Machine Learning Assignment 2

Bohan YANG, 2330016056

October 21, 2025

Question 1

Let $\mathcal{D} = \{X_1, \ldots, X_n\}$ be i.i.d. samples from a Poisson distribution with rate parameter λ, i.e. $X_i \sim \text{Poisson}(\lambda)$.

Denote $S = \sum_{i=1}^n X_i$ and $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$.

(a)

The likelihood function is

$$L(\lambda) = \prod_{i=1}^n P(X_i \mid \lambda) = \prod_{i=1}^n \frac{\lambda^{X_i} e^{-\lambda}}{X_i!} = \frac{\lambda^S e^{-n\lambda}}{\prod_{i=1}^n X_i!}.$$

The log-likelihood is

$$\ell(\lambda) = S \log \lambda - n\lambda - \sum_{i=1}^n \log(X_i!).$$

Taking the derivative w.r.t. λ and setting it to zero gives

$$\frac{d\ell}{d\lambda} = \frac{S}{\lambda} - n = 0 \quad \Longrightarrow \quad \hat{\lambda}_{\text{MLE}} = \frac{S}{n} = \bar{X}.$$

Because $\mathbb{E}[X_i] = \lambda$, the estimator is unbiased:

$$\mathbb{E}[\hat{\lambda}_{\text{MLE}}] = \mathbb{E}[\bar{X}] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i] = \lambda.$$

(b)

Assume the prior $\lambda \sim \text{Gamma}(\alpha, \beta)$ (rate parameterization) with pdf

$$p(\lambda \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}, \qquad \alpha > 0,\ \beta > 0.$$

The posterior is proportional to the product of the prior and the likelihood:

$$p(\lambda \mid \mathcal{D}) \propto p(\lambda)L(\lambda) \propto \lambda^{\alpha-1}e^{-\beta\lambda}\,\lambda^{S}e^{-n\lambda} = \lambda^{(\alpha+S)-1}e^{-(\beta+n)\lambda}.$$

Hence,

$$\boxed{\lambda \mid \mathcal{D} \sim \mathrm{Gamma}(\alpha+S,\ \beta+n).}$$

(c)

From part (b), we know

$$\lambda \mid \mathcal{D} \sim \mathrm{Gamma}(\alpha', \beta'), \quad \text{where } \alpha' = \alpha + S,\ \beta' = \beta + n.$$

$$p(\lambda \mid \alpha', \beta') = \frac{(\beta')^{\alpha'}}{\Gamma(\alpha')}\lambda^{\alpha'-1}e^{-\beta'\lambda}, \qquad \lambda > 0.$$

To find the mode (i.e., the most probable value of λ), we maximize $p(\lambda \mid \alpha', \beta')$ with respect to λ.

$$\ell(\lambda) = \log p(\lambda \mid \alpha', \beta') = (\alpha'-1)\log\lambda - \beta'\lambda + \text{constant}.$$

$$\frac{d\ell}{d\lambda} = \frac{\alpha'-1}{\lambda} - \beta'.$$

Setting the derivative to zero gives

$$\frac{\alpha'-1}{\lambda} - \beta' = 0 \quad \Longrightarrow \quad \lambda_{\text{mode}} = \frac{\alpha'-1}{\beta'}.$$

This stationary point is only valid if $\alpha' > 1$; otherwise, the pdf is monotonically decreasing and the mode occurs at the boundary $\lambda = 0$.

Therefore, the mode of the posterior distribution (which is the MAP estimator) is

$$\boxed{\hat{\lambda}_{\text{MAP}} = \begin{cases} \dfrac{\alpha'-1}{\beta'}, & \alpha' > 1, \\ 0, & \alpha' \le 1. \end{cases}}$$

Since $\alpha' = \alpha + S$ and $\beta' = \beta + n$, we obtain

$$\boxed{\hat{\lambda}_{\text{MAP}} = \begin{cases} \dfrac{\alpha+S-1}{\beta+n}, & \alpha+S > 1, \\ 0, & \alpha+S \le 1. \end{cases}}$$

Equivalently, using the sample mean $\bar{X} = \dfrac{S}{n}$,

$$\hat{\lambda}_{\text{MAP}} = \frac{n\bar{X} + \alpha - 1}{\beta + n} \qquad (\alpha + S > 1).$$

Question 2

$$y_i \sim \mathcal{N}(\mu, 1), \quad i = 1, 2, \dots, n.$$

(a)

Since the estimator is constant,

$$\mathbb{E}[\hat{\mu}] = 1.$$

Hence,

$$\text{Bias}(\hat{\mu}) = \mathbb{E}[\hat{\mu}] - \mu = 1 - \mu, \qquad \text{Var}(\hat{\mu}) = 0.$$

Therefore, the mean squared error (MSE) is

$$\text{MSE} = (1 - \mu)^2.$$

Interpretation: This estimator ignores the data and is generally not good, except in the special case where the true mean $\mu = 1$.

(b)

Since $y_1 \sim \mathcal{N}(\mu, 1)$,

$$\mathbb{E}[\hat{\mu}] = \mu, \qquad \text{Var}(\hat{\mu}) = 1.$$

Thus,

$$\text{Bias}(\hat{\mu}) = 0, \qquad \text{MSE} = 1.$$

Interpretation: This estimator is unbiased but has high variance, because it uses only one observation instead of all n samples.

(c)

The estimator is defined as

$$\hat{\mu} = \arg\min_{\mu} \sum_{i=1}^{n} (y_i - \mu)^2 + \lambda\mu^2, \qquad \lambda > 0.$$

Taking the derivative and setting it to zero:

$$-2\sum_{i=1}^{n}(y_i - \mu) + 2\lambda\mu = 0 \quad \Rightarrow \quad (n + \lambda)\hat{\mu} = \sum_{i=1}^{n} y_i.$$

Hence,

$$\boxed{\hat{\mu} = \frac{n}{n+\lambda}\,\bar{y}}, \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Expectation:

$$\mathbb{E}[\hat{\mu}] = \frac{n}{n+\lambda}\,\mathbb{E}[\bar{y}] = \frac{n}{n+\lambda}\,\mu.$$

Bias:

$$\text{Bias}(\hat{\mu}) = \mathbb{E}[\hat{\mu}] - \mu = \left(\frac{n}{n+\lambda} - 1\right)\mu = -\frac{\lambda}{n+\lambda}\,\mu.$$

Variance: Since $\text{Var}(\bar{y}) = \frac{1}{n}$,

$$\text{Var}(\hat{\mu}) = \left(\frac{n}{n+\lambda}\right)^2 \text{Var}(\bar{y}) = \left(\frac{n}{n+\lambda}\right)^2 \frac{1}{n} = \frac{n}{(n+\lambda)^2}.$$

MSE:

$$\text{MSE}(\hat{\mu}) = \text{Bias}^2 + \text{Var} = \frac{\lambda^2\mu^2 + n}{(n+\lambda)^2}.$$

Interpretation: This estimator is biased toward 0, but its variance is smaller than that of the unbiased sample mean. It can achieve a lower mean squared error (MSE) when n is small or when we expect μ to be close to 0.

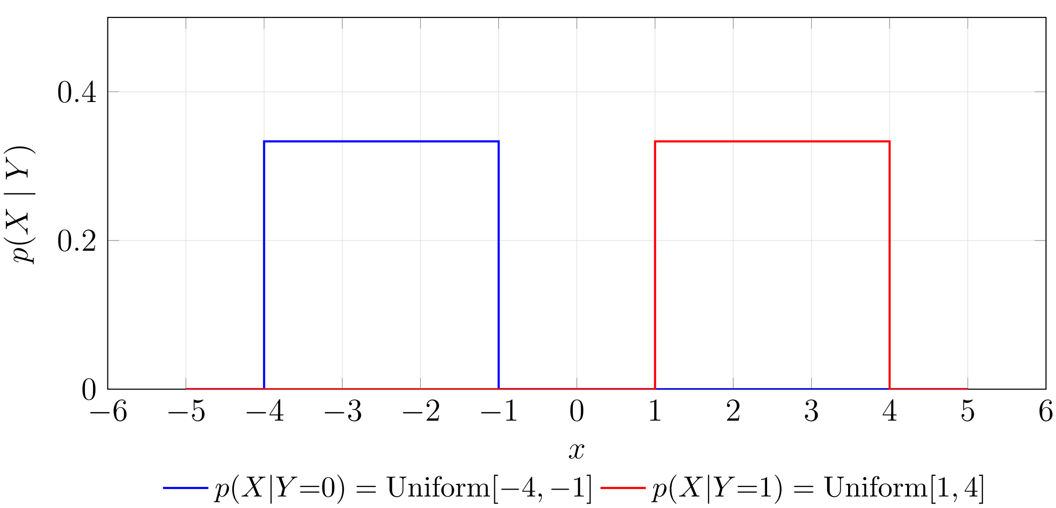
Question 3

(a)

$$p(X \mid Y=1) = \text{Uniform}[1,4], \quad p(X \mid Y=0) = \text{Uniform}[-4,-1].$$

Thus,

$$p(X \mid Y=1) = \begin{cases} \frac{1}{3}, & 1 \le X \le 4, \\ 0, & \text{otherwise,} \end{cases} \qquad p(X \mid Y=0) = \begin{cases} \frac{1}{3}, & -4 \le X \le -1, \\ 0, & \text{otherwise.} \end{cases}$$



(b)

Since the two distributions have disjoint supports and equal priors

$$P(Y=0) = P(Y=1) = \tfrac{1}{2},$$

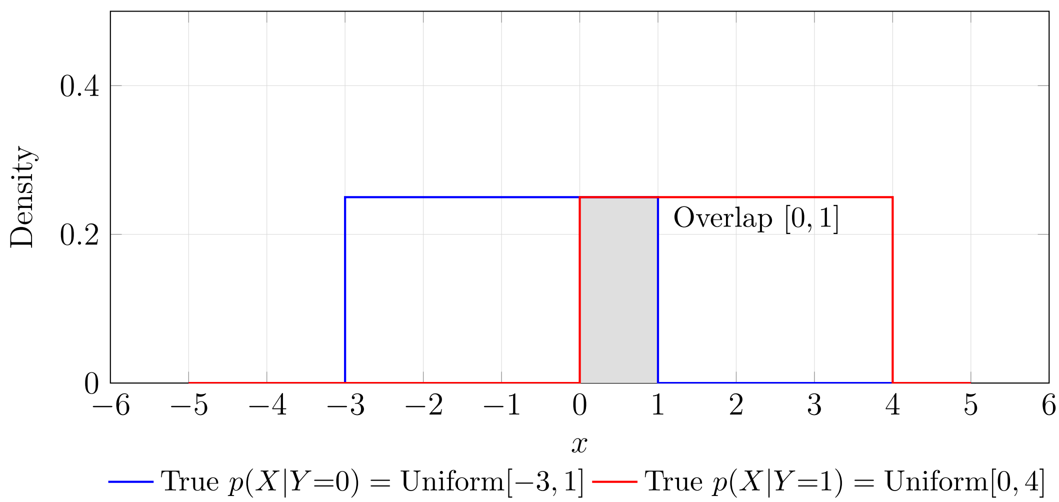
the Bayes classifier simply predicts the class whose support contains x. Therefore, the classifier makes no mistakes:

$$\boxed{\text{Bayes error} = 0.}$$

(c)

$$p(X \mid Y=1) = \text{Uniform}[0,4], \quad p(X \mid Y=0) = \text{Uniform}[-3,1].$$

$$\text{BayesErr} = \frac{1}{2}\int \min\{p(x \mid Y=1),\, p(x \mid Y=0)\}\, dx = \frac{1}{2} \times \frac{1}{4} \times 1 = \boxed{\tfrac{1}{8} = 0.125.}$$



(d)

For a uniform distribution Uniform$[a,b]$, we have:

$$\mu = \frac{a+b}{2}, \qquad \sigma^2 = \frac{(b-a)^2}{12}.$$

Hence,

$$p(X \mid Y=0) \approx \mathcal{N}(-2.5,\, 0.75), \qquad p(X \mid Y=1) \approx \mathcal{N}(2.5,\, 0.75).$$

With equal priors and identical variances, the decision boundary occurs at:

$$x^* = \frac{\mu_0 + \mu_1}{2} = 0.$$

Therefore, the classifier predicts $Y=1$ if $x > 0$, and $Y=0$ otherwise.

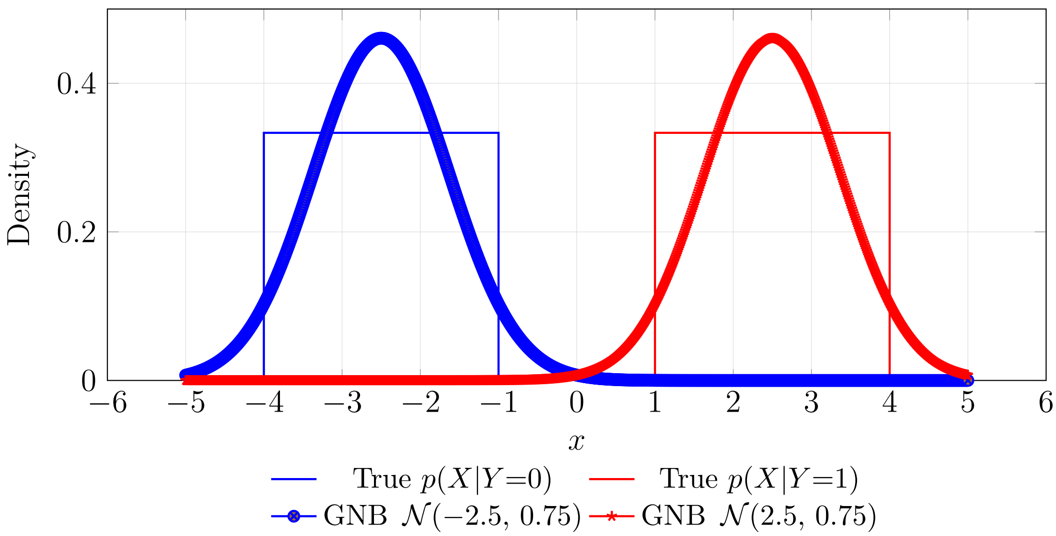
The classification error of GNB is:

$$\mathrm{Err}_{\mathrm{GNB}} = \frac{1}{2}\int_{-4}^{-1} \mathbf{1}\{x > x^*\}\,\frac{1}{3}\,dx + \frac{1}{2}\int_{1}^{4} \mathbf{1}\{x < x^*\}\,\frac{1}{3}\,dx.$$

Since the threshold $x^* = 0$ lies between the two non-overlapping intervals, both integrals are zero and thus:

$$\boxed{\mathrm{Err}_{\mathrm{GNB}} = 0.}$$

Although the GNB model is biased (it cannot represent the true uniform shape), the symmetry of the problem yields zero misclassification error.



(e)

With a finite n, the estimated class means and variances, $\hat{\mu}_k, \hat{\sigma}_k^2$, fluctuate due to sampling noise.

Consequently, the GNB decision rule (and its threshold) fluctuates as well.

This introduces a $\boxed{\textit{variance}}$ component of error on top of the (model) *bias*.

Question 4

Let $Y \in \{0,1\}$ with prior $\pi = P(Y=1)$, and $\mathbf{X} = (X_1,\dots,X_n)$ with conditional independence given Y (naïve Bayes). For each feature i and class $k \in \{0,1\}$, assume a univariate Gaussian:

$$P(X_i \mid Y=k) = \mathcal{N}(\mu_{i,k},\, \sigma_{i,k}^2), \qquad i = 1,\dots,n.$$

Hence

$$p(\mathbf{x} \mid Y=k) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma_{i,k}} \exp\left(-\frac{(x_i-\mu_{i,k})^2}{2\sigma_{i,k}^2}\right).$$

By Bayes' rule,

$$P(Y=1 \mid \mathbf{x}) = \frac{\pi\, p(\mathbf{x} \mid Y=1)}{\pi\, p(\mathbf{x} \mid Y=1) + (1-\pi)\, p(\mathbf{x} \mid Y=0)}.$$

log-odds:

$$\Lambda(\mathbf{x}) := \log \frac{P(Y=1 \mid \mathbf{x})}{P(Y=0 \mid \mathbf{x})} = \log \frac{\pi}{1-\pi} + \log \frac{p(\mathbf{x} \mid Y=1)}{p(\mathbf{x} \mid Y=0)}.$$

(a) $\sigma_{i,0} = \sigma_{i,1} = \sigma_i$

Under the special assumption used in class/readings, each feature has a class-independent standard deviation $\sigma_i > 0$. Then

$$\log \frac{p(\mathbf{x} \mid Y=1)}{p(\mathbf{x} \mid Y=0)} = \sum_{i=1}^{n} \left[-\frac{(x_i-\mu_{i,1})^2}{2\sigma_i^2} + \frac{(x_i-\mu_{i,0})^2}{2\sigma_i^2} \right] = \sum_{i=1}^{n} \frac{1}{2\sigma_i^2} \Big((x_i^2 - 2x_i\mu_{i,1} + \mu_{i,1}^2) - (x_i^2 - 2x_i\mu_{i,0} + \mu_{i,0}^2) \Big).$$

The x_i^2 terms cancel:

$$\log \frac{p(\mathbf{x} \mid Y=1)}{p(\mathbf{x} \mid Y=0)} = \sum_{i=1}^{n} \left(\frac{\mu_{i,1}-\mu_{i,0}}{\sigma_i^2} \right) x_i + \frac{1}{2} \sum_{i=1}^{n} \frac{\mu_{i,0}^2 - \mu_{i,1}^2}{\sigma_i^2}.$$

Therefore the log-odds is linear in $\mathbf{x}$:

$$\Lambda(\mathbf{x}) = w_0 + \sum_{i=1}^{n} w_i x_i, \quad \text{with} \quad w_i = \frac{\mu_{i,1}-\mu_{i,0}}{\sigma_i^2}, \quad w_0 = \log \frac{\pi}{1-\pi} + \frac{1}{2} \sum_{i=1}^{n} \frac{\mu_{i,0}^2 - \mu_{i,1}^2}{\sigma_i^2}.$$

Applying the logistic link,

$$P(Y=1 \mid \mathbf{x}) = \frac{1}{1+\exp\big(-\Lambda(\mathbf{x})\big)} = \sigma\left(w_0 + \sum_{i=1}^{n} w_i x_i \right),$$

which is *exactly* the functional form of logistic regression.

(b) $\sigma_{i,0} \neq \sigma_{i,1}$

$$\log \frac{p(\mathbf{x} \mid Y=1)}{p(\mathbf{x} \mid Y=0)} = \sum_{i=1}^{n} \left[\log \frac{\sigma_{i,0}}{\sigma_{i,1}} - \frac{(x_i-\mu_{i,1})^2}{2\sigma_{i,1}^2} + \frac{(x_i-\mu_{i,0})^2}{2\sigma_{i,0}^2} \right].$$

$$-\frac{(x_i-\mu_{i,1})^2}{2\sigma_{i,1}^2} + \frac{(x_i-\mu_{i,0})^2}{2\sigma_{i,0}^2} = \underbrace{\left(\frac{1}{2\sigma_{i,0}^2} - \frac{1}{2\sigma_{i,1}^2} \right)}_{\text{quadratic coeff.}} x_i^2 + \underbrace{\left(\frac{\mu_{i,1}}{\sigma_{i,1}^2} - \frac{\mu_{i,0}}{\sigma_{i,0}^2} \right)}_{\text{linear coeff.}} x_i + \underbrace{\left(\frac{\mu_{i,0}^2}{2\sigma_{i,0}^2} - \frac{\mu_{i,1}^2}{2\sigma_{i,1}^2} \right)}_{\text{constant}}.$$

$$\Lambda(\mathbf{x}) = \log \frac{\pi}{1-\pi} + \sum_{i=1}^{n} \log \frac{\sigma_{i,0}}{\sigma_{i,1}} + \sum_{i=1}^{n} \left(\frac{1}{2\sigma_{i,0}^2} - \frac{1}{2\sigma_{i,1}^2} \right) x_i^2 + \sum_{i=1}^{n} \left(\frac{\mu_{i,1}}{\sigma_{i,1}^2} - \frac{\mu_{i,0}}{\sigma_{i,0}^2} \right) x_i + \sum_{i=1}^{n} \left(\frac{\mu_{i,0}^2}{2\sigma_{i,0}^2} - \frac{\mu_{i,1}^2}{2\sigma_{i,1}^2} \right).$$

(c) conclusion

unless $\sigma_{i,0} = \sigma_{i,1}$ for every i, the coefficients of x_i^2 do not cancel, so the log-odds is *quadratic* in $\mathbf{x}$.

Hence the form is $\boxed{\textbf{not}}$ the logistic regression form.