

## HW 3

Kevin Lin

2/23/2026

### 1

Given  $X_1, X_2, \dots, X_n$  are  $n$  i.i.d. Bernoulli random variables following:

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

where  $p$  is the probability of  $X$  being 1. We want to find the MLE of  $p$  using the given samples. The likelihood function is given by:

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

To find the MLE, we take the natural logarithm of the likelihood function:

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

Next, we differentiate  $\ell(p)$  with respect to  $p$  and set it to zero to find the critical points:

$$\frac{d\ell(p)}{dp} = \sum_{i=1}^n \left( \frac{x_i}{p} - \frac{1 - x_i}{1 - p} \right) = 0$$

This simplifies to:

$$\sum_{i=1}^n \frac{x_i}{p} = \sum_{i=1}^n \frac{1 - x_i}{1 - p}$$

Multiplying both sides by  $p(1 - p)$  gives:

$$(1 - p) \sum_{i=1}^n x_i = p \sum_{i=1}^n (1 - x_i)$$

Let  $S = \sum_{i=1}^n x_i$ , then we can rewrite the equation as:

$$\begin{aligned}(1-p)S &= p(n-S) \\ S - pS &= pn - pS \\ S &= pn \\ p &= \frac{S}{n} \therefore \\ \hat{p} &= \frac{\sum_{i=1}^n x_i}{n}\end{aligned}$$

Thus, the MLE of  $p$  is the sample mean of the observed data.

## 2

We are given a binary classification problem with outputs  $y \in \{0, 1\}$ , and assume that  $p(y|x)$  is Bernoulli  $\hat{p}(x; \theta^*)$  for some parameter vector  $\theta^* \in \mathbb{R}^d$ , where for each  $\theta$ ,  $\hat{p}(x; \theta)$  returns a number between  $[0, 1]$ .

- (a) We know the cross-entropy loss for each parameter  $\theta$  for a set of  $n$  examples with predictions  $\hat{p}(x_i; \theta)$  and true labels  $y_i$  is:

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

We can derive this loss function from the principle of the MLE:

$$\begin{aligned}\hat{\theta} &= \arg \max_{\theta} P(y_1, y_2, \dots, y_n | x_1, x_2, \dots, x_n; \theta) \\ &= \arg \max_{\theta} \prod_{i=1}^n P(y_i | x_i; \theta) \\ &= \arg \max_{\theta} \prod_{i=1}^n \hat{p}(x_i; \theta)^{y_i} (1 - \hat{p}(x_i; \theta))^{1-y_i} \\ &= \arg \max_{\theta} \sum_{i=1}^n (y_i \ln \hat{p}(x_i; \theta) + (1 - y_i) \ln(1 - \hat{p}(x_i; \theta))) \\ &= \arg \min_{\theta} -\frac{1}{n} \sum_{i=1}^n (y_i \ln \hat{p}(x_i; \theta) + (1 - y_i) \ln(1 - \hat{p}(x_i; \theta))) \\ &= \arg \min_{\theta} CE(\theta)\end{aligned}$$

Thus, minimizing the cross-entropy loss is equivalent to maximizing the likelihood of the observed data.

- (b) Let's instantiate the prediction as the sigmoid function applied to a linear combination of input features  $\hat{p}(x; \theta) = \sigma(\theta^T x)$ , where  $\sigma(z) = \frac{1}{1+e^{-z}}$ .

Then, minimizing the cross-entropy loss is equivalent to minimizing the familiar logistic loss given by:

$$L(\theta) = \frac{1}{n} \sum_{i=1}^n \ln(1 + e^{-\tilde{y}_i \theta^T x_i}) \text{ where } \tilde{y}_i = \begin{cases} 1 & \text{if } y_i = 1 \\ -1 & \text{if } y_i = 0 \end{cases}$$

We can show this by substituting  $\hat{p}(x_i; \theta) = \sigma(\theta^T x_i)$  into the cross-entropy loss:

$$\begin{aligned} CE(\theta) &= -\frac{1}{n} \sum_{i=1}^n (y_i \ln \sigma(\theta^T x_i) + (1 - y_i) \ln(1 - \sigma(\theta^T x_i))) \\ &= -\frac{1}{n} \sum_{i=1}^n \left( y_i \ln \frac{1}{1 + e^{-\theta^T x_i}} + (1 - y_i) \ln \left( 1 - \frac{1}{1 + e^{-\theta^T x_i}} \right) \right) \\ &= -\frac{1}{n} \sum_{i=1}^n \left( y_i \ln \frac{1}{1 + e^{-\theta^T x_i}} + (1 - y_i) \ln \frac{1}{1 + e^{\theta^T x_i}} \right) \\ &= -\frac{1}{n} \sum_{i=1}^n \left( -y_i \ln(1 + e^{-\theta^T x_i}) - (1 - y_i) \ln(1 + e^{\theta^T x_i}) \right) \\ &= \frac{1}{n} \sum_{i=1}^n \left( y_i \ln(1 + e^{-\theta^T x_i}) + (1 - y_i) \ln(1 + e^{\theta^T x_i}) \right) \end{aligned}$$

We know that:

$$\tilde{y}_i = \begin{cases} 1 & \text{if } y_i = 1 \\ -1 & \text{if } y_i = 0 \end{cases}$$

Thus, we can define  $\tilde{y}_i$  such that:

$$\begin{aligned} \tilde{y}_i &= 2y_i - 1 \therefore \\ y_i &= \frac{1 + \tilde{y}_i}{2}, 1 - y_i = \frac{1 - \tilde{y}_i}{2} \end{aligned}$$

Substituting this back into the expression for  $CE(\theta)$  gives:

$$CE(\theta) = \frac{1}{n} \sum_{i=1}^n \left( \frac{1 + \tilde{y}_i}{2} \ln(1 + e^{-\theta^T x_i}) + \frac{1 - \tilde{y}_i}{2} \ln(1 + e^{\theta^T x_i}) \right)$$

Thus, if we plug in  $\tilde{y}_i = 1$ , we get:

$$CE(\theta) = \frac{1}{n} \sum_{i=1}^n \ln(1 + e^{-\theta^T x_i})$$

And if we plug in  $\tilde{y}_i = -1$ , we get:

$$CE(\theta) = \frac{1}{n} \sum_{i=1}^n \ln(1 + e^{\theta^T x_i})$$

which is equivalent to the logistic loss function:

$$L(\theta) = \begin{cases} \frac{1}{n} \sum_{i=1}^n \ln(1 + e^{-\theta^T x_i}) & \text{if } \tilde{y}_i = 1 \\ \frac{1}{n} \sum_{i=1}^n \ln(1 + e^{\theta^T x_i}) & \text{if } \tilde{y}_i = -1 \end{cases}$$

### 3

- (a) We know that data from class A are Gaussian  $\mathcal{N}(\mu_1, I_d)$  and data from class B are also Gaussian  $\mathcal{N}(\mu_2, I_d)$ , where the prior probabilities of both classes are equal. We can derive the posterior probability  $P(y|x)$  for each class using Bayes' theorem:

$$\begin{aligned} P(y|x) &= \frac{P(x|y)P(y)}{P(x)} \\ P(x) &= P(x|y=0)P(y=0) + P(x|y=1)P(y=1) \\ P(y=0|x) &= \frac{P(x|y=0)P(y=0)}{P(x|y=0)P(y=0) + P(x|y=1)P(y=1)} \\ P(y=1|x) &= \frac{P(x|y=1)P(y=1)}{P(x|y=0)P(y=0) + P(x|y=1)P(y=1)} \end{aligned}$$

Since the prior probabilities are equal, we can simplify this to:

$$\begin{aligned} P(y=0|x) &= \frac{P(x|y=0)}{P(x|y=0) + P(x|y=1)} \\ P(y=1|x) &= \frac{P(x|y=1)}{P(x|y=0) + P(x|y=1)} \end{aligned}$$

We can further simplify this by substituting the Gaussian probability density functions for  $P(x|y=0)$  and  $P(x|y=1)$ :

$$\begin{aligned} P(x|y=0) &= \frac{1}{(2\pi)^{d/2}} e^{-\frac{1}{2}\|x-\mu_1\|^2} \\ P(x|y=1) &= \frac{1}{(2\pi)^{d/2}} e^{-\frac{1}{2}\|x-\mu_2\|^2} \end{aligned}$$

Thus, the posterior probabilities can be expressed as:

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

$$P(y = 1|x) = \frac{e^{-\frac{1}{2}\|x-\mu_2\|^2}}{e^{-\frac{1}{2}\|x-\mu_1\|^2} + e^{-\frac{1}{2}\|x-\mu_2\|^2}}$$

The decision rule for the Bayes optimal classifier is to assign  $x$  to class A if  $P(y = 0|x) > P(y = 1|x)$  and to class B otherwise. This can be simplified to:

$$P(y = 0|x) > P(y = 1|x)$$

$$\frac{e^{-\frac{1}{2}\|x-\mu_1\|^2}}{e^{-\frac{1}{2}\|x-\mu_1\|^2} + e^{-\frac{1}{2}\|x-\mu_2\|^2}} > \frac{e^{-\frac{1}{2}\|x-\mu_2\|^2}}{e^{-\frac{1}{2}\|x-\mu_1\|^2} + e^{-\frac{1}{2}\|x-\mu_2\|^2}}$$

$$e^{-\frac{1}{2}\|x-\mu_1\|^2} > e^{-\frac{1}{2}\|x-\mu_2\|^2}$$

$$-\frac{1}{2}\|x - \mu_1\|^2 > -\frac{1}{2}\|x - \mu_2\|^2$$

$$\|x - \mu_1\|^2 < \|x - \mu_2\|^2$$

Thus, the decision boundary is the set of points  $x$  such that:

$$\|x - \mu_1\|^2 = \|x - \mu_2\|^2$$

This is the equation of a hyperplane that is equidistant from the two means  $\mu_1$  and  $\mu_2$ . We can simplify this further by expanding the norms:

$$(x - \mu_1)^T(x - \mu_1) = (x - \mu_2)^T(x - \mu_2)$$

$$x^T x - 2\mu_1^T x + \mu_1^T \mu_1 = x^T x - 2\mu_2^T x + \mu_2^T \mu_2$$

$$-2\mu_1^T x + \mu_1^T \mu_1 = -2\mu_2^T x + \mu_2^T \mu_2$$

$$2(\mu_2 - \mu_1)^T x = \mu_2^T \mu_2 - \mu_1^T \mu_1$$

$$(\mu_2 - \mu_1)^T x = \frac{1}{2}(\mu_2^T \mu_2 - \mu_1^T \mu_1)$$

Given that  $\mu_1 = (0, 0)$  and  $\mu_2 = (1, 1)$ , we can substitute these values into the equation for the decision boundary:

$$(\mu_2 - \mu_1)^T x = \frac{1}{2}(\mu_2^T \mu_2 - \mu_1^T \mu_1)$$

$$(1, 1)^T x = \frac{1}{2}(1^2 + 1^2 - 0^2 - 0^2)$$

$$x_1 + x_2 = 1$$

Dataset	Gaussian Naive Bayes	Linear Logistic Regression
1	Each feature has mixed overlap between 1 and 2. Because GNB assumes that each feature is independent, they are modeled separately and thus the boundary becomes a vertical line.	Dataset is linearly separable so logistic regression performs really well, giving us a clean fit.
2	The decision boundary becomes quadratic, so GNB can create a smooth circled boundary. However because the features are correlated in a slanted ellipse, naive bayes fails to recognize this and thus still somewhat poorly fits, but is still better than logistic regression.	Fails as data is not linearly separable but rather overlaps. Thus, it draws a line that minimizes classification error, but it cannot match the ellipse structure.
3	Each class is radially symmetric, and even though features are not independent (related to radius), GNB can still model it well in circles.	Fails as data is not linearly separable. Thus again, it draws a line that minimizes error but cannot match circular shape as it is constrained to 1 dimension

#### 4

(a) See attached Jupyter notebook for code and plots.

(b)