Statistical Methods for Machine Learning

Derek Li

Contents

1 Introduction

1.1 Terminology

$X = (X_1, \cdots, X_p)$ are p predictor variables and suppose there are n observations, $X_p = (X_{p1}, \cdots, X_{pn})$. Output varible typically denoted by Y.

1.2 Supervised and Unsupervised

Supervised statistical learning involves building a statistical model for predicting or estimating an output based on one or more inputs. For unsupervised statistical learning, there are inputs but no supervising output.

Alternatively we could say for supervised learning, all data is labeled and the algorithms learn to predict the output from the input data; for unsupervised learning, all data is unlabeled and the algorithms learn to inherent structure from the input data.

1.2.1 Supervised Learning

In the regression problem, Y is quantitative, but Xs could be quantitative or qualitative. The goal is to predict a new observation not in the training data set - what is the expected value of y_0, given x_0.

In the classification problem, Y takes values in a finite, unordered set (e.g., binary or dichotomous, categorical), but Xs could be quantitative or qualitative.The goal is to predict a new observation not in the training data set - what is the class probability of y_0, given x_0.

1.2.2 Unsupervised Learning

$X = (X_1, \cdots, X_p)$ are measured variables on n observations and usually p is large. Note that there is no response (Y) to supervise the algorithm.

Two common methods for unsupervised learning:

- Clustering: Understand the relationships between the ***observations.*** Cluster the observations on the basis of the variables measured and identify the group to which each observation belongs (number of groups is unknown).
- Principal components analysis (PCA): Understand the relationships between the ***variables*** and reduce the number of variables to a smaller number.

1.3 Simple Model

Suppose that we observe a quantitative response Y and p different predictors, $X_1, \cdots, X_p$. Assume a general form of a relationship by a model

$$Y = f(X) + \varepsilon,$$

where f is fixed and unknown and ε is a random error term, which is independent of X and has mean zero. ε captures measurement errors and other discrepancies such as omitted predictors.

There are two main reasons that we estimate f : prediction and inference. With a more accurate f we can make predictions of Y at new points $X = x$, understand which components of $X = (X_1, X_2, \cdots, X_p)$ are important in explaining Y, and depending on the complexity of f, understand how each component X_i of X affects Y.

1.3.1 Prediction

We can predict Y using $\widehat{Y} = \widehat{f}(X)$. The accuracy of $\widehat{Y}$ depends on:

- Reducible error: an error introduced by because of $\widehat{f}$ not being the perfect estimate of f but by using the most appropriate algorithm, this error can be reduced.
- Irreducible error: Y is a function of ε that cannot be predicted using X, and we cannot reduce the error introduced by ε.

Theorem 1.1. Consider a given estimate $\widehat{f}$ and a set of predictors X that yields the prediction $\widehat{Y} = \widehat{f}(X)$. Assume $\widehat{f}$ and X are fixed. Then

$$\mathbb{E}\left[\left(Y - \widehat{Y}\right)^2\right] = \underbrace{\left[f(X) - \widehat{f}(X)\right]^2}_{\text{Reducible}} + \underbrace{\text{Var}[\varepsilon]}_{\text{Irreducible}}.$$

Proof. We have

$$\begin{aligned}\mathbb{E}\left[\left(Y - \widehat{Y}\right)^2\right] &= \mathbb{E}\left[\left(f(X) + \varepsilon - \widehat{f}(X)\right)^2\right] \\ &= \mathbb{E}\left[\left(f(X) - \widehat{f}(X)\right)^2\right] + \mathbb{E}[\varepsilon^2] + \mathbb{E}\left[2\varepsilon\left(f(X) - \widehat{f}(X)\right)\right].\end{aligned}$$

Since $\mathbb{E}[\varepsilon] = 0$ and $\varepsilon \perp X$, we have

$$\mathbb{E}\left[2\varepsilon\left(f(X) - \widehat{f}(X)\right)\right] = 0.$$

Besides,

$$\mathbb{E}[\varepsilon^2] = \mathbb{E}[\varepsilon^2] - (\mathbb{E}[\varepsilon])^2 = \text{Var}[\varepsilon].$$

Therefore,

$$\mathbb{E}\left[\left(Y - \widehat{Y}\right)^2\right] = \left[f(X) - \widehat{f}(X)\right]^2 + \text{Var}[\varepsilon].$$

□

1.3.2 Inference

Inference is the goal of classical statistical methods: we want to understanding the relationship between Y and Xs and how strong is the relationship, etc.

1.4 Estimation Methods

We can use linear and non-linear approaches to estimate f. For these methods, we observe a set of n different data points (training data) and train the specific method to find an estimate $\widehat{f}$ for f.

1.4.1 Parametric Methods

Parametric methods involve a two-step model-based approach.

- Assume the functional form of f.
- Use the training data to train the model.

The problem of estimating f reduces to one of estimating a set of parameters without fitting an entirely arbitrary function f so that we call this method parametric.

Assuming a parametric form for f simplifies the problem of estimating f because it is generally much easier to estimate a set of parameters (in the linear model).

Nevertheless, the model we choose will usually not match the true unknown form of f, so that our estimate will be poor. We can try flexible model but it requires estimating a greater number of parameters, which is more complex and can lead to overfitting (following the errors or noise too closely).

1.4.2 Non-Parametric Methods

Non-parametric methods do not make explicit assumptions about the functional form of f, but seek an estimate of f that gets as close to the data points as possible.

These methods can fit a wider range of possible shapes for f but since these methods do not reduce the problem of estimating f to a small number of parameters, a very large number of observations is required in order to obtain an accurate estimate for f.

1.5 Flexibility and Interpretability

Restrictive models (e.g., linear regression) are more interpretable, but flexible models (e.g., splines, boosting) are less interpretable and they can have complicated estimates of f.

Generally, as flexibility increases, interpretability decreases. Also, as flexibility increases, prediction accuracy may decrease due to over-fitting.

1.6 Assessing Model Accuracy

1.6.1 Measuring the Quality of Fit

In the regression setting, the most commonly-used measure is the mean squared error (MSE):

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{f}(\mathbf{x}_i))^2,$$

where $\widehat{f}(\mathbf{x}_i)$ is the prediction that $\widehat{f}$ gives for the ith observation.

Since the MSE above is computed using the training data that was used to fit the model, so we call it training MSE. Nevertheless, we are interested in the accuracy of the predictions that we obtain when we apply our model to previously unseen test data not used to train the model.

Suppose we have new observations $(\mathbf{x}_0, y_0)$.We want to choose the model that gives the lowest test MSE

$$\text{Ave}(y_0 - \widehat{f}(\mathbf{x}_0))^2,$$

as opposed to the lowest training MSE.

Recall that $\text{Var}[\varepsilon]$ is the irreducible error, and it corresponds to the lowest achievable test MSE among all possible methods.

Here is a fundamental property of statistical learning that holds regardless of the data set and the model being used.

Property 1.1. As the flexibility of the model increases, there is a monotone decrease in the training MSE and a U-shape in the test MSE.

Definition 1.1 (Overfitting)**.** When a given model yields a small training MSE but a large test MSE, we are said to be overfitting the data.

Overfitting happens because the model may pick up some patterns that are just caused by random chance rather than by true properties of the unknown function f. The test MSE will be large because the patterns do not exist in the test data.

Note that we almost always expect the training MSE to be smaller than the test MSE since most models seek to minimize the training MSE.

In practice, test observations usually are not available, we still cannot select a model that minimizes the training MSE, because there is no guarantee that the method with the lowest training MSE will also have the lowest test MSE. To estimate test MSE, we may use cross-validation.

1.6.2 The Bias-Variance Trade-Off

Suppose we have fit a model $\widehat{f}(x)$ to some training data and let (x_0, y_0) be a test observation drawn from the population. Suppose the true model is $Y = f(X) + \varepsilon$ and $f(x) = \mathbb{E}[Y|X = x]$, then for given x_0, test MSE is

$$\mathbb{E}\left[\left(y_0 - \widehat{f}(x_0)\right)^2\right] = \text{Var}\left[\widehat{f}(x_0)\right] + \left[\text{Bias}\left(\widehat{f}(x_0)\right)\right]^2 + \text{Var}[\varepsilon].$$

Proof. We have

$$\begin{aligned}
\mathbb{E}\left[\left(y_0 - \widehat{f}(x_0)\right)^2\right] &= \mathbb{E}\left[\left(y_0 - \mathbb{E}\left[\widehat{f}(x_0)\right] + \mathbb{E}\left[\widehat{f}(x_0)\right] - \widehat{f}(x_0)\right)^2\right] \\
&= \mathbb{E}\left[\left(\mathbb{E}\left[\widehat{f}(x_0)\right] - y_0\right)^2\right] + \mathbb{E}\left[\left(\widehat{f}(x_0) - \mathbb{E}\left[\widehat{f}(x_0)\right]\right)^2\right] \\
&= \mathbb{E}\left[\left(\mathbb{E}\left[\widehat{f}(x_0)\right] - f(x_0) - \varepsilon\right)^2\right] + \mathrm{Var}\left[\widehat{f}(x_0)\right] \\
&= \mathbb{E}\left[\left(\mathbb{E}\left[\widehat{f}(x_0)\right] - f(x_0)\right)^2\right] + \mathbb{E}[\varepsilon^2] - (\mathbb{E}[\varepsilon])^2 + \mathrm{Var}\left[\widehat{f}(x_0)\right] \\
&= \left[\mathrm{Bias}\left(\widehat{f}(x_0)\right)\right]^2 + \mathrm{Var}[\varepsilon] + \mathrm{Var}\left[\widehat{f}(x_0)\right].
\end{aligned}$$

□

Here are some comments:

- Expected test MSE can never lie below $\mathrm{Var}(\varepsilon)$, the irreducible error.
- To minimize the expected test error, we need to select a model that simultaneously achieves low variance and low bias.
- Variance refers to the amount by which $\widehat{f}$ would change if we estimated it using a different training data set. Ideally, the estimate for f should not vary too much between training sets. More flexible models have higher variance.
- Bias refers to the error introduced by approximating a real life problem, which may be extremely complicated, by a much simpler model. Generally, more flexible models result in less bias.
- As flexibility increases, the variance will increase and the bias will decrease. The relative rate of change in variance and bias determines whether the test MSE increases or decreases.
- As we increase the flexibility, the bias tends to initially decrease faster than the variance increases so that the expected test MSE declines. At some point increasing flexibility has little impact on the bias but starts to significantly increase the variance so that the test MSE increases.

2 Linear Regression

Linear regression is a simple model to supervised learning. It assumes that the dependence of Y on $X_1, \cdots, X_p$ is linear.

2.1 Simple Linear Regression

We assume a model

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

where β_0 and β_1 are two unknown parameters represents the intercept and slope respectively, and ε is the error term. ε is to catch what we miss with the model. Assume ε is independent of X and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

We have

$$\mathbb{E}[Y|X] = \mathbb{E}[\beta_0 + \beta_1 X + \varepsilon|X] = \beta_0 + \beta_1 + X$$

and

$$\mathrm{Var}[Y|X] = \mathrm{Var}[\beta_0 + \beta_1 X + \varepsilon|X] = \mathrm{Var}[\varepsilon|X] = \sigma^2.$$

Given some estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ for the model coefficients, we can predict Y given X by

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x,$$

where $\widehat{y}$ indicates a prediction of Y given $X = x$.

2.1.1 Interpretation

β_0 is the intercept term, i.e., the expected value of Y when $X = 0$. β_1 is the slope, i.e., the average increase in Y associated with a one-unit increase in X.

2.1.2 Estimation of the Parameters by Least Squares

Let $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$ and $e_i = y_i - \widehat{y}_i$, where $i = 1, \cdots, n$, and n is the sample size and e_i is the ith residual of observation i.

We define residual sum of squares

$$\mathrm{RSS} = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2,$$

and by the least squares method, we choose $\widehat{\beta}_0$ and $\widehat{\beta}_1$ to minimize RSS:

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x},$$

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}.$$

We can prove that $\mathbb{E}\left[\widehat{\beta}_0\right] = \beta_0$ and $\mathbb{E}\left[\widehat{\beta}_1\right] = \beta_1$, i.e., $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are unbiased estimators for β_0 and β_1 respectively.

The unbiased means that if we average the values of $\widehat{\beta}_0$ or $\widehat{\beta}_1$ obtained over a huge number of data sets, the average would exactly equal to β_0 or β_1 and an unbiased estimator does not systematically over or under estimate the true parameter.

2.1.3 Assessing the Accuracy of the Coefficient Estimates

A standard error of a statistic is the estimated standard deviation of the statistic and it reflects how it varies under repeated sampling:

$$\mathrm{SE}\left(\widehat{\beta}_0\right)^2 = \sigma^2\left[\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right], \mathrm{SE}\left(\widehat{\beta}_1\right)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2},$$

where $\sigma^2 = \mathrm{Var}[\varepsilon]$, provided ε_i are independent of each other.

When the standard error increases, i.e., values of the estimators are more spread out, gives an inaccurate representation of the true population parameters.

Here are some comments for SE:

- When the x_i are more spread out, $\mathrm{SE}\left(\widehat{\beta}_1\right)$ is smaller.
- When $\overline{x} = 0, \mathrm{SE}\left(\widehat{\beta}_0\right)^2 = \mathrm{Var}\left[\widehat{\beta}_0\right] = \frac{\sigma^2}{n}$.
- $\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1\overline{x}$ and when $\overline{x} = 0, \widehat{\beta}_0 = \overline{y}, \mathrm{Var}\left[\overline{y}\right] = \frac{\sigma^2}{n}$.

SE can be used to compute confidence intervals. A 95% confidence interval is defined as a range of values s.t. with 95% probability, the range will contain the true unknown value of the parameter. For linear regression, the 95% confidence interval for $\beta_i, i = 0$ or 1, is

$$\left[\widehat{\beta}_i - 1.96 \cdot \mathrm{SE}\left(\widehat{\beta}_i\right), \widehat{\beta}_i + 1.96 \cdot \mathrm{SE}\left(\widehat{\beta}_i\right)\right],$$

i.e., there is approximately a 95% chance that the interval will contain the true value of β_i.

SE can be used to perform hypothesis tests on the coefficients. The most common hypothesis test involves testing if there is some relationship between X and Y:

$$H_0 : \beta_1 = 0$$

versus

$$H_A : \beta_1 \neq 0.$$

To test the null hypothesis, we compute a t-statistic given by

$$t = \frac{\widehat{\beta}_1 - 0}{\mathrm{SE}(\widehat{\beta}_1)},$$

which will have a t distribution with $n-2$ degrees of freedom, assuming $\beta_1 = 0$. We can compute the probability of observing any value equal to $|t|$ or larger under H_0 and we call the probability the p-value.

2.1.4 Assessing the Accuracy of the Model

The residual standard error (RSE) is an estimate of σ, the standard deviation of ε, given by

$$\text{RSE} = \sqrt{\frac{1}{n-2}\text{RSS}}.$$

RSE is an absolute measure of the lack of fit of the model to the data and is measured in the units of Y. RSE $= a$ means actual Y deviate from the true regression line by a units on average. It is not always clear what constitutes a good RSE. The R^2 statistic provides an alternative measure of fit:

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}},$$

where $\text{TSS} = \sum_{i=1}^{n}(y_i - \overline{y})^2$ is the total sum of squares.

R^2 statistic provides an alternative measure of fit and it measures the proportion of variability in Y that can be explained using X. $R^2 = b, b \in [0,1]$ indicates that $b\%$ of the variability in the response is explained by the regression. $R^2 \approx 0$ occurs when the linear model is wrong, or the inherent error σ^2 is high, or both.

In simple linear regression, $R^2 = r^2$, where r is the correlation between X and Y :

$$r = \text{Cor}(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}.$$

In SLR, r^2 can be used to assess the fit.

2.2 Multiple Linear Regression

We assume a model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p + X_p + \varepsilon,$$

where β_j is the average effect on Y of a one unit increase in X_j, holding all other predictors fixed. But predictors usually change together. Also note that claims of causality should be avoided for observational data.

2.2.1 Estimation of Parameters

Given estimates $\widehat{\beta}_0, \widehat{\beta}_1, \cdots, \widehat{\beta}_p$, we can make predictions by

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \cdots + \widehat{\beta}_p x_p.$$

We estimate β_i as the values

$$\min \text{RSS} = \min \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \min \sum_{i=1}^{n} \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \cdots - \widehat{\beta}_p x_{ip}\right)^2.$$

2.2.2 Testing the Relationship

We test

$$H_0 : \beta_1 = \cdots = \beta_p = 0$$

against

$$H_A : \text{at least one } \beta_j \text{ is non} - \text{zero.}$$

We use F-statistic

$$F = \frac{(\text{TSS} - \text{RSS})/p}{\text{RSS}/(n-p-1)} \sim F_{p(n-p-1)}.$$

2.2.3 Deciding on Important Variables

The most direct method is called all subsets or best subsets regression: we compute the least squares fit for all possible subsets and then choose between them based on some criterion that balances training error with model size. There are 2^p models and thus if p is large, we need an automated approach.

- Forward selection. Begin with the null model that contains an intercept but no predictors. Fit p SLRs and add to the null model the variable that results in the lowest RSS. Add to that model the variable that results in the lowest RSS for the new two-variable model. This approach is continued until some stopping rule is satisfied.
- Backward selection. Start with all variables in the model, and remove the variable with the largest p-value, i.e., the variable is the least statistically significant. The new $(p-1)$-variable model is fit and remove the variable with the largest p-value. This procedure continues until a stopping rule is reached. For example, we may stop when all remaining variables have a significant p-value defined by some significance threshold.

2.2.4 Qualitative Predictors

Example 2.1. For the ethnicity (Caucasian, African American or Asian) variable, we create two dummy variables. The first could be

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{th person is Asian} \\ 0 & \text{if } i\text{th person is not Asian} \end{cases},$$

and the second could be

$$x_{i2} = \begin{cases} 1 & \text{if } i\text{th person is Caucasian} \\ 0 & \text{if } i\text{th person is not Caucasian} \end{cases}.$$

We have

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{if } i\text{th person is Asian} \\ \beta_0 + \beta_2 + \varepsilon_i & \text{if } i\text{th person is Caucasian} \\ \beta_0 + \varepsilon_i & \text{if } i\text{th person is AA} \end{cases}.$$

The level with no dummy variable, AA in this example, is known as the baseline.

2.2.5 Extensions of the Linear Model

In practice, there is a synergy effect and in statistics it is referred to as an interaction effect and thus we will introduce interaction term in the model.

Sometimes it is the case that an interaction term has a very small p-value, but the associated main effects do not.

We follow the hierarchy principle: If we include an interaction in a model, we should also include the main effects, even if the p-values associated with their coefficients are not significant. The rationale for the principle is that interactions are hard to interpret without main effects and their meaning is changed.

Here is an example for interactions between qualitative and quantitative variables.

Example 2.2. Without an interaction term, the model takes the form

$$y_i = \beta_0 + \beta_1 x_i + \begin{cases} \beta_2 \\ 0 \end{cases} = \beta_1 x_i + \begin{cases} \beta_0 + \beta_2 \\ \beta_0 \end{cases} .$$

With an interaction, we have

$$y_i = \beta_0 + \beta_1 x_i + \begin{cases} \beta_2 + \beta_3 x_i \\ 0 \end{cases} = \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_i \\ \beta_0 + \beta_1 x_i \end{cases} .$$

2.2.6 Non-Linear Relationships

We can extend the linear model to accommodate non-linear relationships, using polynomial regression.

3 Classification

In classification model, the response variable Y is qualitative. Our goals are to

- Build a classifier $C(X)$ that assigns a class label from C to a future unlabeled observation or assess the uncertainty in each classification.
- Understand the roles of the different predictors among $X = (X_1, \cdots, X_p)$.

3.1 Bayes Optimal Classifier

Suppose the M labels in C are numbered $1, \cdots, M$. Let

$$p_m(\mathbf{x}) = P(Y = m|X = \mathbf{x}), m = 1, \cdots, M.$$

The Bayes optimal classifier at $\mathbf{x}$ is

$$C(\mathbf{x}) = j \text{ if } p_j(\mathbf{x}) = \max\{p_1(\mathbf{x}), \cdots, p_M(\mathbf{x})\},$$

i.e., to choose the label with highest probability and thus minimizing the probability that is makes an error. Hence, Bayes classifier (using the true $p_m(\mathbf{x})$) has smallest error.

3.1.1 Model Accuracy

We measure the performance by training error rate, which is the proportion of mistakes, given by

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \widehat{C}(\mathbf{x}_i)),$$

where I is an indicator function.

Test error rate is associated with a set of test observations of the form $(\mathbf{x}_0, y_0)$ is given by

$$\text{Test Error Rate} = \text{Ave}(I(y_0 \neq \widehat{C}(\mathbf{x}_0))).$$

A good classifier is one where the test error is smallest and actually the Bayes classifier $C(\mathbf{x})$ produces the lowest possible test error rate, called the Bayes error rate and thus we call it optimal Bayes classifier.

3.1.2 Bayes Decision Boundary

Definition 3.1 (Bayes Decision Boundary)**.** The line that represents the points where the probability is exactly 50% is called the Bayes decision boundary.

3.1.3 Limitation

In practice, we do not know the true population probabilities and thus we need to define other classifiers that can approximate $p_m(\mathbf{x})$.

3.2 K-Nearest Neighbors

To estimate the conditional distribution, we can use K-nearest neighbors (KNN) classifier.

Given a $K \in \mathbb{N}$ and a test observation $\mathbf{x}_0$, the KNN classifier first identifies the K points in the training data that are closest to $\mathbf{x}_0$, represented by $\mathcal{N}_0$. It then estimates the conditional probability

$$P(Y = j|X = \mathbf{x}_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = j).$$

Finally, KNN applies Bayes rule and classifies $\mathbf{x}_0$ to the class with te largest probability.

The flexibility is $\frac{1}{K}$.

3.3 Logistic Regression

Logistic regression uses the logistic (sigmoid-S shaped) function:

$$p(X) = P(Y = 1|X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}.$$

Wherefore,

$$\text{logit} = \ln\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X.$$

The monotone transformation is called the log odds or logit transformation of $p(X)$.

3.3.1 Estimation

We use maximum likelihood to estimate the parameters:

$$l(\beta_0, \beta) = \prod_{i:y_i=1} p(x_i) \prod_{i:y_i=0} (1 - p(x_i)),$$

where gives the joint probability of the observed zeros and ones in the data.

In R, we use the glm function.

3.3.2 Interpretation

A one-unit increase in X is associated with an increase in the log odds of Y by β_1 units. $\widehat{\beta}_0$ is typically not of interest.

3.4 Multiple Logistic Regression

We use

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}} \Rightarrow \ln\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p,$$

where $X = (X_1, \cdots, X_p)$ are p predictors. We use the maximum likelihood method to estimate β's.

Definition 3.2 (Confounding Variable)**.** A confounding variable is a factor associated with both the predictor and response.

Definition 3.3 (Interaction)**.** Interaction among variables, also known as effect modification, exists when the effect of one predictor variable on the response depends on the particular level or value of another predictor variable.

3.5 Multi-Classes Logistic Regression

We could use R package glmnet.

3.6 Discriminant Analysis

When we have more than two response classes, when the classes are well-separated (the parameter estimates for the logistic regression model are unstable), or if n is small and the distribution of the predictors X is approximately normal in each of the classes (the linear discriminant model is mode stable), we will use discriminant analysis rather than logistic regression.

Theorem 3.1 (Bayes Theorem)**.** $P(Y = k|X = \mathbf{x}) = \frac{P(X=\mathbf{x}|Y=k)\cdot P(Y=k)}{P(X=\mathbf{x})}$

Let $\pi_k = P(Y = k)$ be the marginal or prior probability that a randomly chosen observation comes from the kth class. Let $f_k(X) = P(X = \mathbf{x}|Y = k)$ be the density for X in class k, then

$$p_k(X) = P(Y = k|X = \mathbf{x}) = \frac{\pi_k f_k(\mathbf{x})}{\sum_{i=1}^{K} \pi_i f_i(\mathbf{x})}.$$

Assume predictors X are continuous variables and conditional class densities of $X = \mathbf{x}$ are multivariate normal distributions for each class k, i.e.,

$$f_k(\mathbf{x}) \sim \mathcal{N}(\mu_k, \Sigma_k), k = 1, \cdots, K.$$

We classify an observation to the class k for which $p_k(\mathbf{x})$ is largest.

3.6.1 Linear Discriminant Analysis

(1) **Case:** $p = 1$. The Gaussian density has the form

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma_k}\right)^2}.$$

Assume $\sigma_k = \sigma$, we have

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma}\right)^2}}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu_l}{\sigma}\right)^2}}.$$

We can consider the term regardless of k as a constant, and have

$$p_k(x) = C\pi_k e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma}\right)^2} \Rightarrow \ln(p_k(x)) = \ln C + \ln \pi_k - \frac{(x-\mu_k)^2}{2\sigma^2},$$

and have

$$\ln(p_k(x)) = C' + \frac{x\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \ln \pi_k.$$

So the objection function, or linear discriminant function is

$$\delta_k(x) = \ln \pi_k - \frac{\mu_k^2}{2\sigma^2} + \frac{x\mu_k}{\sigma^2}x.$$

The Bayes classifier is equivalent to assigning x to the class with the largest discriminant score $\delta_k(x)$. Notice that $\delta_k(x)$ is linear because $\sigma_k^2 = \sigma^2$ for all k.

Example 3.1. Assume $K = 2$ and $\pi_1 = \pi_2 = 0.5$, then the Bayes classifier assigns an observation to class 1 if $\delta_1(x) > \delta_2(x)$. Besides, the decision boundary is at $x = \frac{\mu_1+\mu_2}{2}$, i.e., when $\delta_1(x) = \delta_2(x)$.

In practice, we are not able to calculate the Bayes classifier and LDA approximates the Bayes classifier when we substitute estimated parameters using the training observations.

$$\widehat{\pi}_k = \frac{n_k}{n},\ \widehat{\mu}_k = \frac{1}{n_k}\sum_{i:y_i=k} x_i,\ \widehat{\sigma}^2 = \frac{1}{n-K}\sum_{k=1}^{K}\sum_{i:y_i=k}(x_i - \widehat{\mu}_k)^2 = \sum_{k=1}^{K}\frac{n_k-1}{n-K}\cdot\widehat{\sigma}_k^2,$$

where $\widehat{\sigma}_k^2 = \frac{1}{n_k-1}\sum_{i:y_i=l}(x_i - \widehat{\mu}_k)^2$. Hence,

$$\widehat{\delta}_k(x) = \ln \widehat{\pi}_k - \frac{\widehat{\mu}_k^2}{2\widehat{\sigma}^2} + \frac{\widehat{\mu}_k}{\widehat{\sigma}^2}x.$$

Note that, we need to assume the observations within each class come from a normal distribution with a class-specific mean vector and a common variance σ^2. With the test data, we can compute

$$\text{Bayes Test Error Rate} = \text{Ave}(I(y_0 \neq C(x_0))),$$

and

$$\text{LDA Test Error Rate} = \text{Ave}(I(y_0 \neq \widehat{C}(x_0))).$$

(2) **Case:** $p > 1$. We assume $X = (X_1, \cdots, X_p)$ is multivariate Gaussian distribution with a class-specific mean vector and a common covariance matrix, so that each individual predictor follows a one-dimensional normal distribution. We write

$$X \sim \mathcal{N}(\mu, \Sigma),$$

where μ is a $p \times 1$ mean vector and Σ is a $p \times p$ covariance matrix. The multivariate normal density function of X is

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}e^{-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)}.$$

Similarly, we have the discriminant function

$$\delta_k(\mathbf{x}) = \mathbf{x}^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \ln \pi_k = c_{k0} + c_{k1}x_1 + \cdots + c_{kp}x_p,$$

provided Σ is common to all K classes and $\mathbf{x} = (x_1, \cdots, x_p)$. Bayes classifier assigns an observation $X = \mathbf{x}$ to the class for which $\delta_k(\mathbf{x})$ is largest.

3.6.2 Confusion Matrix

Suppose we have a binary classifier $C(\mathbf{x})$ and $C(\mathbf{x})$ can make two types of errors: $C(\mathbf{x}) = G_1$ when in fact $\mathbf{x}$ is from G_2 and $C(\mathbf{x}) = G_2$ when in fact $\mathbf{x}$ is from G_1. A confusion matrix can display the error information.

Notice that errors that are very lopsided could be a problem. Hence we also consider:

- False positive rate: The fraction of negative example that are classified as positive.
- False negative rate: The fraction of positive example that are classified as negative.
- Sensitivity/True positive rate: The fraction of positive example that are classified as positive.
- Specificity/True negative rate: The fraction of negative example that are classified as negative.

We may modify the threshold value but there is a trade-off between these rates.

3.6.3 ROC and AUC

ROC curve, an acronym for receiver operating characteristics, is for simultaneously displaying the two types of errors for all possible thresholds.

The overall performance of a classifier, summarized over all possible thresholds, is given by the area under the ROC curve, called AUC. Higher AUC is better.

3.6.4 Quadratic Discriminant Analysis

With Gaussian conditional class probabilities but different Σ_k in each class, we get quadratic discriminant analysis (QDA). The discriminant function is

$$\delta_k(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \mu_k)^T \Sigma_k^{-1}(\mathbf{x} - \mu_k) + \ln \pi_k.$$