Problem 1. ([Al] Ex. 3.2-3.3)

We discussed the discriminant functions $g_i(x), i \in [K]$ where K is the number of classes. When $K = 2$ we can also define a single discriminant

$$g(x) = g_1(x) - g_2(x)$$

and we choose C_1 if $g(x) > 0$ and C_2 if $g(x) < 0$.

1. In a two-class problem, the likelihood ratio is

$$\frac{p(x \mid C_1)}{p(x \mid C_2)}$$

Write a discriminant function in terms of the likelihood ratio.

Solution. We could introduce a discriminant function as

$$g(x) = \frac{P(C_1 \mid x)}{P(C_2 \mid x)} = \frac{P(x \mid C_1)}{P(x \mid C_2)} \frac{P(C_1)}{P(C_2)}$$

Then, we choose C_1 if $g(x) \geq 1$ and C_2 if $g(x) < 1$. ■

2. In a two-class problem, the log odds is defined as

$$\log \frac{P(C_1 \mid x)}{P(C_2 \mid x)}$$

Write a discriminant function in terms of the log odds.

Solution. We could introduce a similar discriminant function as

$$g(x) = \log \frac{P(C_1 \mid x)}{P(C_2 \mid x)} = \log \frac{P(x \mid C_1)}{P(x \mid C_2)} + \log \frac{P(C_1)}{P(C_2)}$$

Then, we choose C_1 if $g(x) \geq 0$ and C_2 if $g(x) < 0$. ■

Problem 2. ([AI] Ex. 3.4)

In a two-class, two-action problem, if the loss function is $\lambda_{11} = \lambda_{22} = 0, \lambda_{12} = 10$ and $\lambda_{21} = 5$, write the optimal decision rule. How does the rule change if we add a third action of reject with $\lambda = 1$? [Note: we don't have 0/1 loss for this problem.]

Solution. The expected risks are

$$R(\alpha_1 \mid x) = \lambda_{11} P(C_1 \mid x) + \lambda_{12} P(C_2 \mid x) = 10P(C_2 \mid x)$$
$$R(\alpha_2 \mid x) = \lambda_{21} P(C_1 \mid x) + \lambda_{22} P(C_2 \mid x) = 5P(C_1 \mid x)$$

We choose C_1 if $R(\alpha_1 \mid x) < R(\alpha_2 \mid x)$, or $10P(C_2 \mid x) < 5P(C_1 \mid x)$, $P(C_1 \mid x) > \frac{2}{3}$, choose C_2 if $P(C_1 \mid x) \leq \frac{2}{3}$. The risk of reject is

$$R(\alpha_3 \mid x) = \lambda P(C_1 \mid x) + \lambda P(C_2 \mid x) = \lambda = 1$$

Then, we choose C_1 if

$$\begin{cases} R(\alpha_1 \mid x) < R(\alpha_2 \mid x) \\ R(\alpha_1 \mid x) < R(\alpha_3 \mid x) \end{cases} \Rightarrow \begin{cases} 10P(C_2 \mid x) < 5P(C_1 \mid x) \\ 10P(C_2 \mid x) < 1 \end{cases} \Rightarrow \begin{cases} P(C_1 \mid x) > \frac{2}{3} \\ P(C_1 \mid x) > \frac{9}{10} \end{cases} \Rightarrow P(C_1 \mid x) > \frac{9}{10}$$

We choose C_2 if

$$\begin{cases} R(\alpha_2 \mid x) < R(\alpha_1 \mid x) \\ R(\alpha_2 \mid x) < R(\alpha_3 \mid x) \end{cases} \Rightarrow \begin{cases} 5P(C_1 \mid x) < 10P(C_2 \mid x) \\ 5P(C_1 \mid x) < 1 \end{cases} \Rightarrow \begin{cases} P(C_1 \mid x) < \frac{2}{3} \\ P(C_1 \mid x) < \frac{1}{5} \end{cases} \Rightarrow P(C_1 \mid x) < \frac{1}{5}$$

We reject if

$$\begin{cases} R(\alpha_3 \mid x) \leq R(\alpha_1 \mid x) \\ R(\alpha_3 \mid x) \leq R(\alpha_2 \mid x) \end{cases} \Rightarrow \begin{cases} 1 \leq 10P(C_2 \mid x) \\ 1 \leq 5P(C_1 \mid x) \end{cases} \Rightarrow \begin{cases} P(C_1 \mid x) \leq \frac{9}{10} \\ P(C_1 \mid x) \geq \frac{1}{5} \end{cases} \Rightarrow \frac{1}{5} \leq P(C_1 \mid x) \leq \frac{9}{10}$$

■

Problem 3. (Poisson MLE)

Let X be a random variable. $X \sim$ Poisson (λ) with the density

$$p(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

1. Find $\mathbb{E}[X]$ and $\mathrm{Var}(X)$ if $X \sim$ Poisson (λ).

Solution.

$$\begin{aligned} \mathbb{E}[X] &= \sum_{x \in \mathrm{Img(X)}} x\mathbb{P}(X = x) \\ &= \sum_{x=0}^{\infty} x \frac{\lambda^x e^{-\lambda}}{x!} \\ &= e^{-\lambda} \sum_{x=1}^{\infty} x \frac{\lambda^x}{x!} \\ &= e^{-\lambda} \lambda \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} \\ &= e^{-\lambda} \lambda \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} \\ &= e^{-\lambda} \lambda e^{\lambda} \\ &= \lambda \end{aligned}$$

$$\begin{aligned} \mathbb{E}[X^2] &= \sum_{x \in \mathrm{Img(X)}} x^2 \mathbb{P}(X = x) \\ &= \sum_{x=0}^{\infty} x^2 \frac{\lambda^x e^{-\lambda}}{x!} = e^{-\lambda} \sum_{x=1}^{\infty} x^2 \frac{\lambda^x}{x!} \\ &= e^{-\lambda} \lambda \sum_{x=1}^{\infty} (x - 1 + 1) \frac{\lambda^{x-1}}{(x-1)!} \\ &= e^{-\lambda} \lambda \left(\sum_{x=1}^{\infty} (x-1) \frac{\lambda^{x-1}}{(x-1)!} + \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} \right) \\ &= e^{-\lambda} \lambda \left(\lambda \sum_{x=2}^{\infty} \frac{\lambda^{x-2}}{(x-2)!} + \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} \right) \\ &= e^{-\lambda} \lambda \left(\lambda \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} + \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} \right) \\ &= e^{-\lambda} \lambda (\lambda e^{\lambda} + e^{\lambda}) \\ &= \lambda^2 + \lambda \\ \mathrm{Var}(X) &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \\ &= \lambda^2 + \lambda - \lambda^2 = \lambda \end{aligned}$$

■

2. Consider the sample $\mathcal{X} = \{x_n\}_{n=1}^N$ where $x_n \sim^{i.i.d.}$ Poisson(λ). For the parameter λ above, write the likelihood $l(\lambda \mid \mathcal{X})$ and the log-likelihood $\mathcal{L}(\lambda \mid \mathcal{X})$.

Solution. Since they are i.i.d. samples,

$$l(\lambda \mid \mathcal{X}) = \prod_{n=1}^{N} \frac{\lambda^{x_n} e^{-\lambda}}{x_n!}$$

By taking the logarithm of the likelihood,

$$\begin{aligned}\mathcal{L}(\lambda \mid \mathcal{X}) &= \log\left(\prod_{n=1}^{N} \frac{\lambda^{x_n} e^{-\lambda}}{x_n!}\right) \\ &= \sum_{n=1}^{N} \log\left(\frac{\lambda^{x_n} e^{-\lambda}}{x_n!}\right) \\ &= \sum_{n=1}^{N} \log(\lambda^{x_n}) + \log(e^{-\lambda}) - \log(x_n!) \\ &= -n\lambda + \log(\lambda)\sum_{n=1}^{N} x_n - \sum_{n=1}^{N}\log(x_n!)\end{aligned}$$

■

3. Find the maximum likelihood estimator $\hat{\lambda}_{\text{MLE}}$.

Solution. To maximize $\mathcal{L}(\lambda \mid \mathcal{X})$, we need to solve

$$\hat{\lambda}_{\text{MLE}} = \underset{\lambda}{\operatorname{argmax}} = \underbrace{-n\lambda + \log(\lambda)\sum_{n=1}^{N} x_n - \sum_{n=1}^{N}\log(x_n!)}_{f(\lambda)}$$

By the first order condition of the maximum,

$$\frac{\mathrm{d}f}{\mathrm{d}\lambda} = -n + \frac{1}{\lambda}\sum_{n=1}^{N} x_n = 0 \quad \Rightarrow \quad \hat{\lambda}_{\text{MLE}} = \frac{1}{n}\sum_{n=1}^{N} x_n$$

■

4. Is $\hat{\lambda}_{\text{MLE}}$ biased?

Solution. The bias of the estimator is

$$d_\lambda(\hat{\lambda}_{\text{MLE}}) = \mathbb{E}[\hat{\lambda}_{\text{MLE}}] - \lambda = \mathbb{E}\left[\frac{1}{n}\sum_{n=1}^{N} x_n\right] - \lambda = \frac{1}{n}\sum_{n=1}^{N}\mathbb{E}[x_n] - \lambda = \lambda - \lambda = 0$$

therefore it is unbiased. ■

Problem 4. (Uniform MLE) Let X be a random variable. $X \sim \text{Unif}(\theta)$ with the density

$$p(x) = \begin{cases} \frac{1}{\theta}, & \text{if } 0 \le \text{x} \le \theta \\ 0, & \text{otherwise.} \end{cases}$$

1. Find $\mathbb{E}[X]$ and $\text{Var}(X)$ if $X \sim \text{Unif}(\theta)$.

Solution.

$$\begin{aligned}\mathbb{E}[X] &= \int_0^\theta x\frac{1}{\theta}\,\mathrm{d}x \\ &= \frac{1}{\theta}\left.\frac{x^2}{2}\right|_{x=0}^{\theta} \\ &= \frac{\theta}{2}\end{aligned}$$

$$\begin{aligned}\mathbb{E}[X^2] &= \int_0^\theta x^2\frac{1}{\theta}\,\mathrm{d}x \\ &= \frac{1}{\theta}\left.\frac{x^3}{3}\right|_{x=0}^{\theta} = \frac{\theta^2}{3} \\ \text{Var}(X) &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \\ &= \frac{\theta^2}{3} - \frac{\theta^2}{4} = \frac{\theta^2}{12}\end{aligned}$$

■

2. Consider the sample $\mathcal{X} = \{x_n\}_{n=1}^N$ where $x_n \sim^{i.i.d.} \mathrm{Unif}(\theta)$. For the parameter θ above, write the likelihood $l(\theta \mid \mathcal{X})$ and the log-likelihood $\mathcal{L}(\theta \mid \mathcal{X})$.

Solution. Suppose $I(\cdot)$ is the indicator function. The likelihood function is,

$$l(\theta \mid \mathcal{X}) = \prod_{n=1}^{N} p(x_n \mid \theta) = \frac{1}{\theta^N} I\left(\{x_n\}_{n=1}^N \in [0, \theta]\right) = \frac{1}{\theta^N} I\left(\max\{x_n\}_{n=1}^N \leq \theta\right)$$

By taking the logarithm of the likelihood,

$$\mathcal{L}(\theta \mid \mathcal{X}) = \log\left(\frac{1}{\theta^N} I\left(\max\{x_n\}_{n=1}^N \leq \theta\right)\right) = -N\log(\theta) + \log\left(I\left(\max\{x_n\}_{n=1}^N \leq \theta\right)\right)$$

■

3. Find the maximum likelihood estimator $\hat{\theta}_{\text{MLE}}$.

Solution. When $\theta < \max\{x_n\}_{n=1}^N$, $l(\theta \mid \mathcal{X}) = 0$. When $\theta \geq \max\{x_n\}_{n=1}^N$, $l(\theta \mid \mathcal{X}) = \frac{1}{\theta^N}$.

Since $\frac{1}{\theta^N}$ is monotonically decreasing, the maximum likelihood estimator is $\hat{\theta}_{\text{MLE}} = \max\{x_n\}_{n=1}^N$. ■

4. Is $\hat{\theta}_{\text{MLE}}$ biased?

Solution. In order to take the expectation of $\hat{\theta}_{\text{MLE}}$, we need to find its distribution. The CDF of the estimator is obvious,

$$P(\hat{\theta}_{\text{MLE}} \leq m) = P(\max\{x_n\}_{n=1}^N \leq m) = P(x_1 \leq m, x_2 \leq m, \dots, x_N \leq m) = \underbrace{\left(\frac{m}{\theta}\right)^N}_{F(m)}$$

Then, we could get the PDF by,

$$f(m) = \frac{\mathrm{d}F(m)}{\mathrm{d}m} = \frac{1}{\theta} N \left(\frac{m}{\theta}\right)^{N-1}$$

The bias of the estimator is,

$$d_\theta(\hat{\theta}_{\text{MLE}}) = \mathbb{E}[\hat{\theta}_{\text{MLE}}] - \theta = \int_0^\theta m \frac{1}{\theta} N \left(\frac{m}{\theta}\right)^{N-1} \mathrm{d}m - \theta = \frac{N}{\theta^N} \left.\frac{m^{N+1}}{N+1}\right|_{m=0}^{\theta} - \theta = \frac{N}{N+1}\theta - \theta = -\frac{\theta}{N+1}$$

therefore it is biased. ■

Problem 5. (See [Al] Ch.16.2.2) Find $\hat{q}_{\text{MAP}}$ for the Bernoulli likelihood

$$p(\mathcal{X} \mid q) = \prod_{n=1}^{N} q^{x_n}(1-q)^{1-x_n}$$

with the beta prior

$$p(q) = beta(q \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} q^{\alpha-1}(1-q)^{\beta-1}$$

Solution.

$$\begin{aligned}
\hat{q}_{\text{MAP}} &= \operatorname*{argmax}_q \mathbb{P}(q \mid \mathcal{X}) = \operatorname*{argmax}_q \log \mathbb{P}(q \mid \mathcal{X}) = \operatorname*{argmax}_q \log \frac{\mathbb{P}(\mathcal{X} \mid q)\mathbb{P}(q)}{\mathbb{P}(\mathcal{X})} = \operatorname*{argmax}_q \log \mathbb{P}(\mathcal{X} \mid q)\mathbb{P}(q) \\
&= \operatorname*{argmax}_q \log \prod_{n=1}^{N} \mathbb{P}(x_n \mid q)\mathbb{P}(q) = \operatorname*{argmax}_q \sum_{n=1}^{N} \log \mathbb{P}(x_n \mid q) + \log \mathbb{P}(q) \\
&= \operatorname*{argmax}_q \underbrace{\sum_{n=1}^{N} x_n \log q + (1-x_n)\log(1-q) + \log \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} + (\alpha-1)\log q + (\beta-1)\log(1-q)}_{\mathcal{L}}
\end{aligned}$$

By the first order condition of the maximum,

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial q} &= \sum_{n=1}^{N} \frac{\partial}{\partial q} x_n \log q + \frac{\partial}{\partial q}(1-x_n)\log(1-q) + \frac{\partial}{\partial q} \log \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} + \frac{\partial}{\partial q}(\alpha-1)\log q + \frac{\partial}{\partial q}(\beta-1)\log(1-q) \\
&= \frac{1}{q}\sum_{n=1}^{N} x_n - \frac{1}{1-q}\sum_{n=1}^{N}(1-x_n) + 0 + \frac{\alpha-1}{q} - \frac{\beta-1}{1-q}
\end{aligned}$$

Let $\frac{\partial \mathcal{L}}{\partial q} = 0$ and we have

$$\begin{aligned}
\frac{1}{q}\sum_{n=1}^{N} x_n - \frac{1}{1-q}\sum_{n=1}^{N}(1-x_n) + \frac{\alpha-1}{q} - \frac{\beta-1}{1-q} &= 0 \\
q\left(\sum_{n=1}^{N}(1-x_n) + \beta - 1\right) &= (1-q)\left(\sum_{n=1}^{N} x_n + \alpha - 1\right) \\
q\left(\sum_{n=1}^{N}(1-x_n) + \sum_{n=1}^{N} x_n + \beta - 1 + \alpha - 1\right) &= \sum_{n=1}^{N} x_n + \alpha - 1 \\
q\,(N + \beta + \alpha - 2) &= \sum_{n=1}^{N} x_n + \alpha - 1 \\
q &= \frac{\sum_{n=1}^{N} x_n + \alpha - 1}{N + \beta + \alpha - 2}
\end{aligned}$$

We have

$$\hat{q}_{\text{MAP}} = \frac{\sum_{n=1}^{N} x_n + \alpha - 1}{N + \beta + \alpha - 2}$$

■

Problem 6. (Exponential family) A probability distribution in the exponential family is given by

$$p(\boldsymbol{x} \mid \boldsymbol{\eta}) = h(\boldsymbol{x}) \exp\left(\boldsymbol{\eta}^\top T(\boldsymbol{x}) - A(\boldsymbol{\eta})\right)$$

where $\boldsymbol{\eta}$ is the parameter vector.

1. Prove that $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{I})$ with identity covariance (where $\boldsymbol{\mu}$ is the parameter) is in the exponential family.

   Solution. Suppose $\boldsymbol{x} \in \mathbb{R}^d$. For $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{I})$, we have,

$$\begin{aligned}
p(\boldsymbol{x} \mid \boldsymbol{\mu}) &= (2\pi)^{\frac{-d}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top(\boldsymbol{x}-\boldsymbol{\mu})\right) \\
&= (2\pi)^{\frac{-d}{2}} \exp\left(\boldsymbol{\mu}^\top \boldsymbol{x} - \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\mu}\right) \\
&= (2\pi)^{\frac{-d}{2}} \exp\left(\langle \boldsymbol{\mu}, \boldsymbol{x} \rangle + \left\langle \operatorname{vec}\left(-\frac{1}{2}\boldsymbol{I}\right), \operatorname{vec}\left(\boldsymbol{x}\boldsymbol{x}^\top\right)\right\rangle - \frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\mu}\right)
\end{aligned}$$

Then, we write $h(\boldsymbol{x}) = (2\pi)^{\frac{-d}{2}}$, $T(\boldsymbol{x}) = \begin{bmatrix} \boldsymbol{x} \\ \text{vec}\left(\boldsymbol{x}\boldsymbol{x}^\top\right) \end{bmatrix}$, $\boldsymbol{\eta} = \begin{bmatrix} \boldsymbol{\mu} \\ \text{vec}\left(-\frac{1}{2}\boldsymbol{I}\right) \end{bmatrix}$, $A(\boldsymbol{\eta}) = \frac{1}{2}\boldsymbol{\mu}^\top\boldsymbol{\mu} = \frac{1}{2}(\boldsymbol{\eta}^\top\boldsymbol{\eta} - \frac{d}{4})$ ■

2. Prove that

$$\nabla_{\boldsymbol{\eta}} A = \mathbb{E}_{\mathbf{x}\sim p(\boldsymbol{x}|\boldsymbol{\eta})}[T(\mathbf{x})].$$

Hint: Use the fact that $\int p(\boldsymbol{x} \mid \boldsymbol{\eta}) d\boldsymbol{x} = 1$ to get an expression of A first.

Solution. As the Hint shows,

$$\begin{aligned} \int h(\boldsymbol{x}) \exp\left(\boldsymbol{\eta}^\top T(\boldsymbol{x}) - A(\boldsymbol{\eta})\right) \mathrm{d}\boldsymbol{x} = 1 \\ \exp(-A(\boldsymbol{\eta})) \int h(\boldsymbol{x}) \exp\left(\boldsymbol{\eta}^\top T(\boldsymbol{x})\right) \mathrm{d}\boldsymbol{x} = 1 \\ A(\boldsymbol{\eta}) = \log \int h(\boldsymbol{x}) \exp\left(\boldsymbol{\eta}^\top T(\boldsymbol{x})\right) \mathrm{d}\boldsymbol{x} \end{aligned}$$

Then, we take the derivative,

$$\begin{aligned} \nabla_{\boldsymbol{\eta}} A &= \frac{\partial}{\partial \boldsymbol{\eta}^\top}\left(\log \int h(\boldsymbol{x}) \exp\left(\boldsymbol{\eta}^\top T(\boldsymbol{x})\right) \mathrm{d}\boldsymbol{x}\right) \\ &= \frac{\int T(\boldsymbol{x}) h(\boldsymbol{x}) \exp\left(\boldsymbol{\eta}^\top T(\boldsymbol{x})\right) \mathrm{d}\boldsymbol{x}}{\int h(\boldsymbol{x}) \exp\left(\boldsymbol{\eta}^\top T(\boldsymbol{x})\right) \mathrm{d}\boldsymbol{x}} \\ &= \int T(\boldsymbol{x}) h(\boldsymbol{x}) \exp\left(\boldsymbol{\eta}^\top T(\boldsymbol{x}) - A(\boldsymbol{\eta})\right) \mathrm{d}\boldsymbol{x} \\ &= \mathbb{E}_{\mathbf{x}\sim p(\boldsymbol{x}|\boldsymbol{\eta})}[T(\mathbf{x})] \end{aligned}$$

■

3. Verify Part 2 using the example in Part 1.

Solution. We first consider

$$(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top])_{ij} = \mathbb{E}[(\boldsymbol{x}\boldsymbol{x}^\top)_{ij}] = \mathbb{E}[x_i x_j] = \text{cov}(x_i, x_j) + \mathbb{E}[x_i]\mathbb{E}[x_j] = \Sigma_{ij} + \mu_i\mu_j$$

$$\mathbb{E}\left[\text{vec}(\boldsymbol{x}\boldsymbol{x}^\top)\right] = \mathbb{E}\left[\begin{pmatrix} x_1^2 \\ \vdots \\ x_1 x_d \\ \vdots \\ x_d^2 \end{pmatrix}\right] = \begin{pmatrix} \mathbb{E}\left[x_1^2\right] \\ \vdots \\ \mathbb{E}\left[x_1 x_d\right] \\ \vdots \\ \mathbb{E}\left[x_d^2\right] \end{pmatrix} = \begin{pmatrix} \Sigma_{11} + \mu_1^2 \\ \vdots \\ \Sigma_{1d} + \mu_1\mu_d \\ \vdots \\ \Sigma_{dd} + \mu_d^2 \end{pmatrix} = \text{vec}\left(\mathbb{E}\left[\boldsymbol{x}\boldsymbol{x}^\top\right]\right)$$

Therefore,

$$\mathbb{E}[T(\boldsymbol{x})] = \mathbb{E}\begin{bmatrix} \boldsymbol{x} \\ \text{vec}\left(\boldsymbol{x}\boldsymbol{x}^\top\right) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu} \\ \text{vec}\left(\boldsymbol{I} + \boldsymbol{\mu}\boldsymbol{\mu}^\top\right) \end{bmatrix}$$

We can verify that $\nabla_{\boldsymbol{\eta}} A = \mathbb{E}[T(\boldsymbol{x})]$ by direct differentiation. ■