EE5907/EE5027 Week 1: Probability Review Solutions

Exercise 2.6

(a) According to Bayes Rule,

$$\vec{P}(H|e_1,e_2) = \frac{P(H,e_1,e_2)}{P(e_1,e_2)} = \frac{P(e_1,e_2|H)P(H)}{P(e_1,e_2)} \tag{1}$$

thus (ii) is sufficient for calculation

(b) Given $E_1 \perp E_2|H$, $P(e_1,e_2|H) = P(e_1|H)P(e_2|H)$
From (a), we have

$$\vec{P}(H|e_1,e_2) = \frac{P(e_1,e_2|H)P(H)}{P(e_1,e_2)} \tag{2}$$

From $E_1 \perp E_2|H$, we have

$$\vec{P}(H|e_1,e_2) = \frac{P(e_1|H)P(e_2|H)P(H)}{P(e_1,e_2)} \tag{3}$$

Eq. (3) corresponds to terms in (i). In addition, we can calculate $P(e_1,e_2)$ by $\sum_H(P(e_1,e_2|H)P(H))$, so (iii) is also sufficient.
To conclude, (i),(ii),(iii) are all sufficient.

Exercise 2.7

Proof by counter example:

(I) Let X_1 and X_2 be outcomes of independent coin toss (1 means head, 0 means tails). $X_3 = X_1 \oplus X_2$, where $\oplus$ is XOR operator. $p(X_3|X_1,X_2) \neq p(X_3)$ since X_1 and X_2 determines X_3 deterministically, so X_1, X_2, X_3 are not mutually independent. However, $p(X_3|X_1) = p(X_3)$, $p(X_3|X_2) = p(X_3)$, so X_1, X_2, X_3 are pairwise independent.

(II) Consider a tetrahedron die where three of the faces are colored red, green, and blue respectively. On the fourth face, include all the three colors. Roll the dice and define following events:

X_1 : "red appeared on the face the dice landed on"

X_2 : "green appeared on the face the dice landed on"

X_3 : "blue appeared on the face the dice landed on"

Therefore

$$P(X_i, X_j) = \frac{1}{4} = P(X_i)P(X_j)$$

$$\begin{aligned} P(X_1, X_2, X_3) &= \frac{1}{4} \\ &\neq P(X_1)P(X_2)P(X_3) = \frac{1}{8} \end{aligned}$$

Therefore X_1, X_2, X_3 are pairwise independent, but not mutually independent.

Exercise 2.8

Proof

($\Rightarrow$) Given $X \perp Y|Z$, we have $p(x, y|z) = p(x|z)p(y|z)$. Let $g(x, z) = p(x|z)$ and $h(y, z) = p(y|z)$, then $p(x, y|z) = g(x, z)h(y, z)$.

($\Leftarrow$) Suppose $p(x, y|z) = g(x, z)h(y, z)$. Integrate both sides over x (or summation if x is discrete)

$$\begin{aligned} &\int p(x, y|z)dx = \int g(x, z)dx \times h(y, z) \\ &\implies p(y|z) = G(z)h(y, z), \end{aligned} \tag{4}$$

where $G(z) = \int g(x, z)dx$.

Integrate both sides over y (or summation if y is discrete)

$$\begin{aligned} &\int p(x, y|z)dy = g(x, z) \times \int h(y, z)dy \\ &\implies p(x|z) = g(x, z)H(z), \end{aligned} \tag{5}$$

where $H(z) = \int h(y, z)dy$

Finally, let's integrate with respect to both x and y:

$$\begin{aligned} 1 = \int\int p(x,y|z)dxdy &= \int\int g(x,z)h(y,z)dxdy \qquad (6) \\ &= \int g(x,z)dx \int h(y,z)dy = G(z)H(z) \qquad (7) \end{aligned}$$

Therefore

$$\begin{aligned} p(x,y|z) = g(x,z)h(y,z) = \frac{p(x|z)}{G(z)}\frac{p(y|z)}{H(z)} \text{ using Eq. (4) and Eq. (5)} \\ = p(x|z)p(y|z) \text{ using Eq. (7)} \qquad (8) \end{aligned}$$

Exercise 2.10

According to the "change of variable formula"

$$p_y(y) = p_x(x)\left|\frac{dx}{dy}\right|$$

In this case, $y = \frac{1}{x} \quad \Rightarrow \quad \frac{dy}{dx} = -\frac{1}{x^2} \quad \Rightarrow \quad \frac{dx}{dy} = -x^2$

$$p_y(y) = \frac{b^a}{\Gamma(a)}x^{a-1}e^{-xb} \cdot |-x^2| \qquad (9)$$

Substitute x by $1/y$

$$p_y(y) = \frac{b^a}{\Gamma(a)}y^{-a+1}e^{-\frac{b}{y}} \cdot y^{-2} = \frac{b^a}{\Gamma(a)}y^{-(a+1)}e^{-\frac{b}{y}} \qquad (10)$$

Since $IG(x|\text{shape} = a, \text{scale} = b) = \frac{b^a}{\Gamma(a)}x^{-(a+1)}e^{-\frac{b}{x}}$, y is $IG(a,b)$

Exercise 2.12

According to definition

$$
\begin{aligned}
I(X;Y) \triangleq KL(p(X,Y)||p(X)p(Y)) &= \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\
&= \sum_x \sum_y p(x,y) \log \frac{p(y|x)}{p(y)} \\
&= \sum_x \sum_y p(x,y) \log p(y|x) - \sum_x \sum_y p(x,y) \log p(y) \\
&= \sum_x \sum_y p(x)p(y|x) \log p(y|x) - \sum_y p(y) \log p(y) \\
&= -\sum_x p(x) H(Y|X=x) + H(Y) \\
&= -H(Y|X) + H(Y)
\end{aligned}
$$

By symmetry of the above derivation, $I(X;Y) = H(X) - H(X|Y)$.

Exercise 2.16

According to definition of Beta distribution

$$
\text{Beta}(x|a,b) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} \;\; 0 \leq x \leq 1, \tag{11}
$$

where

$$
B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} = \int_0^1 x^{a-1}(1-x)^{b-1} dx \tag{12}
$$

Mean:

$$
\begin{aligned}
E[X] &= \int_0^1 x \times \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}\, dx \\
&= \frac{1}{B(a,b)} \int_0^1 x^a (1-x)^{b-1}\, dx \\
&= \frac{B(a+1,b)}{B(a,b)} \qquad \text{from Eq. (12)} \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} \\
&= \frac{\Gamma(a+b)\Gamma(a+1)}{\Gamma(a)\Gamma(a+b+1)} \\
&= \frac{\Gamma(a+b)\Gamma(a)a}{\Gamma(a)\Gamma(a+b)(a+b)} \\
&= \frac{a}{a+b},
\end{aligned}
$$

where we have used the property that $\Gamma(t+1) = t\Gamma(t)$.

To compute variance, we first compute

$$E[X^2] = \frac{B(a+2,b)}{B(a,b)} = \frac{\Gamma(a+b)\Gamma(a+2)\Gamma(b)}{\Gamma(a)\Gamma(b)\Gamma(a+b+2)} = \frac{a(a+1)}{(a+b)(a+b+1)}$$

Then

$$\begin{aligned}
\text{variance} &= E[X^2] - E^2[X] \\
&= \frac{a(a+1)}{(a+b)(a+b+1)} - \left(\frac{a}{a+b}\right)^2 \\
&= \frac{a(a+1)(a+b) - a^2(a+b+1)}{(a+b)^2(a+b+1)} \\
&= \frac{ab}{(a+b)^2(a+b+1)}
\end{aligned}$$

To obtain mode, we want:

$$\begin{aligned}
\underset{x}{\operatorname{argmax}} \operatorname{Beta}(x|a,b) &= \underset{x}{\operatorname{argmax}} \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} \\
&= \underset{x}{\operatorname{argmax}} \log \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} \\
&= \underset{x}{\operatorname{argmax}} (a-1)\log x + (b-1)\log(1-x)
\end{aligned}$$

Differentiating with respect to x and set to 0, we get:

$$\begin{aligned}
&\frac{a-1}{x} - \frac{b-1}{1-x} = 0 \\
\implies &(a-1)(1-x) = (b-1)x \\
\implies &x = \frac{a-1}{a+b-2}
\end{aligned}$$

There are several cases here:

- When $a > 1$ and $b > 1$, the distribution is concave, and so mode is $\frac{a-1}{a+b-2}$.
- When $a = b = 1$, we have a uniform distribution, so the mode is any value between 0 and 1.
- When $a = b$ and both are less than 1, then we get a convex distribution symmetric about 0.5, so the modes are at 0 and 1.
- When $a > b$ and $b \leq 1$, then the distribution is convex and mode is 1.
- When $b > a$ and $a \leq 1$, then the distribution is convex and mode is 0.

EE5907/EE5027 Week 2: Probabilistic Estimation + Conjugate Priors

Exercise 3.1

The likelihood is given by

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

Hence the log-likelihood is given by

$$\log p(\mathcal{D}|\theta) = N_1 \log\theta + N_0 \log(1-\theta) \tag{2}$$

To optimize the log-likelihood, we get

$$\operatorname*{argmax}_\theta p(\mathcal{D}|\theta) = \operatorname*{argmax}_\theta (N_1 \log\theta + N_0 \log(1-\theta)) \tag{3}$$

Differentiating with respect to θ and set to 0, we get:

$$\begin{aligned} &\frac{N_1}{\theta} - \frac{N_0}{1-\theta} = 0 \\ &\Longrightarrow N_1(1-\theta) = N_0\theta \\ &\Longrightarrow \theta = \frac{N_1}{N_1+N_0} \\ &\Longrightarrow \theta = \frac{N_1}{N} \end{aligned}$$

Hence, $\hat{\theta}_{MLE} = \frac{N_1}{N}$

Exercise 3.6

The Poisson distribution can be represented as:

$$\mathcal{D} = (x_1, x_2, \cdots, x_n), \mathcal{D} \sim Poi(\lambda) \tag{4}$$

The likelihood is given by

$$p(\mathcal{D}|\lambda) = \prod_{i=1}^{n} e^{-\lambda}\frac{\lambda^{x_i}}{x_i!} \tag{5}$$

To optimize the log-likelihood, we get

$$\begin{aligned}
\hat{\lambda}_{MLE} &\triangleq \underset{\lambda}{\text{argmax}} \log \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} \\
&= \underset{\lambda}{\text{argmax}} \sum_{i=1}^{n} \log \left(e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} \right) \\
&= \underset{\lambda}{\text{argmax}} \sum_{i=1}^{n} (-\lambda + x_i \log \lambda - \log x_i!) \\
&= \underset{\lambda}{\text{argmax}} \left(-n\lambda + \sum_{i=1}^{n} x_i \log \lambda - \sum_{i=1}^{n} \log x_i! \right) \\
&= \underset{\lambda}{\text{argmax}} \left(-n\lambda + \sum_{i=1}^{n} x_i \log \lambda \right)
\end{aligned}$$

Differentiating with respect to λ and set to 0, we get:

$$\begin{aligned}
& -n + \frac{1}{\lambda} \sum_{i=1}^{n} x_i = 0 \\
\Longrightarrow\ & \hat{\lambda}_{MLE} = \frac{1}{n} \sum_{i=1}^{n} x_i
\end{aligned}$$

Exercise 3.7

a. Multiply the likelihood by the conjugate prior given in the question, we get the following posterior:

$$\begin{aligned}
& p(\lambda|\mathcal{D}) \propto p(\mathcal{D}|\lambda)p(\lambda) \propto e^{-n\lambda} \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} \lambda^{a-1} e^{-\lambda b} \\
\Longrightarrow\ & p(\lambda|\mathcal{D}) \propto \frac{1}{\prod_{i=1}^{n} x_i!} e^{-(n+b)\lambda} \lambda^{a-1+\sum_{i=1}^{n} x_i} \\
\Longrightarrow\ & p(\lambda|\mathcal{D}) \propto \lambda^{a-1+\sum_{i=1}^{n} x_i} e^{-(n+b)\lambda} \\
\Longrightarrow\ & p(\lambda|\mathcal{D}) = Ga \left(\lambda|\, a + \sum_{i=1}^{n} x_i, n+b \right)
\end{aligned}$$

b. Given the mean of Gamma distribution $Ga\,(a,b)$ is $\frac{a}{b}$, we can get the mean of $p(\lambda|\mathcal{D})$ to be

$$\bar{\theta} = \frac{a + \sum_{i=1}^{n} x_i}{n+b} \tag{6}$$

Given that $a \to 0$ and $b \to 0$, we have

$$\lim_{a\to 0, b\to 0} \frac{a + \sum_{i=1}^n x_i}{n+b} = \frac{\sum_{i=1}^n x_i}{n} = \bar{x}$$

Hence, the posterior mean converges to the ML solution.

Exercise 3.12

a. The posterior of the Bernoulli

$$p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta)$$

if $\theta = 0.5$,

$$\begin{aligned} &p(\mathcal{D}|\theta)p(\theta) = 0.5^{N+1} \\ \implies &\log p(\mathcal{D}|\theta)p(\theta) = (N+1)\log 0.5 \end{aligned}$$

if $\theta = 0.4$,

$$\begin{aligned} &p(\mathcal{D}|\theta)p(\theta) = 0.4^{N_1} 0.6^{N-N_1} 0.5 \\ \implies &\log p(\mathcal{D}|\theta)p(\theta) = N_1 \log 0.4 + (N - N_1)\log 0.6 + \log 0.5 \end{aligned}$$

if $\theta =$others,

$$p(\mathcal{D}|\theta)p(\theta) = 0$$

For 0.5 to win out over 0.4,

$$\begin{aligned} &(N+1)\log 0.5 > N_1 \log 0.4 + (N - N_1)\log 0.6 + \log 0.5 \\ \implies &N \log \frac{0.5}{0.6} > N_1 \log \frac{0.4}{0.6} \\ \implies &\frac{N_1}{N} > \frac{\log 5/6}{\log 2/3} = \frac{\log 1.2}{\log 1.5} = 0.4497 \text{ because } \log 2/3 \text{ is negative} \end{aligned}$$

Therefore, we have

$$\hat{\theta}_{MAP} = \begin{cases} 0.4 & \text{if } \frac{N_1}{N} < \frac{\log 1.2}{\log 1.5} \\ 0.5 & \text{if } \frac{N_1}{N} > \frac{\log 1.2}{\log 1.5} \end{cases}$$

Note that N_1/N can never be exactly equal to $\frac{\log 1.2}{\log 1.5}$ because $\frac{\log 1.2}{\log 1.5}$ is irrational.

b. If N is large, then the MAP estimate (with the usual beta prior) will approach the true value of 0.41. However, the biased-coin prior will still lead to an estimate of 0.4, resulting in a difference of 0.01 from the true value. Therefore the biased-coin prior does not lead to a consistent estimator.

   If N is small, the unbiased coin prior might possibly be off by a lot. For example, if $N = 1$ and the outcome of the coin toss is head. Then if $\alpha = \beta = 1$, the unbiased coin prior would lead to a MAP estimate of $\hat{\theta} = 1$. On the other hand, the biased coin prior will lead to a MAP estimate of 0.5, which is not that different from 0.41.

Exercise 3.14

a. Denote the counts of each alphabet by N_j. If we use a $\text{Dir}(\alpha)$ prior for θ, the posterior predictive is just

$$p(x = k|\mathcal{D}) = \frac{\alpha_k + N_k}{\sum_{k'}(\alpha_{k'} + N_{k'})}$$

   Substitute $a_k = 10$ and the number of "e" is 260, we have

$$\begin{aligned} p(x_{2001} = \text{e}|\mathcal{D}) &= \frac{10 + 260}{270 + 2000} \\ &= 0.119 \end{aligned}$$

b. Similar to part (a), we easily derive that

$$\begin{aligned} p(x_{2001} = \text{p}|\mathcal{D}) &= \frac{10 + 87}{270 + 2000} \\ &= 0.043 \end{aligned}$$

   Then we have,

$$\begin{aligned} p\left(x_{2001} = p, x_{2002} = a\right| D) &= P\left(x_{2002} = a\right| x_{2001} = p, D) \, P\left(x_{2001} = P\right| D) \\ &= \frac{\alpha_j + N_j}{\sum_{j'}\left(\alpha_{j'} + N_{j'}\right)} P\left(x_{2001} = P|D\right) \\ &= \frac{10 + 100}{270 + 2001} * 0.043 \\ &= 0.00207 \end{aligned}$$

EE5907/EE5027 Week 3: Univariate Gaussian + Naive Bayes

Q1: Mixed Observations Naive Bayes

(i)

$$\begin{aligned}
p(y|x_1 = 0) &= \frac{p(y)p(x_1 = 0|y)}{p(x_1 = 0)} \\
&\propto p(y)p(x_1 = 0|y) \\
&\propto p(y) \qquad \text{because feature is uninformative.} \\
&= [0.5 \;\; 0.25 \;\; 0.25]
\end{aligned}$$

x_1 is uninformative because $\theta = [0.5 \;\; 0.5 \;\; 0.5]$, i.e., the probability of getting a head is the same for all classes. Note that x_1 will also be uninformative if $\theta = [0.4 \;\; 0.4 \;\; 0.4]$.

(ii)

$$\begin{aligned}
p(y|x_2 = 0) &= \frac{p(y)p(x_2 = 0|y)}{p(x_2 = 0)} \\
&\propto p(y)p(x_2 = 0|y) \\
&= \pi_y \frac{1}{\sqrt{2\pi\sigma_y^2}} e^{-\frac{1}{2\sigma_y^2}(0-\mu_y)^2} \\
&\propto \pi_y e^{-0.5\mu_y^2} \\
&= [0.5e^{-0.5} \;\; 0.25e^{0} \;\; 0.25e^{-0.5}] \\
&= [0.3033 \;\; 0.25 \;\; 0.1516]
\end{aligned}$$

Therefore

$$\begin{aligned}
p(y|x_2 = 0) &= [0.3033 \;\; 0.25 \;\; 0.1516]/(0.3033 + 0.25 + 0.1516) \\
&= [0.4302 \;\; 0.3547 \;\; 0.2151]
\end{aligned}$$

(iii)

$$\begin{aligned}
p(y|x_1 = 0, x_2 = 0) &= p(y|x_2 = 0) \qquad \text{because } x_1 \text{ is uninformative.} \\
&= [0.4302 \;\; 0.3547 \;\; 0.2151]
\end{aligned}$$

Exercise 3.20

(a) Since vector x is a D bit vector, and the value of each bit is binary, the vector x can take on one of 2^D possible configurations. For the full model, each of the 2^D configuration is free to take on any values, except for the one constraint that the sum of probabilities need to sum to 1. Therefore, the full model of $p(x|y=c)$ has $2^D - 1$ parameters.

(b) The naive bayes is likely to give a lower test error. With few training samples N, the naive bayes is less likely to overfit.

(c) The full model will perform better. The full model is a more accurate model. With large N, we can reliably estimate all the parameters without overfitting.

(d) **Fitting full model**
For each class c, there is a $2 \times \cdots \times 2$ matrix (D-dimensional matrix) of probability M_c we are trying to estimate. We start by initializing each M_c to 0. For each datapoint, we then add the count to the appropriate M_c (based on the class label of the datapoint). We finally normalize each M_c so that the resulting matrix sums to 1. See Algorithm 1.

Algorithm 1 Fitting Full Model

```
%Initialization
for each $c \in C$ do
    $N_c \leftarrow 0$
    $M_c \leftarrow 0$
end for

%Count
for $n = 1 : N$ do
    index ← ComputeIndexIntoM($x_n$)
    $M_{y_n}$(index) ← $M_{y_n}$(index) +1
    $N_{y_n} \leftarrow N_{y_n} + 1$
end for

%Normalization
for each $c \in C$ do
    $M_c \leftarrow M_c / N_c$
    $N_c \leftarrow N_c / N$
end for
```

The initialization requires $O(C2^D)$ operations, the counting requires $O(ND)$ operations and the normalization requires $O(C2^D)$ operations. Therefore the computational complexity of the full model is $O(ND) + O(2^D C)$.

Fitting naive Bayes Model
The process of fitting naive Bayes Model is the same as full model, but the computation complexities for "initialization", "counting" and "normalization" are $O(CD)$, $O(ND)$, and $O(CD)$ respectively. So, the computational complexity of the Naive Bayes Model is $O\left(ND\right) + O\left(DC\right)$.

(e) **Full model**

$$\underset{y}{\operatorname{argmax}}\, p(y|x_1, \cdots, x_D, \theta) = \underset{y}{\operatorname{argmax}}\, p\left(x_1, \cdots, x_D \middle| y, \theta\right) p(y|\theta)$$

For a given x, we can compute the index in $O(D)$ time. For each class c, we can then grab the corresponding entry in each M_c, multiply with $p(c)$ to obtain the posterior probability for class c (this takes O(C) time). We can then compare the posterior probability among the different classes and return the class with the highest posterior posterior probability (this takes O(C) time). Therefore the runtime is $O(D+C)$.

Naive Bayes Model

$$\begin{aligned}
&\underset{y}{\operatorname{argmax}}\, p\left(y \mid x, \theta\right) \\
=&\underset{y}{\operatorname{argmax}}\, p\left(x \mid y, \theta\right) P\left(y \mid \theta\right) \\
=&\underset{y}{\operatorname{argmax}}\, p\left(x_1 \mid y, \theta\right) p\left(x_2 \mid y, \theta\right) \cdots p\left(x_D \mid y, \theta\right) p\left(y \mid \theta\right)
\end{aligned}$$

For a given x and given class c, we have to multiply D likelihood values with $p(c)$ to obtain the posterior probability (this takes O(CD) time). We can then compare the posterior probability among the different classes and return the class with the highest posterior posterior probability (this takes O(C) time). Therefore the runtime is $O(CD)$ time.

(f)

$$p(y|x_v,\theta) = \frac{p(x_v|y,\theta)p(y|\theta)}{p(x_v|\theta)} = \frac{p(y|\theta)\sum_{x_h} p(x_v,x_h|y,\theta)}{p(x_v|\theta)} \tag{1}$$

the optimization is only based on the numerator $p(y|\theta)\sum_{x_h} p(x_v,x_h|y,\theta)$

Full model $\sum_{x_h} p(x_v,x_h|y,\theta)$ cannot be simplified further, so we have to compute $p(x_v,x_h|y,\theta)$ where x_v corresponds to the test case, and x_h corresponds to a particular configuration (with 2^h possible configurations) and sum them all together. So the computational complexity is $O(2^h(C+D))$ (assuming for each configuration x_v, x_h, we compute the index and use it for all classes c).

Naive Bayes model
$\sum_{x_h} p(x_v,x_h|y,\theta) = p(x_v|y,\theta)$, so we can simply ignore the hidden variables. So the computational complexity is $O\left(CV\right)$, where V is the number of visible variables.

Q3: Posterior Predictive Distribution for Exponential Distribution

(a)
- (i)
  -

$$\begin{aligned}
\lambda_{ML} &= \underset{\lambda}{\operatorname{argmax}}\, p(\{x_1,\cdots,x_N\}|\lambda) \\
&= \underset{\lambda}{\operatorname{argmax}} \prod_{n=1}^{N} p(x_n|\lambda) \\
&= \underset{\lambda}{\operatorname{argmax}} \prod_{n=1}^{N} \lambda e^{-\lambda x_n} \\
&= \underset{\lambda}{\operatorname{argmax}}\, N\log\lambda - \lambda\sum_{n=1}^{N} x_n
\end{aligned}$$

Differentiating with respect to λ, we get

$$\frac{N}{\lambda} - \sum_{n=1}^{N} x_n$$

Setting it to zero, we get

$$\frac{N}{\lambda} - \sum_{n=1}^{N} x_n = 0$$

$$\lambda = \frac{N}{\sum_{n=1}^{N} x_n}$$

- (ii)
  - The problem with this approach is that using the plug-in estimator to predict new data x_{N+1} will be overly confident.
  - One could place a prior on λ and then compute the posterior predictive distribution of new data x_{N+1}

(b) (i)

$$\begin{aligned}
p(\lambda|D) &= \frac{p(D|\lambda)p(\lambda)}{p(D)} \\
&\propto p(D|\lambda)p(\lambda) \\
&= \lambda^N e^{-\lambda \sum_{n=1}^{N} x_n} \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\lambda\beta} \\
&\propto \lambda^{N+\alpha-1} e^{-\lambda\left[\beta+\sum_{n=1}^{N} x_n\right]} \\
&= \text{Gamma}(\lambda; \alpha', \beta'),
\end{aligned}$$

where $\alpha' = N + \alpha$ and $\beta' = \beta + \sum_{n=1}^{N} x_n$

(ii) A direct approach is to compute

$$p(x_{n+1}|D) = \int p(x_{n+1}|\lambda)p(\lambda|D)d\lambda,$$

by exploiting conjugate properties of the Poisson and Gamma distributions. An easier approach is to exploit Bayes' rule to compute the evidence:

$$\begin{aligned}
p(x_{N+1}|D) &= \frac{p(x_{N+1}|\lambda, D)p(\lambda|D)}{p(\lambda|x_{N+1}, D)} \\
&= \frac{p(x_{N+1}|\lambda)p(\lambda|D)}{p(\lambda|x_{N+1}, D)} \\
&= \frac{\text{exponential}(x_{N+1}; \lambda)\text{Gamma}(\lambda; \alpha', \beta')}{\text{Gamma}(\lambda; \alpha'', \beta'')}
\end{aligned}$$

To evaluate above, observe that $\alpha'' = \alpha + N + 1$ and $\beta'' = \beta + \sum_{n=1}^{N+1} x_n$

$$\begin{aligned}
p(x_{n+1}|D) &= \frac{\lambda e^{-\lambda x_{N+1}} \frac{(\beta+\sum_{n=1}^{N} x_n)^{\alpha+N}}{\Gamma(\alpha+N)} \lambda^{\alpha+N-1} e^{-\lambda(\beta+\sum_{n=1}^{N} x_n)}}{\frac{(\beta+\sum_{n=1}^{N+1} x_n)^{\alpha+N+1}}{\Gamma(\alpha+N+1)} \lambda^{\alpha+N} e^{-\lambda(\beta+\sum_{n=1}^{N+1} x_n)}} \\
&= \frac{(\beta + \sum_{n=1}^{N} x_n)^{\alpha+N}}{\Gamma(\alpha+N)} \times \frac{\Gamma(\alpha+N+1)}{(\beta+\sum_{n=1}^{N+1} x_n)^{\alpha+N+1}} \\
&= (\alpha+N) \frac{(\beta + \sum_{n=1}^{N} x_n)^{\alpha+N}}{(\beta+\sum_{n=1}^{N+1} x_n)^{\alpha+N+1}}
\end{aligned}$$

EE5907/EE5027 Week 4: Logistic Regression Solutions

Exercise 8.3

a. Given $\sigma(a) = \frac{1}{1+e^{-a}}$

$$\begin{aligned}
\frac{d\sigma(a)}{da} &= -\frac{1}{(1+e^{-a})^2}\frac{d(1+e^{-a})}{da} \\
&= \frac{1}{(1+e^{-a})^2}e^{-a} \\
&= \frac{1}{(1+e^{-a})^2}(1+e^{-a}-1) \\
&= \frac{1}{1+e^{-a}} - \frac{1}{(1+e^{-a})^2} \\
&= \sigma(a) - \sigma^2(a) \\
&= \sigma(a)(1-\sigma(a))
\end{aligned}$$

b. Given $NLL(w) = -\sum_{i=1}^{N}[y_i \log \mu_i + (1-y_i)\log(1-\mu_i)]$ where $\mu_i = \sigma(w^T x_i)$, we have

$$\begin{aligned}
g = \frac{d}{dw}NLL(w) &= \frac{d}{dw}\left(-\sum_{i=1}^{N}[y_i \log \mu_i + (1-y_i)\log(1-\mu_i)]\right) \\
&= \left(-\sum_{i=1}^{N}\frac{d}{d\mu_i}[y_i \log \mu_i + (1-y_i)\log(1-\mu_i)]\frac{d\mu_i}{d(w^T x_i)}\frac{dw^T x_i}{dw}\right) \\
&= -\sum_{i=1}^{N}\left[\left(\frac{y_i}{\mu_i} - \frac{1-y_i}{1-\mu_i}\right)\frac{d\mu_i}{d(w^T x_i)}\frac{dw^T x_i}{dw}\right] \\
&= -\sum_{i=1}^{N}\left[\frac{y_i - y_i\mu_i - \mu_i + \mu_i y_i}{\mu_i(1-\mu_i)}\mu_i(1-\mu_i)x_i\right] \\
&= \sum_i(\mu_i - y_i)x_i
\end{aligned}$$

c. The Hessian is positive definite if $z^T Hz$ is positive for all $z \in \mathbb{R}^n$ and $z \neq \vec{0}$.

$$z^T Hz = z^T X^T SXz = (Xz)^T S(Xz)$$

Since X is full rank and $z \neq \vec{0}$, then $Xz \neq \vec{0}$. Let the i-th entry of Xz be a_i, then

$$(Xz)^T S(Xz) = \sum_{i=1}^{n} a_i^2 \mu_i (1 - \mu_i)$$

Since $0 < \mu_i < 1$, therefore $a_i^2 \mu_i (1 - \mu_i) \geq 0$ for all i with at least one term greater than 0 because $Xz \neq \vec{0}$. Therefore $z^T H z$ is greater than 0, and H is positive definite.

Exercise 8.6

a. False. According to the proof in Exercise 8.3, the Hessian is positive definite and thus $NLL(w)$ is convex. $\lambda ||w||_2^2$ is also convex. Thus $J(w)$ is also convex because the sum of two convex functions is convex. Therefore there is only one local optimum and that local optimum is also a global optimum.

b. False. For L2 regularization, we tend to not get sparse estimates. The intuitive reason is that as a number x decreases from 10 to 5 to 0, x^2 changes from 100 to 25 to 0. Therefore decreasing x by a constant amount (10 to 5 to 0) yields diminishing returns (100 to 25 to 0). In contrast, $L1$ regularization tends to lead to sparse estimates because as x decreases from 10 to 5 to 0, $|x|$ changes from 10 to 5 to 0, thus there is no such diminishing returns.

c. True. If the training data is linearly separable, the MLE is obtained when $||w|| \to \infty$ without any regularization. Thus some weight ω_j might become infinite.

d. True. The log likelihood of the training set will decrease as λ increases because more weight is given to the regularization so w is not as "free" to fit the data. Therefore, the$NLL(w)$ of training set will increase as λ increases.

e. False. As we increase λ, we can potentially improve results on the test set and so the $NLL(w)$ of the test set can initially decrease. However, as we increase λ, $NLL(w)$ of both training and test sets can become bad (i.e., increase).

Exercise 8.7

a. The decision boundary will satisfy

$$p(y = 1|x) = \text{sigm}(\omega^T x) = 0.5 \implies \omega^T x = 0$$

Thus we have $\omega_0 + \omega_1 x_1 + \omega_2 x_2 = 0$, which implies the possible decision boundary should be a straight line. The rough decision boundary is displayed as the red line in Figure 1. There is zero classification error made on the training set. Note that although there are many possible decision boundaries that result in zero classification error, but there is a unique value of w (and hence unique decision boundary) that minimize $J(w)$.

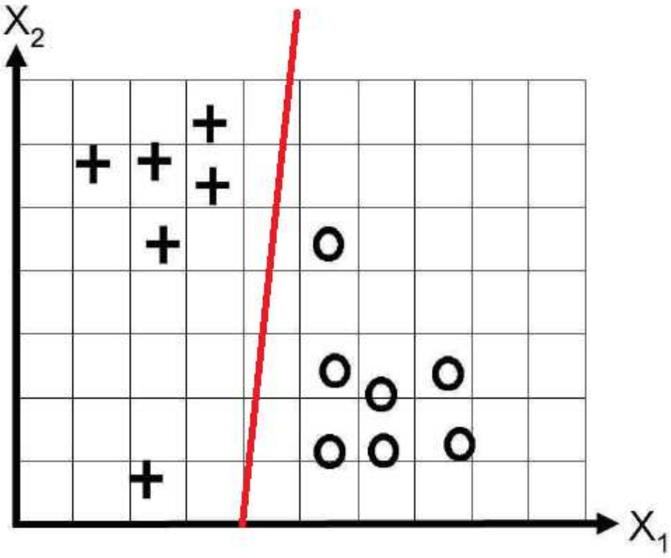

Figure 1: Decision boundary for part (a)

b. Suppose ω_0 is regularized all the way to zero. Then we have $x_2 = -\frac{\omega_1}{\omega_2} x_1$. Thus a decision boundary will definitely pass through the origin. The red line in Figure 2 shows the possible decision boundary. There is one classification error made on the training set. Again note that while there are many decision boundaries that result in one classification error, there is a unique decision boundary that minimize $J(w)$.

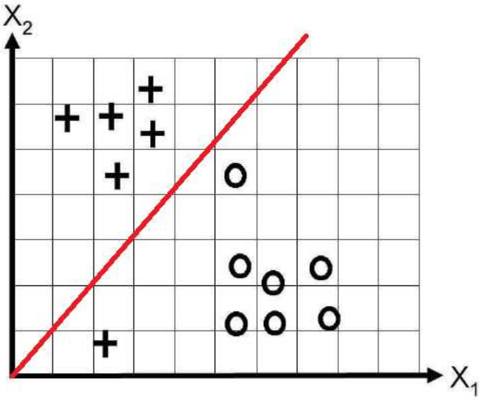

Figure 2: Decision boundary for part (b)

c. Suppose ω_1 is heavily regularized. Then we have $x_2 = -\frac{\omega_1}{\omega_2}x_1 - \frac{\omega_0}{\omega_2}$ where the coefficient of x_1 is approximately zero and $-\frac{\omega_0}{\omega_2}$ is a constant . Thus the decision boundary will be horizontal. The red line in Figure 3 shows a possible decision boundary. There are two classification errors made on the training set. Again note that while there are many decision boundaries that result in two classification errors, there is a unique decision boundary that minimize $J(w)$.

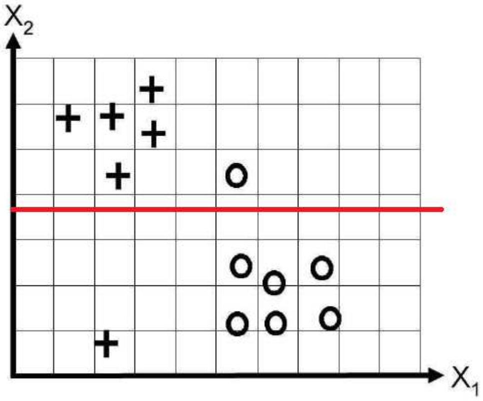

Figure 3: Decision boundary for part (c)

d. Suppose ω_2 is heavily regularized. Then we have $x_1 = -\frac{\omega_2}{\omega_1}x_2 - \frac{\omega_0}{\omega_1}$ where the coefficient of x_2 is approximately zero and $-\frac{\omega_0}{\omega_1}$ is a constant . Thus the decision boundary will be vertical. The red line in Figure 3 shows a possible decision boundary. There is zero classification error made on the training set. Note that although there are many possible decision boundaries that result in zero classification error, but there is a unique value of w (and hence unique decision boundary) that minimize $J(w)$.

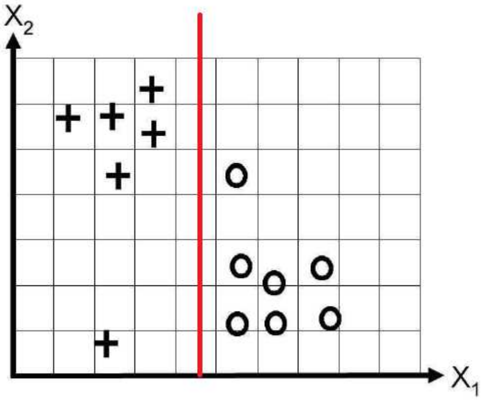

Figure 4: Decision boundary for part (d)

EE5907/EE5027 Week 5: Non-parametric Solutions

Q1: Parzen's Window

- $p_h(x) = \frac{1}{4}\sum_{n=1}^{4} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_n - x)^2}{2}}$
- Plugging x = 2 and the data, we get

$$\begin{aligned} p_1(x=2) &= \frac{1}{4\sqrt{2\pi}}\left[e^{-\frac{(2-1)^2}{2}} + e^{-\frac{(2-3)^2}{2}} + e^{-\frac{(2-4)^2}{2}} + e^{-\frac{(2-10)^2}{2}}\right] \\ &= \frac{1}{4\sqrt{2\pi}}\left[e^{-0.5} + e^{-0.5} + e^{-2} + e^{-32}\right] \\ &= 0.134483 \end{aligned}$$

- Plugging x = 5 and the data we get

$$\begin{aligned} p_1(x=5) &= \frac{1}{4\sqrt{2\pi}}\left[e^{-\frac{(5-1)^2}{2}} + e^{-\frac{(5-3)^2}{2}} + e^{-\frac{(5-4)^2}{2}} + e^{-\frac{(5-10)^2}{2}}\right] \\ &= \frac{1}{4\sqrt{2\pi}}\left[e^{-8} + e^{-2} + e^{-0.5} + e^{-12.5}\right] \\ &= 0.0740 \end{aligned}$$

Q2: KNN

- The 3 closest datapoints for x_5 are x_1, x_2, and x_3.
- Therefore $p(y=1|x) = 1/3$ and $p(y=0|x) = 2/3$
- Therefore the datapoint should be classified as class 0
- The 3 closest datapoints for x_6 are x_2 (or x_1), x_3, and x_4.
- Therefore $p(y=1|x) = 2/3$ and $p(y=0|x) = 1/3$ (Note that x_1 and x_2 are equidistant, so I am also ok with $p(y=1|x) = p(y=0|x) = 1/2$)
- Therefore the datapoint should be classified as class 1

EE5907/EE5027 Week 6: Bayesian Statistics Solutions

Exercise 5.1

Given that $p(\theta) = \sum_k p(z = k)p(\theta|z = k) = \sum_{z=1}^{K} p(\theta, z)$, where $p(\theta|\mathcal{D})$ is conjugate, and $p(z = k)$ are the (prior) mixing weights, we get

$$\begin{aligned}
p(\theta|\mathcal{D}) &= \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} \\
&= \frac{p(\mathcal{D}|\theta)\sum_z p(\theta, z)}{p(\mathcal{D})} \\
&= \sum_z \frac{p(\mathcal{D}|\theta)p(\theta, z)}{p(\mathcal{D})} \\
&= \sum_z \frac{p(\mathcal{D}|\theta, z)p(\theta, z)}{p(\mathcal{D})} \quad \text{because } z \text{ and } D \text{ are conditionally independent given } \theta \\
&= \sum_z \frac{p(D, \theta, z)}{p(\mathcal{D})} \\
&= \sum_z \frac{p(\mathcal{D}, z)p(\theta|\mathcal{D}, z)}{p(\mathcal{D})} \\
&= \sum_z p(z|\mathcal{D})p(\theta|\mathcal{D}, z)
\end{aligned}$$

Exercise 5.3

a. The minimum risk is obtained if the posterior expected loss

$$\begin{aligned}
\rho(a|x) = \mathbb{E}_{p(y|x)}[L(y, a)] &= \sum_y L(y, a)p(y|x) \\
&= \begin{cases} 0 \cdot p(y = a|x) + \sum_{y \neq a} \lambda_s p(y|x) & 1 \leq a \leq C \\ \sum_y \lambda_r p(y|x) & a = C + 1 \end{cases} \\
&= \begin{cases} \sum_{y \neq a} \lambda_s p(y|x) & 1 \leq a \leq C \\ \lambda_r & a = C + 1 \end{cases}
\end{aligned}$$

is minimized over a. Therefore, for optimal a to be j, where $1 \leq j \leq C$, we require

$$\rho(a = j|x) \leq \rho(a = i|x) \text{ for all } i \neq j,$$

which can be split into two cases: (1) $1 \leq i \leq C$ and $i \neq j$, and (2) $i = C + 1$ and $i \neq j$:

Case 1: $1 \leq i \leq C$ and $i \neq j$, then

$$\begin{aligned}
\rho(a=j|x) &\leq \rho(a=i|x) \\
\sum_{y \neq j} \lambda_s p(y|x) &\leq \sum_{y \neq i} \lambda_s p(y|x) \\
\lambda_s p(y=i|x) &\leq \lambda_s p(y=j|x) \\
p(y=i|x) &\leq p(y=j|x)
\end{aligned}$$

Case 2: $i = C+1$ and $i \neq j$

$$\begin{aligned}
\rho(a=j|x) &\leq \rho(a=i|x) \\
\sum_{y \neq j} \lambda_s p(y|x) &\leq \sum_{y} \lambda_r p(y|x) \\
\lambda_s(1-p(y=j|x)) &\leq \lambda_r \\
1-p(y=j|x) &\leq \frac{\lambda_r}{\lambda_s} \\
p(y=j|x) &\geq 1-\frac{\lambda_r}{\lambda_s}
\end{aligned}$$

b. For $\frac{\lambda_r}{\lambda_s} = 0$, there is zero cost in rejection, so we should always reject. As $\lambda_r/\lambda_s \to 1$, the cost of rejection is the same as the cost of guessing the wrong label. Therefore we should always guess the class label with the highest posterior probability and never guess the reject option.

Exercise 5.7

Given that the expectation over Δ is with respect to

$$p(\Delta|\mathcal{D}) = \sum_{m \in M} p(\Delta|m,\mathcal{D})p(m|\mathcal{D})$$

We have for Bayes model averaging:

$$\begin{aligned}
&\mathbb{E}_{p(\Delta|\mathcal{D})}\left[L\left(\Delta, p^{BMA}\right)\right] \\
&= \mathbb{E}_{p(\Delta|\mathcal{D})}\left[-\log p^{BMA}(\Delta)\right] \\
&= -\int \sum_{m \in \mathcal{M}} p(\Delta|m,\mathcal{D})p(m|\mathcal{D}) \log \sum_{m \in \mathcal{M}} p(\Delta|m,\mathcal{D})p(m|\mathcal{D})\, d\Delta
\end{aligned}$$

Similarly, for plugin approximation:

$$\begin{aligned}
&\mathbb{E}_{p(\Delta|\mathcal{D})}\left[L\left(\Delta, p^{M}\right)\right] \\
&= \mathbb{E}\left[-\log p^{M}(\Delta)\right] \\
&= \mathbb{E}_{p(\Delta|\mathcal{D})}\left[-\log p(\Delta|m',\mathcal{D})\right] \\
&= -\int \sum_{m \in M} p(\Delta|m,\mathcal{D})p(m|\mathcal{D}) \log p(\Delta|m',\mathcal{D})\, d\Delta
\end{aligned}$$

The difference between the two approaches is given by

$$\begin{aligned}
&\mathbb{E}\left[L\left(\Delta, p^M\right)\right] - \mathbb{E}\left[L\left(\Delta, p^{BMA}\right)\right] \\
&= \int \sum_{m\in M} p(\Delta|m,\mathcal{D})p(m|\mathcal{D}) \left[-\log p(\Delta|m',\mathcal{D}) + \log \sum_{m\in M} p(\Delta|m,\mathcal{D})p(m|\mathcal{D})\right] d\Delta \\
&= \int p^{BMA}(\Delta) \log \frac{p^{BMA}(\Delta)}{p^M(\Delta)}\, d\Delta \\
&= \mathbb{KL}\left(p^{BMA}||p^M\right) \geq 0
\end{aligned}$$

Hence we have

$$\mathbb{E}\left[L\left(\Delta, p^M\right)\right] \geq \mathbb{E}\left[L\left(\Delta, p^{BMA}\right)\right]$$

Exercise 5.8

a. $p(x,y|\theta) = p(x|\theta)p(y|x,\theta) = p(x|\theta_1)p(y|x,\theta_2)$ Given that

$$p(x|\theta_1) = \begin{cases} 1-\theta_1 & x=0 \\ \theta_1 & x=1 \end{cases}$$

and $p(y|x,\theta_2)$ is given by

	$y=0$	$y=1$
$x=0$	θ_2	$1-\theta_2$
$x=1$	$1-\theta_2$	θ_2

Thus $p(x,y|\theta)$ is

	$y=0$	$y=1$
$x=0$	$(1-\theta_1)\theta_2$	$(1-\theta_1)(1-\theta_2)$
$x=1$	$\theta_1(1-\theta_2)$	$\theta_1\theta_2$

b. In the dataset, (0,0) appears 2 times, (0,1) appears 1 times, (1,0) appears 2 times and (1,1) appears 2 times, i.e.,

	$y=0$	$y=1$
$x=0$	2	1
$x=1$	2	2

thus

$$\begin{aligned}
p(\mathcal{D}|\theta_1,\theta_2) &= [(1-\theta_1)\theta_2]^2 \times [(1-\theta_1)(1-\theta_2)] \times [\theta_1(1-\theta_2)]^2 \times [\theta_1\theta_2]^2 \\
&= (1-\theta_1)^3\theta_1^4 \times (1-\theta_2)^3\theta_2^4 \\
\log p(\mathcal{D}|\theta_1,\theta_2) &= 3\log(1-\theta_1) + 4\log\theta_1 + 3\log(1-\theta_2) + 4\log\theta_2
\end{aligned}$$

Take the partial derivatives and set to 0, we get

$$\begin{aligned}
\frac{\partial \log p(\mathcal{D}|\theta_1,\theta_2)}{\partial\theta_1} &= -\frac{3}{1-\theta_1} + \frac{4}{\theta_1} = 0 \\
\frac{\partial \log p(\mathcal{D}|\theta_1,\theta_2)}{\partial\theta_2} &= -\frac{3}{1-\theta_2} + \frac{4}{\theta_2} = 0
\end{aligned}$$

We have $\theta_1 = \theta_2 = \frac{4}{7}$. Hence

$$p(\mathcal{D}|\hat{\theta}, M_2) = \left(1 - \frac{4}{7}\right)^3 \left(\frac{4}{7}\right)^4 \left(1 - \frac{4}{7}\right)^3 \left(\frac{4}{7}\right)^4 = \left(\frac{4}{7}\right)^8 \left(\frac{3}{7}\right)^6$$

c. In the model with 4 parameters, we have

$$\hat{\theta}^{ML} = \underset{\theta}{\operatorname{argmax}}\, p(\mathcal{D}|\theta) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\ \theta_{11}^2 \cdot \theta_{10}^2 \cdot \theta_{00}^2 \cdot \theta_{01}$$

We can take the derivatives and so on, but in this case, this is essentially the multinomial distribution, and so the ML estimate corresponds to fraction of empirical count in each category, i.e., $\hat{\theta}_{11} = \hat{\theta}_{10} = \hat{\theta}_{00} = \frac{2}{7}$ and $\hat{\theta}_{01} = 1/7$. Therefore

$$p(\mathcal{D}|\hat{\theta}, M_4) = \left(\frac{2}{7}\right)^2 \cdot \left(\frac{2}{7}\right)^2 \cdot \left(\frac{2}{7}\right)^2 \cdot \frac{1}{7} = \left(\frac{2}{7}\right)^6 \frac{1}{7}$$

d. For leave-one-out cross-validation, when $x = 0, y = 1$ is left out, model M_4 will assign 0 probability to $x = 0, y = 1$, and so $L(M_4) = -\infty$. On the other hand, $L(M_2)$ is finite, so CV will pick M_2.

e. For M_2 model, we have

$$\text{BIC}(M_2, \mathcal{D}) = \log p(\mathcal{D}|\hat{\theta}_{MLE}) - \frac{\text{dof}(M_2)}{2} \log N = \log \left(\frac{4}{7}\right)^8 \left(\frac{3}{7}\right)^6 - \frac{2}{2} \log 7 \approx -11.51$$

For M_4 model, we have

$$\text{BIC}(M_4, \mathcal{D}) = \log p(\mathcal{D}|\hat{\theta}_{MLE}) - \frac{\text{dof}(M_4)}{2} \log N = \log \left(\frac{2}{7}\right)^6 \frac{1}{7} - \frac{3}{2} \log 7 \approx -12.38$$

Since $\text{BIC}(M_2, \mathcal{D}) > \text{BIC}(M_4, \mathcal{D})$. BIC will also prefer M_2 model.

Exercise 5.9

L1 loss is defined as:

$$L(y, a) = |y - a|$$

The posterior expected loss is given by

$$\rho(a|x) = \mathbb{E}[L(y, a)|x] = \sum_y |y - a| p(y|x) = \sum_{y \geq a} (y - a) p(y|x) + \sum_{y < a} (a - y) p(y|x)$$

Differentiating with respect to a, we get

$$\frac{\partial \rho(a|x)}{\partial a} = -\sum_{y \geq a} p(y|x) + \sum_{y < a} p(y|x) = -P(y \geq a|x) + P(y < a|x) = 0$$
$$\implies P(y < a|x) = P(y \geq a|x) = 0.5$$

Thus posterior median minimizes L1 loss. This assumes y is discrete. The derivation is similar for when y is continuous by replacing sum with integral, except differentiation is slightly more tricky since we have to differentiate inside the integral (see https://en.wikipedia.org/wiki/Differentiation_-under_the_integral_sign).

Q6: Using an imperfect oracle

- (i) Here are the expected loss

  $$\begin{aligned} R(\alpha_0) &= 0.4 * 6 = 2.4 \\ R(\alpha_1) &= 0.6 * 6 = 3.6 \\ R(\alpha_h) &= 2 \end{aligned}$$

  Therefore we should choose α_h

- (ii) Here are the expected loss

  $$\begin{aligned} R(\alpha_0) &= 0.1 * 6 = 0.6 \\ R(\alpha_1) &= 0.9 * 6 = 5.4 \\ R(\alpha_h) &= 2 \end{aligned}$$

  Therefore we should choose α_0

- (iii) The general expected loss is given by

  $$\begin{aligned} R(\alpha_0) &= 6p_1 \\ R(\alpha_1) &= 6(1 - p_1) \\ R(\alpha_h) &= 2 \end{aligned}$$

  We should choose α_0 if $R(\alpha_0) < R(\alpha_1) \implies p_1 < 0.5$ **and** $R(\alpha_0) < R(\alpha_h) \implies p_1 < 1/3$

  We should choose α_1 if $R(\alpha_1) < R(\alpha_0) \implies p_1 > 0.5$ and $R(\alpha_1) < R(\alpha_h) \implies p_1 > 2/3$

  Therefore we should choose α_0 if $p_1 < 1/3$, α_1 if $p_1 > 2/3$ and α_h otherwise.

- (iv) If the human is correct 95% of the time, then the general expected cost of α_h is $0.95 \times 2 + 0.05 \times 8 = 2.3$

  Therefore, we should choose α_0 if $p_1 < 0.5$ and $R(\alpha_0) < R(\alpha_h) \implies p_1 < 2.3/6 = 0.383$

  And we should choose α_1 if $p_1 > 0.5$ and $R(\alpha_1) < R(\alpha_h) \implies 6(1 - p_1) < 2.3 = 1 - 0.38 = 0.617$

  Therefore we should choose α_0 if $p_1 < 0.383$, α_1 if $p_1 > 0.617$ and α_h otherwise.