STA 141A

Fall 2016

Lecture Note : Classification

1 Classification problem

A classification problem is the problem of deciding whether observation from a population belongs to one of a fixed number of (and known) subpopulations. Some examples:

- A patient is brought the emergency room of a hospital with chest pain. Could this be due to a heart attack or due to indigestion ?
- An online banking service needs to decide whether a transaction being performed is valid or fraudulent.
- Based on satellite imagery of patch of earth, one needs to decide whether the patch is covered by vegetation, or barren ground, or a water body.

In each of these cases, and in many similar instances, the decision making process is *learned* by making use of *training data* that contain the *class labels*.

Mathematically, such a data set can be generically represented as $\{(X_i, Y_i) : i = 1, \ldots, n\}$, where X_i is the set of measurements on the i-th subject and Y_i is the corresponding class label, that is the indicator of the subpopulation from which X_i (or the corresponding subject) was drawn. If there are $K \geq 2$ subpopulations (or classes) then Y_i can take values from $1, \ldots, K$. These numbers are referred to as *class labels*. Given such a data set, and given a new observation X from the same population, a *classification problem* is the task of *predicting the class label* of X, that is deciding on which subpopulation the corresponding subject belongs to.

- **Misclassification error:** Since the problem of classification is a problem of decision making, there is a scope of wrong decisions, that is, given X with *unobserved* class label Y, assigning it the class label $\hat{Y}$ (a value between 1 and K, in a K-class classification problem) that is different from Y.
- **Misclassification error rate:** This is the probability of assigning incorrect class label to the observation, i.e., $\mathbb{P}(\hat{Y} \neq Y)$. Given a *test data set*, $\{(X_i, Y_i) : i = 1, \ldots, m\}$, suppose we predict the class labels for X_i's for any specific classification procedure. Then the estimated misclassification rate will be

$$\frac{1}{m} \sum_{i=1}^{m} I(\hat{Y}_i \neq Y_i) \tag{1}$$

  where $\hat{Y}_i$ is the predicted class label for X_i. Here $I(\cdot)$ refers to the indicator function, i.e., its value is 1 if the expression inside the parentheses is valid, and is zero otherwise.

1.1 Bayes classifier

Suppose that X is a continuous random variable (or random vector), and that, if X belongs to class j, then the probability density function of X is $f_j(x)$, for $j = 1, \ldots, K$. Notice that $f_j(x)$ can

be seen as the conditional probability density function of X given $Y = j$. Suppose also that the probability of an observation belonging to class j is π_j, where $\pi_1, \ldots, \pi_K > 0$ and $\sum_{k=1}^{K} \pi_k = 1$. Thus, π_j indicates the fraction of subjects in the population that belong to class j.

Then, using the *Bayes' Theorem* for conditional probability calculation, it can be derived that the probability of $Y = j$ (i.e., the observation is in class j) when $X = x$ (a given value) is

$$\mathbb{P}(Y = j|X = x) = \frac{\pi_j f_j(x)}{\sum_{k=1}^{K} \pi_k f_k(x)}. \tag{2}$$

It can be shown that, the optimal classifier, in the sense of minimizing the (theoretical) misclassification rate, is the one that maximizes $\mathbb{P}(Y = j|X = x)$ over the possible values $1, \ldots, K$ of the index j. This optimal classifier is called the *Bayes classifier*. In other words, when $X = x$, the Bayes classifier assigns the value j to the predicted class label $\hat{Y}$, if and only if,

$$\mathbb{P}(Y = j|X = x) > \mathbb{P}(Y = k|X = x) \qquad \text{for all } k \neq j, \tag{3}$$

$$\text{that is,} \qquad \pi_j f_j(x) > \pi_k f_k(x) \qquad \text{for all } k \neq j. \tag{4}$$

In other words, according to Bayes classifier, observation x gets the label $\hat{Y} = j$ if the probability of the true label Y being equal to j, given the value x, is the maximum.

Notice that, derivation of the Bayes classifier requires knowledge of the *class probabilities* $\pi_1, \ldots, K$, and the *class-specific probability density functions* $f_1(x), \ldots, f_K(x)$. In practice, neither the former nor the latter may be known. Typically, one can estimate π_j's from the training data by assuming that the training data constitutes a random sample from the population. However, estimation of f_j's is more challenging. Often, one assumes certain models for these probability densities. This is the approach taken *Linear Discriminant Analysis (LDA)*. In contrast, one may either try to model the conditional probabilities $\mathbb{P}(Y = j|X = x)$ directly, as is done for *Logistic Regression*, or one may try to estimate these probabilities in a model-free or nonparametric way, as is done for the *k-Nearest Neighbor (kNN) classifier*.

In the following sections, we briefly describe the key features of these three classification methods, assuming for simplicity that X is one-dimensional, and primarily focusing on the case $K = 2$ (also known as binary classification problem).

2 Linear Discriminant Analysis (LDA)

This approach to classification, which results in *linear decision boundaries*, i.e., the function of X separating the predicted classes are linear, explicitly models the distribution of X within each class or subpopulation. The key assumption is that the distribution of X for each of the K subpopulations is a normal distribution, and that these distributions have different mean but the same variance. Symbolically, the class-specific probability density function $f_j(x)$ is the density function of $N(\mu_j, \sigma^2)$ distribution, where $\mu_1, \ldots, \mu_K$ are all different, and $\sigma > 0$. In that case, by using the formula (2), it can be derived that the *log-odds ratio* of X belonging to class j versus

class k, for any $j \neq k$, is

$$\begin{aligned}
\log \frac{\mathbb{P}(Y=j|X=x)}{\mathbb{P}(Y=k|X=x)} &= \log \frac{\pi_j \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_j)^2\right)}{\pi_k \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)} \\
&= \log\left(\frac{\pi_j}{\pi_k}\right) - \frac{1}{2\sigma^2}\left((x-\mu_j)^2 - (x-\mu_k)^2\right) \\
&= \left(x\frac{\mu_j}{\sigma^2} - \frac{\mu_j^2}{2\sigma^2} + \log \pi_j\right) - \left(x\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log \pi_k\right) \\
&= \delta_j(x) - \delta_k(x). \qquad (5)
\end{aligned}$$

The function $\delta_j(x) = x\frac{\mu_j}{\sigma^2} - \frac{\mu_j^2}{2\sigma^2} + \log \pi_j$, is referred to as the *discriminant function* for class j. Notice that, due to the representation (5),

$$\begin{aligned}
& & \mathbb{P}(Y=j|X=x) &> \mathbb{P}(Y=k|X=x) \\
& \text{if and only if,} & \frac{\mathbb{P}(Y=j|X=x)}{\mathbb{P}(Y=k|X=x)} &> 1 \\
& \text{if and only if,} & \log \frac{\mathbb{P}(Y=j|X=x)}{\mathbb{P}(Y=k|X=x)} &> 0 \qquad (6) \\
& \text{if and only if,} & \delta_j(x) &> \delta_k(x). \qquad (7)
\end{aligned}$$

Therefore, the optimal classification rule, i.e., the Bayes classifier, reduces to the following rule, which is known as the classification rule for LDA:

> Predicted class label $\hat{Y}$ for observation x is the value j that maximizes the value of the discriminant function $\delta_j(x)$.

The "linear" in LDA arises from the fact that $\delta_j(x)$ is a linear function of x. Since LDA is derived from the Bayes classifier, it is indeed the Bayes classifier when the subpopulations are normal having means μ_j's and common variance σ^2.

- **Decision boundary:** Since decision making in LDA simply involves comparing the values of the discriminant function $\delta_j(x)$, the value of x for which $\delta_j(x) = \delta_k(x)$, for any pair of classes $j \neq k$, constitutes the decision boundary between the classes j and k. Denoting x^*_{jk} to be the solution of the equation $\delta_j(x) = \delta_k(x)$, this means that, when deciding between classes j and k, we need to decide whether x is greater than or smaller than x^*_{jk}, assuming that $\mu_j^2 > \mu_k^2$. When $\pi_j = \pi_k$ (the class probabilities are equal), it is easily seen that

$$x^*_{jk} = \frac{\mu_j^2 - \mu_k^2}{2(\mu_j - \mu_k)} = \frac{\mu_j + \mu_k}{2} \ .$$

- **Estimated discriminant function and decision boundary:** In practice, the class probabilities $\pi_1, \ldots, \pi_K$ and the means $\mu_1, \ldots, \mu_K$ and variance σ^2 are unknown. We therefore estimate them from the observed (training) data. Assume that the number of observations belonging to class j is n_j, for $j = 1, \ldots, K$, so that $n = \sum_{k=1}^K n_k$. Then the following estimates

are used in LDA:

$$\begin{aligned}
\hat{\pi}_j &= \frac{n_j}{n}, \qquad j = 1, \ldots, K \\
\hat{\mu}_j &= \frac{1}{n_j} \sum_{i:Y_i=j} X_i, \qquad j = 1, \ldots, K \\
\hat{\sigma}^2 &= \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:Y_i=k} (X_i - \hat{\mu}_k)^2.
\end{aligned}$$

Note that $\hat{\mu}_j$ is simply the sample mean of X restricted to the j-th class, and $\hat{\sigma}^2$ is the *pooled sample variance*. Replacing π_j, μ_j and σ^2 by $\hat{\pi}_j$, $\hat{\mu}_j$ and $\hat{\sigma}^2$, respectively, in the expression for the discriminant function $\delta_j(x)$, we obtain the estimated discriminant function $\hat{\delta}_j(x)$. Correspondingly, the estimated class boundary between classes j and k (assuming that $n_j = n_k$, that is both classes have same number of measurements in the training data) becomes $\hat{x}^*_{jk} = (\hat{\mu}_j + \hat{\mu}_k)/2$.

3 Logistic Regression

Logistic regression is a classification procedure that explicitly models the log-odds ratio of an observation belonging to a certain class versus another class as a linear function of the observation. Specifically, this approach assumes that

$$\log \frac{\mathbb{P}(Y = j|X = x)}{\mathbb{P}(Y = K|X = x)} = \alpha_{0j} + \alpha_{1j} x, \qquad j = 1, \ldots, K, \tag{8}$$

for unknown constants $(\alpha_{0j}, \alpha_{1j})$, with $\alpha_{0K} = \alpha_{1K} = 0$. This means that for any $j \neq k$,

$$\log \frac{\mathbb{P}(Y = j|X = x)}{\mathbb{P}(Y = k|X = x)} = (\alpha_{0j} - \alpha_{0k}) + (\alpha_{1j} - \alpha_{1k})x,$$

which is also a linear function of x.

For simplicity, for the rest of this subsection, we only focus on the binary classification problem, i.e., when $K = 2$. In this case, we simply rename the parameters as $\beta_0 = \alpha_{01}$ and $\beta_1 = \alpha_{11}$. Also, we define $p(x) = \mathbb{P}(Y = 1|X = x)$ (conditional probability of being in class 1, when the true observation is x). Then, since $\mathbb{P}(Y = 1|X = x) = 1 - p(x)$, equation (8) simplifies to

$$\log \frac{\mathbb{P}(Y = 1|X = x)}{\mathbb{P}(Y = 2|X = x)} == \log \left(\frac{p(x)}{1 - p(x)} \right) = \beta_0 + \beta_1 x. \tag{9}$$

From (9), we deduce that the conditional probability $p(x)$ can be expressed as

$$p(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} \tag{10}$$

The function $e^a/(1 + e^a)$ is known as the logistic function (of its argument a), which gives rise to the name logistic regression.

- **Estimation:** Given (training) data $\{(X_i, Y_i) : i = 1, \ldots, n\}$, for a binary ($K = 2$ classes) classification problem, the estimation problem in logistic regression involves finding appropriate estimates for the parameters (β_0, β_1). Specifically, the model assumes that the conditional distribution of the indicator function $I(Y = 1)$ given $X = x$ is a Binomial$(1, p(x))$ distribution with success probability $p(x)$ being given by (10). This, coupled with the assumption that the observations $\{(X_i, Y_i) : i = 1, \ldots, n\}$ are independent, allows us to write down the *likelihood function* for the parameters (β_0, β_1). Then the estimates $(\hat{\beta}_0, \hat{\beta}_1)$ are obtained by the *maximum likelihood procedure*, whose implementation requires nonlinear optimization, and can only be done numerically [see Section 4.3.2 of James, Witten, Hastie and Tibshirani (2013) for further details].

- **Predicting class label:** Once we obtain the estimates $(\hat{\beta}_0, \hat{\beta}_1)$, the classification problem (referred to as prediction problem in logistic regression) involves predicting the class label for observation x. From the description (9), making use of the Bayes classifier rule again, the predicted class label $\hat{Y}$ is 1 or 2 depending on whether $\hat{\beta}_0 + \hat{\beta}_1 x$ is positive (corresponds to $\hat{p}(x) > 0.5$) or negative (corresponds to $\hat{p}(x) < 0.5$). Here $\hat{p}(x)$ is the estimated conditional probability for class 1, expressed as

$$\hat{p}(x) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x}} \, .$$

- **Logistic Regression vs. LDA:** There is a connection between logistic regression and LDA. Notice that, if the assumptions behind the LDA approach hold, then the logistic model (8) is satisfied, subject to some reparametrization. In particular, both methods result in linear decision boundaries. However, the logistic regression model simply models the conditional probability of the observation belonging to certain class, given its value, and does not require making any distributional assumption on the subpopulations. In particular, one may consider using the logistic regression model for classification even when X is discrete valued, in which case the normality assumption clearly does not hold. Nevertheless, in practice one can also use LDA simply as a classification rule, without bothering about the validity of the underlying distributional assumptions.

4 k-Nearest Neighbor (kNN) classifier

The idea of the k-Nearest Neighbor (kNN) classifier is to first estimate the conditional probabilities $\mathbb{P}(Y = j | X = x)$ in a nonparametric or model-free manner, and then using the Bayes classification rule based on these estimated probabilities.

- **Estimation of conditional probabilities:** Given a fixed integer value of $k \geq 1$, find the set of k nearest neighbors of x within the training data. That is, find the indexes $i_1, \ldots, i_k$ (depending on x) such that values of $|X_{i_1} - x|, \ldots, |X_{i_k} - x|$ are smaller that the distances $|X_i - x|$ for any $i \in \{1, \ldots, n\} \setminus \{i_1, \ldots, i_k\}$. Let $\mathcal{N}_k(x)$ denote the set $\{i_1, \ldots, i_k\}$. Then

$$\widehat{\mathbf{P}}(Y = j | X = x) = \frac{1}{k} \sum_{i : i \in \mathcal{N}_k(x)} I(Y_i = j), \qquad j = 1, \ldots, K. \tag{11}$$

  In other words, for each $j = 1, \ldots, K$, $\widehat{\mathbf{P}}(Y = j | X = x)$ is simply the fraction of observations within the k nearest neighbors of x that belong to class j.

- **Predicting class label:** The predicted value $\hat{Y}$ is the value of j for which $\widehat{\mathbf{P}}(Y = j|X = x)$ is the maximum.

- **Choice of k:** The parameter k controls the quality of the classifier. A small value of k leads very rugged classification boundaries, even though this may be close to the decision boundary of the optimal classifier (Bayes classifier). In other words, for small k, the kNN classifier has high variability. A large value of k leads to much smoother classification boundaries, that are close to linear, but may not correspond to the optimal decision boundary (of the Bayes classifier). Thus, for larger k, the classifier has large bias. This phenomenon is known as a *Bias-Variance trade-off*. In practice, we typically use the method of *cross validation* to choose the best value of k for a given data set.

Reference

1. James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer.