Machine Learning

Instructors: Mohammadreza A. Dehaqani, Babak n. Arabi

Shahzad Momayez, Roshak Masoudi Nejad

Fall 2025

Homework 1

Question 1: Minimum Risk Classification with Rejection Option

In many pattern classification problems we have the option to either assign a pattern to one of c classes, or reject it as unrecognizable—if the cost to reject is not too high. Let the cost of classification be defined as:

$$\lambda(\omega_i|\omega_j) = \begin{cases} 0, & \text{if } \omega_i = \omega_j, \quad \text{(i.e., Correct Classification)} \\ \lambda_r, & \text{if } \omega_i = \omega_0, \quad \text{(i.e., Rejection)} \\ \lambda_s, & \text{Otherwise, (i.e., Substitution Error)} \end{cases}$$

Show that for the minimum risk classification, the decision rule should associate a test vector x with class ω_i, if

$$P(\omega_i|x) \geq P(\omega_j|x) \quad \forall j \quad \text{and} \quad P(\omega_i|x) \geq 1 - \frac{\lambda_r}{\lambda_s},$$

and reject otherwise.

Solution: Average risk is choosing class ω_i:

$$R(\omega_i|x) = \sum_{j=1}^{c} \lambda(\omega_i|\omega_j) P(\omega_j|x) = 0^* P(\omega_i|x) + \sum_{j=1, j\neq i}^{c} \lambda_s P(\omega_j|x)$$

where $\lambda(\omega_i|\omega_j)$ is used to mean the cost of choosing class ω_i where the true class is ω_j.
Hence,

$$R(\omega_i|x) = \lambda_s(1 - P(\omega_i|x))$$

Associate x with the class ω_i if the highest posterior class probability and the average risk are less than the cost of rejection:

$$\lambda_s(1 - P(\omega_i|x)) \leq \lambda_r$$

$$P(\omega_i|x) \geq 1 - \frac{\lambda_r}{\lambda_s}$$

Question 2: Naive Bayes

Consider a **Naive Bayes** classification problem with 3 classes and 2 features. One of these features follows a Bernoulli distribution, and the other follows a Gaussian distribution. The features are denoted by $X = [X_1, X_2]^T$, and the class by Y.

Prior distribution

$$P[Y = 0] = 0.5, \quad P[Y = 1] = 0.25, \quad P[Y = 2] = 0.25$$

This homework 1 is due some due date, and the version of the problem set is December 16, 2025.

Feature distributions

$$p_{X_1|Y}(x_1 \mid Y = c) = Ber(x_1; \theta_c)$$
$$p_{X_2|Y}(x_2 \mid Y = c) = Normal(x_2; \mu_c, \sigma_c^2)$$

Parameter assumptions

$$\sigma_c^2 = \begin{cases} 1 & \text{if } c = 0 \\ 1 & \text{if } c = 1 \\ 1 & \text{if } c = 2 \end{cases} \quad \mu_c = \begin{cases} -1 & \text{if } c = 0 \\ 0 & \text{if } c = 1 \\ 1 & \text{if } c = 2 \end{cases} \quad \theta_c = \begin{cases} 0.5 & \text{if } c = 0 \\ 0.75 & \text{if } c = 1 \\ 0.5 & \text{if } c = 2 \end{cases}$$

Questions (a) Compute $p_{Y|X_1,X_2}(y \mid x_1 = 0, x_2 = 0)$. (The answer should be a vector in $\mathbb{R}^3$ whose elements sum to 1.)

(b) Compute $p_{Y|X_1}(y \mid x_1 = 0)$.

(c) Compute $p_{Y|X_2}(y \mid x_2 = 1)$.

(d) Analyze the pattern observed in the results from the previous parts.

Solution:

Part (a): Compute $p_{Y|X_1,X_2}(y|x_1 = 0, x_2 = 0)$. This is the posterior distribution of Y given $X_1 = 0$ and $X_2 = 0$.

We can apply **Bayes' Theorem**:

$$p_{Y|X_1,X_2}(y|x_1 = 0, x_2 = 0) = \frac{p_{X_1,X_2|Y}(x_1 = 0, x_2 = 0|y) \cdot P[Y = y]}{p_{X_1,X_2}(x_1 = 0, x_2 = 0)}$$

* The **likelihood** $p_{X_1,X_2|Y}(x_1 = 0, x_2 = 0|y)$ is the product of the individual feature likelihoods:

$$p_{X_1|Y}(x_1 = 0|y) \cdot p_{X_2|Y}(x_2 = 0|y)$$

* For X_1, since $X_1 \sim$ Bernoulli(θ_y):

$$p_{X_1|Y}(x_1 = 0|y) = 1 - \theta_y$$

* For X_2, since $X_2 \sim$ Normal$(\mu_y, 1)$ (since $\sigma_y^2 = 1$):

$$p_{X_2|Y}(x_2 = 0|y) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(\mu_y - 0)^2}{2}\right)$$

Now calculate for each class y:

* For $y = 0$, we have $\theta_0 = 0.5$ and $\mu_0 = -1$:

$$p_{X_1|Y}(x_1 = 0|0) = 1 - 0.5 = 0.5$$

$$p_{X_2|Y}(x_2 = 0|0) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(-1)^2}{2}\right) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\right)$$

So,

$$p_{X_1,X_2|Y}(x_1 = 0, x_2 = 0|0) = 0.5 \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\right)$$

* Similarly, for $y = 1$, $\theta_1 = 0.75$ and $\mu_1 = 0$:

$$p_{X_1|Y}(x_1 = 0|1) = 1 - 0.75 = 0.25$$

$$p_{X_2|Y}(x_2 = 0|1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(0)^2}{2}\right) = \frac{1}{\sqrt{2\pi}}$$

So,

$$p_{X_1,X_2|Y}(x_1 = 0, x_2 = 0|1) = 0.25 \cdot \frac{1}{\sqrt{2\pi}}$$

* For $y = 2$, $\theta_2 = 0.5$ and $\mu_2 = 1$:

$$p_{X_1|Y}(x_1 = 0|2) = 1 - 0.5 = 0.5$$

$$p_{X_2|Y}(x_2 = 0|2) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(1)^2}{2}\right) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\right)$$

So,

$$p_{X_1,X_2|Y}(x_1 = 0, x_2 = 0|2) = 0.5 \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\right)$$

* Now we need the **normalizing constant** $p_{X_1,X_2}(x_1 = 0, x_2 = 0)$, which is the total probability:

$$p_{X_1,X_2}(x_1 = 0, x_2 = 0) = \sum_{y=0}^{2} p_{X_1,X_2|Y}(x_1 = 0, x_2 = 0|y) \cdot P[Y = y]$$

Plugging in the values from the prior distribution and the likelihoods calculated above, we get the normalizing constant.

Part (b): Compute $p_{Y|X_1}(y|x_1 = 0)$. This is the posterior distribution of Y given only $X_1 = 0$. We use Bayes' Theorem again:

$$p_{Y|X_1}(y|x_1 = 0) = \frac{p_{X_1|Y}(x_1 = 0|y) \cdot P[Y = y]}{p_{X_1}(x_1 = 0)}$$

Here, the likelihood for $X_1 = 0$ is simply:

$$p_{X_1|Y}(x_1 = 0|y) = 1 - \theta_y$$

The normalizing constant is:

$$p_{X_1}(x_1 = 0) = \sum_{y=0}^{2} p_{X_1|Y}(x_1 = 0|y) \cdot P[Y = y]$$

Part (c): Compute $p_{Y|X_2}(y|x_2 = 1)$. This is similar to part (b), but now given $X_2 = 1$. We apply Bayes' Theorem:

$$p_{Y|X_2}(y|x_2 = 1) = \frac{p_{X_2|Y}(x_2 = 1|y) \cdot P[Y = y]}{p_{X_2}(x_2 = 1)}$$

The likelihood for $X_2 = 1$ is:

$$p_{X_2|Y}(x_2 = 1|y) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(\mu_y - 1)^2}{2}\right)$$

The normalizing constant is:

$$p_{X_2}(x_2 = 1) = \sum_{y=0}^{2} p_{X_2|Y}(x_2 = 1|y) \cdot P[Y = y]$$

Part (d): Analyze the Pattern Observed in the Results. Once you calculate the posterior distributions for each part, you can analyze the patterns based on the likelihoods, priors, and the normalizing constants. For example, the effects of the Gaussian and Bernoulli distributions on the posterior could differ, and the influence of different priors on the classification results can also be analyzed.

Question 3: Multinomial Naive Bayes

Consider a Multinomial Naive Bayes model for a two-class text classification problem. Assume that the total number of words in our dictionary (the total number of model features) is d. For a text input sample x, the values $c_1, c_2, \ldots, c_p$ form the feature vector. That is, each c_i represents the number of times the i-th word appears in the document. The parameters of this model are as follows: (y is the model output or the class label of the sample.)

$$P_y = P(y = 1)$$

$$P_{i|y=1} = P(\text{word } i \text{ appears in a specific document position} \mid y = 1)$$

$$P_{i|y=.} = P(\text{word } i \text{ appears in a specific document position} \mid y = .)$$

(a) Write an expression for the conditional probability $P(y = 1 \mid x)$ for a text sample x in terms of the model parameters.

(b) Show that the decision boundary of the trained model is linear.

(c) Show that the conditional probability written in part (a) is a logistic function:

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-(\theta^T x + \theta_0)}}$$

Solution:

Part (a): Expression for the conditional probability (P(y = 1 | x)). For a Multinomial Naive Bayes model, we use Bayes' Theorem:

$$P(y = 1|x) = \frac{P(y = 1)P(x|y = 1)}{P(x)}$$

Here:
* $P(y = 1)$ is the prior probability of the class $y = 1$. * $P(x|y = 1)$ is the likelihood, which is the probability of observing the feature vector x given the class label $y = 1$. Since each word is considered conditionally independent, the likelihood for each feature c_i is:

$$P(x|y = 1) = \prod_{i=1}^{d} P(c_i|y = 1)$$

So the expression for the conditional probability becomes:

$$P(y = 1|x) = \frac{P(y = 1) \prod_{i=1}^{d} P(c_i|y = 1)}{P(x)}$$

Where $P(x)$ is the evidence or the total probability of observing the feature vector x, which can be computed by summing over all possible class labels:

$$P(x) = P(y = 1) \prod_{i=1}^{d} P(c_i|y = 1) + P(y = 0) \prod_{i=1}^{d} P(c_i|y = 0)$$

Thus, the conditional probability $P(y = 1|x)$ can be written as:

$$P(y = 1|x) = \frac{P(y = 1) \prod_{i=1}^{d} P(c_i|y = 1)}{P(y = 1) \prod_{i=1}^{d} P(c_i|y = 1) + P(y = 0) \prod_{i=1}^{d} P(c_i|y = 0)}$$

Part (b): Show that the decision boundary is linear. The decision boundary is where the model's prediction changes between the two classes. We can find this by setting the conditional probability equal for both classes, i.e., $P(y=1|x) = P(y=0|x)$.

Using Bayes' Theorem, we have:

$$\frac{P(y=1)\prod_{i=1}^{d} P(c_i|y=1)}{P(y=1)\prod_{i=1}^{d} P(c_i|y=1) + P(y=0)\prod_{i=1}^{d} P(c_i|y=0)} = \frac{P(y=0)\prod_{i=1}^{d} P(c_i|y=0)}{P(y=1)\prod_{i=1}^{d} P(c_i|y=1) + P(y=0)\prod_{i=1}^{d} P(c_i|y=0)}$$

Cross-multiplying and simplifying:

$$P(y=1)\prod_{i=1}^{d} P(c_i|y=1) = P(y=0)\prod_{i=1}^{d} P(c_i|y=0)$$

Taking the logarithm of both sides:

$$\log P(y=1) + \sum_{i=1}^{d} \log P(c_i|y=1) = \log P(y=0) + \sum_{i=1}^{d} \log P(c_i|y=0)$$

This simplifies to:

$$\sum_{i=1}^{d} \log \frac{P(c_i|y=1)}{P(c_i|y=0)} = \log \frac{P(y=0)}{P(y=1)}$$

Thus, the decision boundary is defined by a linear function of the feature vector x. The boundary is linear because the logarithms of the likelihoods and priors yield a linear expression.

Part (c): Show that the conditional probability is a logistic function. To show that the conditional probability $P(y=1|x)$ is a logistic function, we take the natural logarithm of the odds ratio between the two classes:

$$\log \frac{P(y=1|x)}{P(y=0|x)} = \log \frac{P(y=1)}{P(y=0)} + \sum_{i=1}^{d} \log \frac{P(c_i|y=1)}{P(c_i|y=0)}$$

This can be expressed as a linear function of the features x, and therefore:

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

Where $f(x)$ is a linear function of x, i.e.,

$$f(x) = \theta^T x + \theta_0$$

Thus, the conditional probability can be written as:

$$P(y=1|x) = \frac{1}{1+e^{-(\theta^T x+\theta_0)}}$$

This is exactly the logistic function.

Question 4: MLE

(a) Let $X_1, X_2, ..., X_N$ be data drawn independently from a uniform distribution over a diamond-shaped area with edge length $\sqrt{2}\theta$ in $\mathbb{R}^2$, where $\theta \in \mathbb{R}^+$ (see Figure 1). Thus, $X_i \in \mathbb{R}^2$ and the distribution is

$$p(x \mid \theta) = \begin{cases} \frac{1}{2\theta^2}, & \text{if } \|x\|_1 \le \theta, \\ 0, & \text{otherwise,} \end{cases}$$

where $\|x\|_1 = |x_1| + |x_2|$ is the L_1 norm. Find the maximum likelihood estimate of θ.

(b) Let $X_1, X_2, ..., X_n$ be a random sample from a Gamma(α, β) population. Find the maximum likelihood estimate (MLE) of β, assuming α is known.

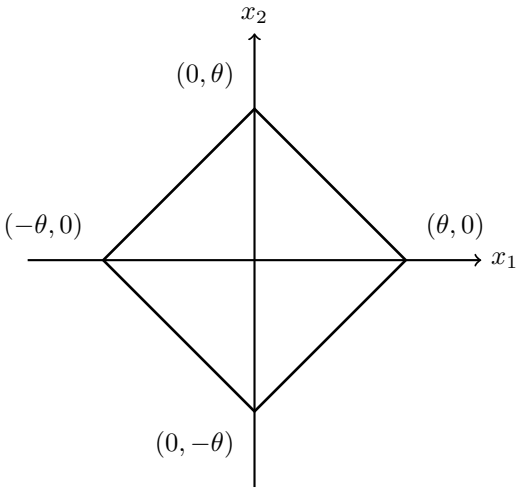

Figure 1: Area $\|x\|_1 \leq \theta$

Solution:

Part (a): As $X_1, X_2, \ldots, X_N$ are samples that have been drawn from the distribution, their probability of occurrence must be non-zero. Therefore, each of $X_1, X_2, \ldots, X_N$ must obey the condition $\|x\| \leq \theta$. Knowing this, we can write the likelihood function as:

$$L(X_1, X_2, \ldots, X_N; \theta) = \frac{1}{(2\theta)^N} \mathbb{1}\left\{ \max_{1 \leq i \leq N} \|X_i\| \leq \theta \right\}$$

To maximize likelihood, we want θ to be as small as possible with the constraint of

$$\max_{1 \leq i \leq N} \|X_i\| \leq \theta, \text{ otherwise the likelihood drops to 0.}$$

So the MLE of θ is

$$\hat{\theta} = \max_{1 \leq i \leq N} \|X_i\|$$

Part (b): The likelihood function is

$$L(\beta \mid \mathbf{x}) = \prod_{i=1}^{n} \frac{1}{\Gamma(\alpha)\beta^\alpha} x_i^{\alpha-1} \exp\left(-\frac{x_i}{\beta}\right)$$

$$= \left(\frac{1}{\Gamma(\alpha)}\right)^n \frac{1}{\beta^{n\alpha}} \left(\prod_{i=1}^{n} x_i\right)^{\alpha-1} \exp\left(-\frac{\sum_{i=1}^n x_i}{\beta}\right)$$

$$\log L(\beta \mid \mathbf{x}) = -n \log \Gamma(\alpha) - n\alpha \log \beta + (\alpha - 1) \sum_{i=1}^{n} \log x_i - \frac{\sum_{i=1}^n x_i}{\beta}$$

$$\frac{\partial \log L(\beta \mid \mathbf{x})}{\partial \beta} = -\frac{n\alpha}{\beta} + \frac{\sum_{i=1}^n x_i}{\beta^2} = 0 \quad \Rightarrow \quad \hat{\beta} = \frac{\bar{x}}{\alpha}$$

Setting the partial derivative equal to zero and solving gives

$$\hat{\beta} = \frac{\bar{x}}{\alpha}.$$

To check that this is a maximum,

$$\frac{\partial^2 \log L(\beta \mid \mathbf{x})}{\partial \beta^2} = \frac{n\alpha}{\beta^2} - \frac{2\sum_{i=1}^n x_i}{\beta^3}$$

At $\beta = \hat{\beta}$,

$$\left.\frac{\partial^2 \log L(\beta \mid \mathbf{x})}{\partial \beta^2}\right|_{\beta=\hat{\beta}} = \frac{n\alpha}{(\bar{x}/\alpha)^2} - \frac{2n\bar{x}}{(\bar{x}/\alpha)^3} = -\frac{n\alpha^3}{\bar{x}^2} < 0$$

Because $\frac{\bar{x}}{\alpha}$ is the only extremum and it corresponds to a global maximum,

$$\boxed{\hat{\beta} = \frac{\bar{x}}{\alpha}}$$

is the maximum likelihood estimator (MLE) of β.

Question 5: MAP

MAP of Multinational Distribution

You have just got a loaded 6-sided dice from your statistician friend. Unfortunately, he does not remember its exact probability distribution $p_1, p_2, \ldots, p_6$. He remembers, however, that he generated the vector $(p_1, p_2, \ldots, p_6)$ from the following Dirichlet distribution:

$$\mathbb{P}(p_1, p_2, \ldots, p_6) = \frac{\Gamma\left(\sum_{i=1}^6 u_i\right)}{\prod_{i=1}^6 \Gamma(u_i)} \prod_{i=1}^6 p_i^{u_i - 1}\, \delta\left(\sum_{i=1}^6 p_i - 1\right),$$

where he chose $u_i = i$ for all $i = 1, \ldots, 6$. Here Γ denotes the gamma function, and δ is the Dirac delta function.
To estimate the probabilities $p_1, p_2, \ldots, p_6$, you roll the dice 1000 times and then observe that side i occurred n_i times, where

$$\sum_{i=1}^6 n_i = 1000.$$

(a) Conjugacy of the Dirichlet Prior:
Prove that the Dirichlet distribution is a conjugate prior for the multinomial distribution.
(b) Posterior Distribution:
Given the data $\{n_1, n_2, \ldots, n_6\}$, find the posterior distribution of the side probabilities:

$$\mathbb{P}(p_1, p_2, \ldots, p_6 \mid n_1, n_2, \ldots, n_6).$$

Solution:

Part (a): We solve the problem with the above notation for the 6-sided dice. If you solve it for the more general case, that is also fine. Our data is $n_1, n_2, \ldots, n_6$, and the distribution we study is multinomial distribution. Therefore, the likelihood function is

$$P(n_1, n_2, \ldots, n_6 | p_1, p_2, \ldots, p_6) = \prod_{i=1}^6 p_i^{n_i}$$

To see that the Dirichlet distribution is conjugate prior to the multinomial distribution, we need to calculate the product of the likelihood and the Dirichlet distribution.

$$\begin{aligned} P(n_1, n_2, \ldots, n_6, p_1, p_2, \ldots, p_6) &= \left(\prod_{i=1}^6 p_i^{n_i}\right) \frac{\Gamma\left(\sum_{i=1}^6 u_i\right)}{\prod_{i=1}^6 \Gamma(u_i)} \prod_{i=1}^6 p_i^{u_i - 1} \delta\left(\sum_{i=1}^6 p_i - 1\right) \\ &= \frac{\Gamma\left(\sum_{i=1}^6 u_i\right)}{\prod_{i=1}^6 \Gamma(u_i)} \prod_{i=1}^6 p_i^{n_i + u_i - 1} \delta\left(\sum_{i=1}^6 p_i - 1\right) \end{aligned}$$

where const doesn't depend on $p_1, p_2, \ldots, p_6$. Now we can see that the posterior and the prior are both Dirichlet distributions, thereby, Dirichlet prior is a conjugate prior for the multinomial distribution.

Part (b): From the above equation, we have that the posterior is

$$P\left(\{p_i\}_{i=1}^6 \middle| \{n_i\}_{i=1}^6\right) = \frac{\Gamma\left(\sum_{i=1}^6 n_i + \mu_i\right)}{\prod_{i=1}^6 \Gamma(n_i + \mu_i)} \prod_{i=1}^6 p_i^{n_i+\mu_i-1} \delta\left(\sum_{i=1}^6 p_i - 1\right)$$

where equation (2) is the posterior.