Homework 2 – DSC 240
Winter 2026
Diana Freeman: A16574234

Question: 1

Let $X_1, X_2, \ldots, X_n$ be n independent and identically distributed (i.i.d.) random variables following a Bernoulli distribution with parameter p:

$$P(X = x; p) = p^x (1-p)^{1-x}, \quad x \in \{0, 1\}$$

To find the Maximum Likelihood Estimator (MLE) for p, we follow these steps:

1. The Likelihood Function

The likelihood function $L(p)$ is the joint probability of the observed data:

$$L(p) = \prod_{i=1}^{n} P(X_i = x_i; p) = \prod_{i=1}^{n} p^{x_i} (1-p)^{1-x_i}$$

2. The Log-Likelihood Function

To simplify the differentiation, we take the natural logarithm of the likelihood function:

$$\ell(p) = \ln L(p) = \ln \left(p^{\sum x_i} (1-p)^{n - \sum x_i} \right)$$

$$\ell(p) = \left(\sum_{i=1}^{n} x_i \right) \ln(p) + \left(n - \sum_{i=1}^{n} x_i \right) \ln(1-p)$$

3. Finding the Maximum

We take the first derivative of the log-likelihood with respect to p and set it to zero:

$$\frac{d}{dp} \ell(p) = \frac{\sum x_i}{p} - \frac{n - \sum x_i}{1-p} = 0$$

Multiplying through by $p(1-p)$ to clear the denominators:

$$(1-p) \sum x_i - p(n - \sum x_i) = 0$$

$$\sum x_i - p \sum x_i - np + p \sum x_i = 0$$

$$\sum x_i - np = 0$$

4. The MLE Estimator

Solving for p gives us the Maximum Likelihood Estimator, denoted as $\hat{p}$:

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

The estimator $\hat{p}$ is the sample mean of the observations.

Question: 2

(a)

Starting with the parameter $\hat{p}_i = \hat{p}(x_i; \theta)$ and $p(y|x)$ following a Bernoulli distribution.

For a single observation (x_i, y_i), the probability mass function is:

$$P(y_i|x_i;\theta) = \hat{p}_i^{y_i}(1-\hat{p}_i)^{1-y_i}$$

Assuming i.i.d. examples, the likelihood function $L(\theta)$ is

$$L(\theta) = \prod_{i=1}^{n} P(y_i|x_i;\theta) = \prod_{i=1}^{n} \hat{p}_i^{y_i}(1-\hat{p}_i)^{1-y_i}$$

We apply the natural logarithm to obtain the log-likelihood function, $\ell(\theta)$, which converts the product into a sum:

$$\ell(\theta) = \log L(\theta) = \log\left(\prod_{i=1}^{n} \hat{p}_i^{y_i}(1-\hat{p}_i)^{1-y_i}\right)$$

$$\ell(\theta) = \sum_{i=1}^{n} \left(y_i \log(\hat{p}_i) + (1-y_i)\log(1-\hat{p}_i)\right)$$

The principle of Maximum Likelihood Estimation (MLE) seeks to maximize $\ell(\theta)$. In machine learning, we typically frame problems as minimizing a loss function. Maximizing a function is equivalent to minimizing its negative. Therefore, we minimize the Negative Log-Likelihood (NLL). Taking the average over the n examples yields the Cross-Entropy (CE) loss:

$$\mathrm{CE}(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \left(y_i \log(\hat{p}_i) + (1-y_i)\log(1-\hat{p}_i)\right)$$

(b)

Instantiate the prediction as $\hat{p}_i = \sigma(\theta^\top x_i) = \frac{1}{1+e^{-\theta^\top x_i}}$. Let $z_i = \theta^\top x_i$.

First, let us express $\hat{p}_i$ and $1 - \hat{p}_i$ in terms of exponentials:

$$\hat{p}_i = \frac{1}{1 + e^{-z_i}}$$

$$1 - \hat{p}_i = 1 - \frac{1}{1 + e^{-z_i}} = \frac{e^{-z_i}}{1 + e^{-z_i}} = \frac{1}{1 + e^{z_i}}$$

Now, we evaluate the log terms from the Cross-Entropy loss function:

$$\log(\hat{p}_i) = -\log(1 + e^{-z_i})$$

$$\log(1 - \hat{p}_i) = -\log(1 + e^{z_i})$$

Let us examine the individual loss term for the i-th example, $l_i = -[y_i \log(\hat{p}_i) + (1 - y_i)\log(1 - \hat{p}_i)]$, considering the two possible cases for $y_i \in \{0, 1\}$ and mapping them to $\tilde{y}_i \in \{-1, 1\}$:

- **Case 1 ($y_i = 1 \implies \tilde{y}_i = 1$):**
  The loss simplifies to:

  $$l_i = -[1 \cdot \log(\hat{p}_i) + 0] = -\log(\hat{p}_i) = \log(1 + e^{-z_i})$$

  Since $\tilde{y}_i = 1$, we can rewrite this as $\log(1 + e^{-\tilde{y}_i z_i})$.

- **Case 2 ($y_i = 0 \implies \tilde{y}_i = -1$):**
  The loss simplifies to:

  $$l_i = -[0 + 1 \cdot \log(1 - \hat{p}_i)] = -\log(1 - \hat{p}_i) = \log(1 + e^{z_i})$$

  Since $\tilde{y}_i = -1$, note that $z_i = -(-\tilde{y}_i)z_i = -\tilde{y}_i z_i$. Thus, we can also rewrite this as $\log(1 + e^{-\tilde{y}_i z_i})$.

In both cases, the individual loss term takes the exact same form: $\log(1 + e^{-\tilde{y}_i z_i})$. Substituting $z_i = \theta^\top x_i$ and averaging over the n examples, we recover the logistic loss:

$$L(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + e^{-\tilde{y}_i \theta^\top x_i}\right)$$

Thus, minimizing the cross-entropy loss with a sigmoid activation is equivalent to minimizing the logistic loss.

Question 3

The Bayes optimal classifier is defined by the decision rule $h^*(x) = \arg\max_{y \in \{A,B\}} P(y|x)$. By Bayes' theorem, the posterior probability for class A is given by:

$$P(A|x) = \frac{p(x|A)P(A)}{p(x|A)P(A) + p(x|B)P(B)}$$

The decision boundary occurs where $P(A|x) = P(B|x)$, or equivalently, where the log-likelihood ratio plus the log-prior ratio equals zero:

$$\ln\left(\frac{p(x|A)}{p(x|B)}\right) + \ln\left(\frac{P(A)}{P(B)}\right) = 0$$

The multivariate Gaussian density is defined as:

$$p(x|\mu, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)$$

(a) Equal Priors and Identity Covariance

Parameters: $\mu_1 = [0,0]^\top, \mu_2 = [1,1]^\top, \Sigma_A = \Sigma_B = I_d, P(A) = P(B) = 0.5$.

1. Analytical Expression for $P(y|x)$: Since the covariance matrices are identity, $|\Sigma| = 1$ and $\Sigma^{-1} = I_d$.

$$P(A|x) = \frac{1}{1 + \exp\left(\frac{1}{2}\|x-\mu_1\|^2 - \frac{1}{2}\|x-\mu_2\|^2\right)}$$

2. Mathematical Work for Decision Rule:

$$-\frac{1}{2}(x-\mu_1)^\top(x-\mu_1) + \frac{1}{2}(x-\mu_2)^\top(x-\mu_2) + \ln(1) = 0$$

$$-\frac{1}{2}(x^\top x - 2\mu_1^\top x + \mu_1^\top \mu_1) + \frac{1}{2}(x^\top x - 2\mu_2^\top x + \mu_2^\top \mu_2) = 0$$

$$(\mu_1 - \mu_2)^\top x + \frac{1}{2}(\|\mu_2\|^2 - \|\mu_1\|^2) = 0$$

Substituting $\mu_1 = [0,0]^\top$ and $\mu_2 = [1,1]^\top$:

$$([0,0] - [1,1])\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \frac{1}{2}(2-0) = 0 \implies -x_1 - x_2 + 1 = 0$$

Decision Boundary: $x_1 + x_2 = 1$. The boundary is **linear**.

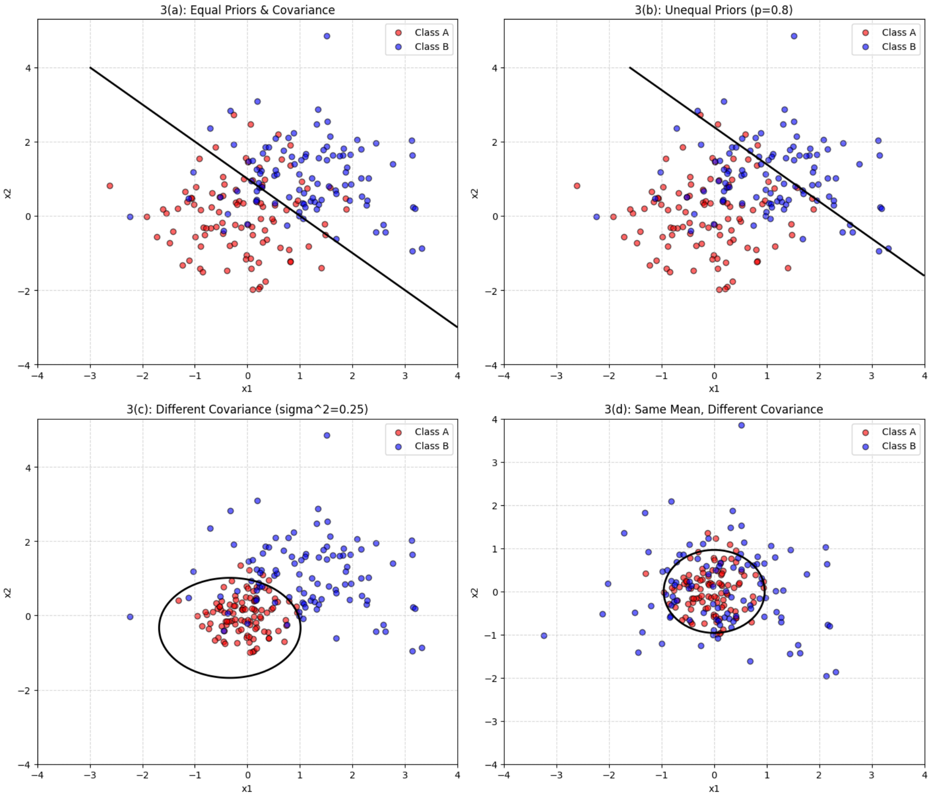

Figure 1: Decision Boundary Plots

(b) Unequal Prior Probabilities

Parameters: $\mu_1 = [0,0]^\top, \mu_2 = [1,1]^\top, \Sigma = I_d, P(A) = 0.8, P(B) = 0.2.$

1. Analytical Expression for $P(y|x)$:

$$P(A|x) = \frac{0.8 \cdot p(x|A)}{0.8 \cdot p(x|A) + 0.2 \cdot p(x|B)}$$

2. Mathematical Work for Decision Rule: The log-prior ratio is $\ln(0.8/0.2) = \ln(4)$.

$$(\mu_1 - \mu_2)^\top x + \frac{1}{2}(\|\mu_2\|^2 - \|\mu_1\|^2) + \ln(4) = 0$$

$$-x_1 - x_2 + 1 + \ln(4) = 0$$

Decision Boundary: $x_1 + x_2 = 1 + \ln(4)$. The boundary remains **linear**.

(c) Different Covariance Matrices

Parameters: $\mu_1 = [0,0]^\top, \mu_2 = [1,1]^\top, \Sigma_A = 0.25 I_d, \Sigma_B = I_d, P(A) = 0.5.$

1. Analytical Expression for $P(y|x)$: For class A, $|\Sigma_A|^{1/2} = (0.25)^{d/2} = 0.25$ for $d = 2$, and $\Sigma_A^{-1} = 4I_d$.

$$P(A|x) = \frac{1}{1 + \frac{1}{0.25}\exp\left(\frac{1}{2}(4\|x - \mu_1\|^2 - \|x - \mu_2\|^2)\right)}$$

2. Mathematical Work for Decision Rule:

$$\ln\left(\frac{1/0.25}{1}\right) - \frac{1}{2}(4\|x-\mu_1\|^2) + \frac{1}{2}\|x-\mu_2\|^2 = 0$$

$$\ln(4) - 2(x_1^2 + x_2^2) + \frac{1}{2}((x_1-1)^2 + (x_2-1)^2) = 0$$

$$\ln(4) - 2x_1^2 - 2x_2^2 + \frac{1}{2}(x_1^2 - 2x_1 + 1 + x_2^2 - 2x_2 + 1) = 0$$

$$-1.5x_1^2 - 1.5x_2^2 - x_1 - x_2 + (1 + \ln 4) = 0$$

Decision Boundary: The presence of x^2 terms indicates the boundary is **quadratic** (circular).

(d) Identical Means, Different Variances

Parameters: $\mu_1 = \mu_2 = [0,0]^\top, \Sigma_A = 0.25I_d, \Sigma_B = I_d, P(A) = 0.5$.

1. Analytical Expression for $P(y|x)$:

$$P(A|x) = \frac{1}{1 + 4\exp\left(\frac{1}{2}(4\|x\|^2 - \|x\|^2)\right)} = \frac{1}{1 + 4\exp(1.5\|x\|^2)}$$

2. Mathematical Work for Decision Rule: Using the derivation from (c) with $\mu_1 = \mu_2 = 0$:

$$\ln(4) - 1.5(x_1^2 + x_2^2) = 0 \implies x_1^2 + x_2^2 = \frac{\ln(4)}{1.5}$$

Decision Boundary: This is a **circle** centered at the origin.

(e) Realization by Gaussian Naive Bayes (GNB)

Answer: Yes, the Gaussian Naive Bayes assumes that features are independent within each class, which implies a diagonal covariance matrix. In all parts (a) through (d), the covariance matrices are identity or scaled identity matrices (I_d or $\sigma^2 I_d$), both of which are diagonal. Therefore, there exists a parameter vector within the GNB hypothesis space that exactly matches the Bayes optimal classifier h^*.

Question 4

Data set 1

The Gaussian Naive Bayes classifier did not draw a useful boundary line for the first data set. The GNB assumes that features are conditionally independent and could not account for the diagonal correlation. Instead, it misclassified many points with a vertical line.

The Logistic regression was successful in creating a decision boundary that maximizes the likelihood of the labels, successfully finding the diagonal gap between the two clusters.

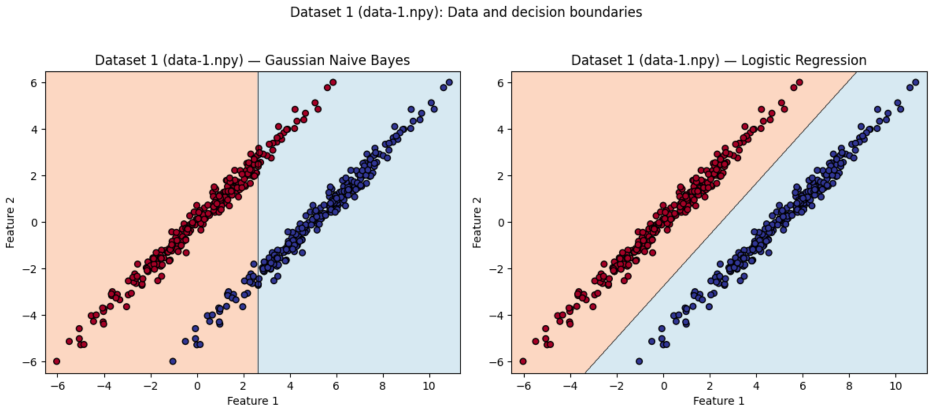

Figure 2: Data Set 1

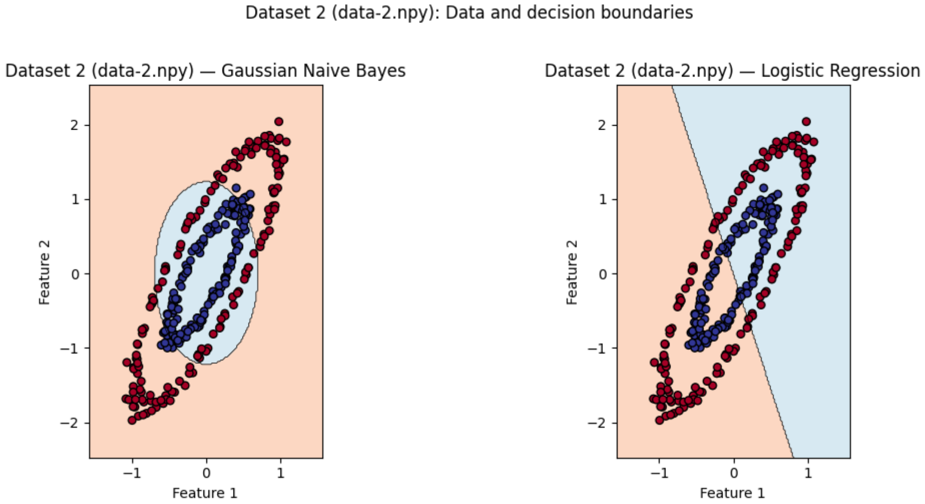

Figure 3: Data set 2

Data set 2

Similar to the first data set, the GNB classifier did not draw a useful boundary line. The GNB assumes that features are conditionally independent forcing it to create an axis-aligned elliptical boundary. This circular boundary fails to capture the diagonal orientation of the data, leading to significant overlap and misclassifications in the center.

The Logistic regression failed because LR finds the best possible straight line, and this dataset is not linearly separable, but more circular.

Data set 3

The Gaussian Naive Bayes classifier did draw a useful boundary line. This data set is axis-aligned and the two classes have significantly different variances. Whats interesting is that the features are technically independent in this specific circular orientation but GNB's ability to model different variances for each class allows it to create a quadratic decision boundary that perfectly encapsulates the inner blue circle.

Similar to the second data set, the Logistic regression failed because it finds the best possible straight line, and this dataset is not linearly separable, but more circular.

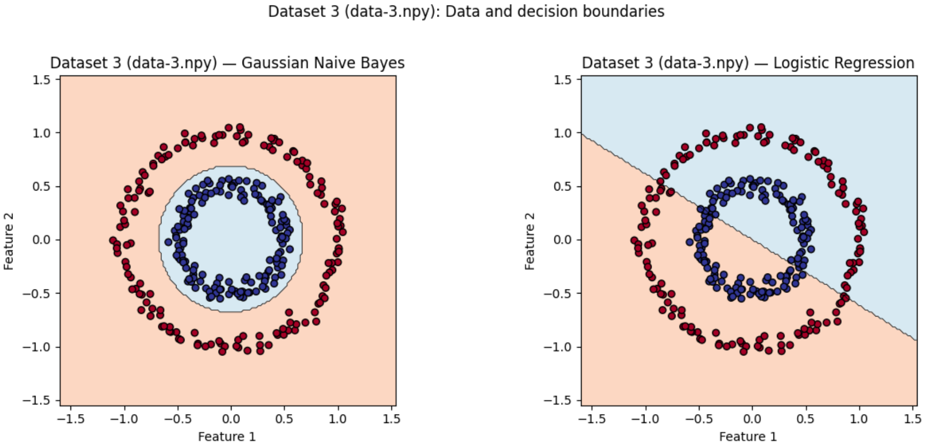

Figure 4: Data set 3