p(A)p(G|A,C)p(C|A) \\
p(A|G,C) &= \frac{p(A,G,C)}{p(G,C)} = \frac{p(A)p(G|A,C)p(C|A)}{\sum_A p(A)p(G|A,C)p(C|A)} \neq \frac{p(A,C)}{p(C)} = p(A|C).
\end{aligned}
$$

Thus, there is insufficient evidence that A and G are conditionally independent given C.

3.3: The set is the null set since there are no nodes independent of C_2 given the shaded nodes. To see why, we can apply the Bayes ball algorithm. From C_2, we can use active causal chain triplets to reach any node on the left side of the shaded node in each row. In addition, we can use the active common effect triplet of $C_2C_3B_3$ to reach B_3 from C_2. Then from B_3, we can use active causal chain triplets to reach any node on the right side of the shaded node in each row.