Lab1

ECE 368

Question 1: Classifying Spam vs non-Spam Emails

- We want to solve a binary classification problem for detecting spam vs non-spam emails.
- We have a training set containing N emails, and each email n is represented by $\{\mathbf{x}_n, y_n\}, n = 1, 2, \ldots, N$, where y_n is the class label which takes the value

$$y_n = \begin{cases} 1 & \text{if email } n \text{ is spam,} \\ 0 & \text{if email } n \text{ is non-spam (also called ham),} \end{cases}$$

  and $\mathbf{x}_n$ is a feature vector of the email n.
- Let $\mathcal{W} = \{w_1, w_2, \ldots, w_D\}$ be the set of the words (called the vocabulary) that appear at least once in the training set.
- The feature vector $\mathbf{x}_n$ is defined as a D-dimensional vector $\mathbf{x}_n = [x_{n1}, x_{n2}, \ldots, x_{nD}]$, where each entry $x_{nd}, d = 1, 2, \ldots, D$ is the number of occurrences of word w_d in email n. Thus the total number of words in email n can be expressed as $l_n = x_{n1} + x_{n2} + \ldots + x_{nD}$.

Question 1: Classifying Spam vs non-Spam Emails

What is the Naïve Bayes Classifier

- Recall the Bayes rule

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

Question 1: Classifying Spam vs non-Spam Emails

What is the Naïve Bayes Classifier

- Recall the Bayes rule

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

- We assume that the prior class distribution $p(y_n)$ is modeled as

$$p(y_n = 1) = \pi,$$
$$p(y_n = 0) = 1 - \pi,$$

where π is a fixed parameter (e.g., 0.5).

Question 1: Classifying Spam vs non-Spam Emails

What is the Naïve Bayes Classifier

- Recall the Bayes rule

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

- We assume that the prior class distribution $p(y_n)$ is modeled as

$$p(y_n = 1) = \pi,$$
$$p(y_n = 0) = 1 - \pi,$$

  where π is a fixed parameter (e.g., 0.5).
- $p(x|y)$ is unknown in this formula and we need to learn it from the data.

Question 1: Classifying Spam vs non-Spam Emails

What is the Naïve Bayes Classifier

- Recall the Bayes rule

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

- We assume that the prior class distribution $p(y_n)$ is modeled as

$$p(y_n = 1) = \pi,$$
$$p(y_n = 0) = 1 - \pi,$$

  where π is a fixed parameter (e.g., 0.5).
- $p(x|y)$ is unknown in this formula and we need to learn it from the data.
- **Assumption**

$$p(x = [x_{n1}, x_{n2}, \ldots, x_{nD}]|y) = p(x_1|y)p(x_2|y)\ldots p(x_D|y).$$

- We have a discrete probability space. Why?
- We want to learn $P(x = w_i|y = j)$ for $i \in \{1, \ldots, D\}$ and $j \in \{0, 1\}$ from the training data.

What is the probabilistic model: Mutlinomial Distribution

$$p\left(\mathbf{x}_n \mid y_n\right)=\frac{\left(x_{n 1}+x_{n 2}+\ldots+x_{n D}\right)!}{\left(x_{n 1}\right)!\left(x_{n 2}\right)!\ldots\left(x_{n D}\right)!} \prod_{d=1}^D p\left(w_d \mid y_n\right)^{x_{n d}} .$$

Objectives:

1. You want to use maximum likelihood estimates for learning $p(x=w_i|y=j)$ for $i \in \{1,\ldots,D\}$ and $j \in \{0,1\}$.
2. The maximum likelihood estimates are not the most appropriate to use when the probabilities are very close to 0 or to 1. For example, some words that occur in one class may not occur at all in the other class. In this problem, we use the technique of Laplace smoothing to deal with this problem.
3. What is the technique of Laplace smoothing?
4. After learning $p(x=w_i|y=j)$ for $i \in \{1,\ldots,D\}$ and $j \in \{0,1\}$ we want to use it for classification of the test set.
5. The classification is based on MAP rule.

$$\hat{y}_n=\begin{cases}1 & \text{if} p(y=1|x) \geq p(y=0|x), \\ 0 & \text{if } p(y=1|x)<p(y=0|x),\end{cases}$$

6. There are two types of errors in classifying unlabeled emails: Type 1 error is defined as the event that a spam email is misclassified as ham; Type 2 error is defined as the event that a ham email is misclassified as spam. How to tradeoff these two errors?

Question 2:Linear/Quadratic Discriminant Analysis for Height/Weight Data

- We want to solve a binary classification problem.
- Let $\mathbf{x}_n = [h_n, w_n]$ be the feature vector, where h_n denotes the height and w_n denotes the weight of a person indexed by n. Let y_n denote the class label. Here $y_n = 1$ is male, and $y_n = 2$ is female. We model the class prior as $p(y_n = 1) = \pi$ and $p(y_n = 2) = 1 - \pi$. For this problem, let $\pi = 0.5$.
- When the feature vector is real-valued (instead of binary), a Gaussian vector model is appropriate, i.e.,

$$p(\mathbf{x}|y_n = c) \propto \frac{1}{|\Sigma_c|} e^{-\frac{1}{2}(\mathbf{x}-\mu_c)^T \Sigma_c^{-1} (\mathbf{x}-\mu_c)}, \quad c \in \{f, m\}. \tag{1}$$

- For LDA, a common covariance matrix is shared by both classes, which is denoted by $\mathbf{\Sigma}$; for QDA, different covariance matrices are used for male and female, which are denoted by $\mathbf{\Sigma}_m$ and $\mathbf{\Sigma}_f$, respectively.
- For LDA: estimate $\boldsymbol{\mu}_m, \boldsymbol{\mu}_f$, and $\mathbf{\Sigma}$.
- For QDA: estimate $\boldsymbol{\mu}_m, \boldsymbol{\mu}_f, \mathbf{\Sigma}_m$, and $\mathbf{\Sigma}_f$.

LDA and QDA

- **Training**: We want to use the ML to estimate the LDA/QDA parameters.
- Based on the Bayes classifier, we then want to plot the decision boundary in both cases. **What is the difference between LDA and QDA?**
- **Testing**: Compute the misclassification rate for both cases.