Classification

December 29, 2024

Table of Contents

Introduction

Classification - Introduction

- Linear regression: Y is quantitative.
- But: very often Y can be qualitative (often called categorical).
- Qualitative variables take values in an unordered set C, such as:
  `eye color` $\in$ *brown*, *blue*, *green*
  `email` $\in$ *spam*, *ham*.
- Given a feature vector X and a qualitative response Y taking, the classification task is to build a function that takes as input the feature vector X and predicts its value for Y.
- For that classification methods often estimate the probabilities that X belongs to each category in C for the basis of classification.
- We are now looking at approaches for predicting qualitative responses, also known as classification methods.

Classification - Introduction

- Examples of Classification problems:
  1. A person arrives at the emergency room with a set of symptoms that could possibly be attributed to one of three medical conditions. Which of the three conditions does the individual have?
  2. An online banking service must be able to determine whether or not a transaction being performed on the site is fraudulent, on the basis of the user's IP address, past transaction history, and so forth.
  3. On the basis of DNA sequence data for a number of patients with and without a given disease, a biologist would like to figure out which DNA mutations are deleterious (disease-causing) and which are not.
- Similarities to the linear regression problem:
  - There is a set of training observations $(x_1, y_1), ..., (x_n, y_n)$
  - We want our classifier to perform well not only on the training data, but also on unobserved test data. (Recall lecture 1)

Classification - Introduction

- We will use the `Default` data set.
- Predict, whether an individual will default on his/her credit card balance, given annual income and monthly credit card balance.

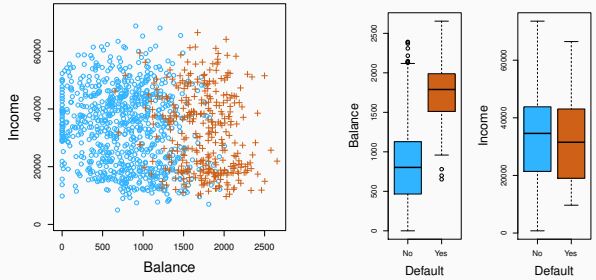


Figure 1: The Default data set. Left: The annual incomes and monthly credit card balances of a number of individuals. The individuals who defaulted on their credit card payments are shown in orange, and those who did not are shown in blue. Center: Boxplots of balance as a function of default status. Right: Boxplots of income as a function of default status.

Classification - Introduction

- Can we use Linear Regression?
- *Reason 1:*
  - Suppose for the Default classification task we code:

$$Y = \begin{cases} 0, & \text{if } \texttt{No} \\ 1, & \text{if } \texttt{Yes} \end{cases}.$$

  - Can we simply perform a linear regression of Y on X and classify as "Yes" if $\hat{Y} > 0.5$ - interpreting them as probabilities?
  - In this special case of a binary outcome, it can be shown that $X\hat{\beta}$ is in fact an estimate of $Pr(Y = \texttt{default}|X)$
  - **BUT**: Linear regression might produce probabilities less than zero or bigger than one - which makes them hard to be interpreted as probabilities (see next slide). Logistic regression is more appropriate.

Classification - Introduction

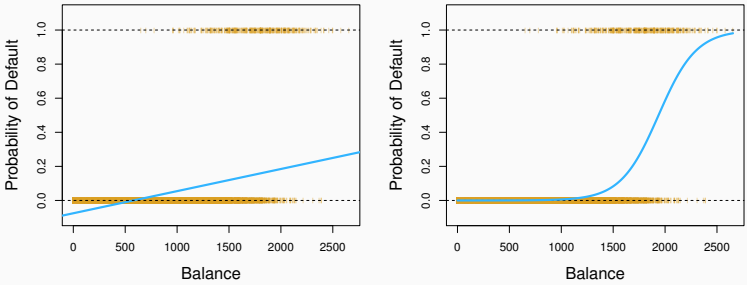


Figure 2: The orange marks indicate the response Y, either 0 or 1. Linear regression does not estimate $Pr(Y = 1|X)$ well. Logistic regression seems well suited to the task.

Classification - Introduction

- *Reason 2:*
  - Now suppose we have a response variable with three possible values. A patient presents at the emergency room, and we must classify them according to their symptoms

$$Y = \begin{cases} 1, & \text{if } \texttt{stroke}; \\ 2, & \text{if } \texttt{drug overdose}; \\ 3, & \text{if } \texttt{epilleptic seizure} \end{cases}$$

  - This coding suggests an ordering, and in fact implies that the difference between `stroke` and `drug overdose` is the same as between `drug overdose` and `epileptic seizure`.
  - We could simply reorder the encoding, which would lead to a completely different model. Linear regression is not appropriate here.
  - Only if the response has a natural ordering (mild, moderate, severe) linear regression could be used.
  - Multiclass Logistic Regression is more appropriate

Classification - Introduction

- Two reasons why linear regression is not suitable for qualitative responses:
  1. A regression method cannot accommodate a qualitative response with more than two classes
  2. a regression method will not provide meaningful estimates of $Pr(Y|X)$, even with just two classes
- In the following we will discuss the following model:
  - Logistic Regression
  - Linear Discriminant Analysis

Simple Logistic Regression

Logistic Regression

- Reconsider again the default data set.
- Instead of modelling Y directly, logistic regression models the probability that Y belongs to a particular category.

$$Pr(\texttt{default} = \texttt{Yes}|\ \texttt{balance}) \tag{1}$$

- $Pr(\texttt{default} = \texttt{Yes}|\ \texttt{balance}) \ \in \ [0, 1]$
- We might conclude that $\texttt{default} = \texttt{Yes}$ for any individual for whom $Pr(\texttt{balance}) > 0.5$ or if the company is more conservatively $Pr(\texttt{balance}) > 0.1$.

Logistic Regression

- How should we model this relationship?
  - Let's write $p(X) = Pr(Y = 1|X)$ for short and consider using `balance` to predict `default`. Logistic regression uses the logistic function:

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{(1 + e^{\beta_0+\beta_1 X})} \tag{2}$$

  - Note: ($e \approx 2.71828$ is a mathematical constant [Euler's number.])
  - It is easy to see that no matter the values for $\beta_0, \beta_1,$ or X: $p(X)$ will have values between 0 and 1 (see Figure 2, the logistic function will produce an *S-shape*)
  - A bit of rearrangement results in:

$$\left(\frac{p(X)}{(1 - p(X))}\right) = e^{\beta_0+\beta_1 X} \tag{3}$$

  - $\frac{p(X)}{(1-p(X))}$ is called the *odds* and can take any value between 0 and ∞, indicating very high and very low probabilities, respectively.
  - Odds are stated with regard to likelihoods. How likely is one event compared to another.
  - Example:
    - On average 1 in 5 people will default, implies an odds of $\frac{0.2}{1-0.2} = \frac{1}{4}$, since $p(X) = 0.2$.

Questions:

a) On average, what fraction of people with an odds of 0.37 of defaulting on their credit card payment will in fact default?

b) Suppose that an individual has a 16% chance of defaulting on her credit card payment. What are the odds that she will default?

Logistic Regression

- Taking the logarithm leads to:

$$\log\left(\frac{p(X)}{(1-p(X))}\right) = \beta_0 + \beta_1 X \tag{4}$$

- This monotone transformation is called the *log odds* or logit transformation of $p(X)$ (by log we mean natural *log* : *ln.*).
- Interpretation:
  - Linear regressions: β_1 = average change in Y, when we increase X by one unit.
  - Logistic regressions: a one unit increase in X increases the log odds by β_1. Even though (6) does not imply a straight line, given a positive β_1 and increasing X is associated with a higher $p(X)$.
  - Further: the amount by which $p(X)$ changes due to a one unit change in X depends on the current value of X (see Figure 2).
- Note: no linear relationship between $p(X)$ and X, thus β_1 does not correspond to the change in $p(X)$.

Logistic Regression

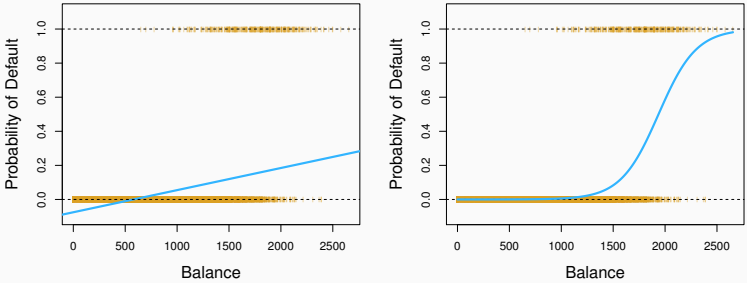


Figure 3: The orange marks indicate the response Y, either 0 or 1. Linear regression does not estimate $Pr(Y = 1|X)$ well. Logistic regression seems well suited to the task.

Simple Logistic Regression

Estimation

Logistic Regression - Estimation

- For the linear regression we used OLS, now we use *maximum likelihood* to estimate the parameters.
- That is, we pick β_0 and β_1 in such a way that the predicted probability $\hat{p}(X)$ of default for each individual corresponds as closely as possible to the individually observed default status.

$$l(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'})) \tag{5}$$

with

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{(1 + e^{\beta_0+\beta_1 X})} \tag{6}$$

- Intuitively: maximum likelihood provides us with estimates such that $p(X)$ is close to one for all individuals that defaulted and close to zero for those who did not.

Logistic Regression - Estimation

- Most statistical packages can fit linear logistic regression models by maximum likelihood.

	Coefficient	Std. Error	Z-Statistic	P-Value
`Intercept`	-10.6513	0.3612	-29.5	<0.0001
`balance`	0.0055	0.0002	24.9	<0.0001

Table 1: For the Default data, estimated coefficients of the logistic regression model that predicts the probability of default using balance. A one-unit increase in balance is associated with an increase in the log odds of default by 0.0055 units.

- Notes:
  - Hypothesis Tests and Confidence Intervals apply, just like in the linear regression framework.

Using the data from the previous slide, answer the following questions:

a) Compute an approximate 95% Confidence Interval for β_1.

b) Can you reject the null hypothesis $H_0 : \beta_1 = 0$ for $\alpha = 5\%$?

c) Interpret the estimated coefficient value for β_0.

d) Interpret the estimated coefficient value for β_1.

Simple Logistic Regression

Prediction

Logistic Regression - Prediction

- What is our estimated probability of default for someone with a balance of \$1000?

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 1000}}{1+e^{-10.6513+0.005\times 1000}} = 0.006$$

- With a balance of \$2000?

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 2000}}{1+e^{-10.6513+0.005\times 2000}} = 0.586$$

Logistic Regression - Prediction

- As with linear regression, we cann also incorporate qualitative variables into the model.
- For example, consider `student` with a 0/1 encoding as the predictor:

	Coefficient	Std. Error	Z-Statistic	P-Value
`Intercept`	-3.5041	0.0707	-49.55	$<$0.0001
`student[Yes]`	0.4049	0.1150	3.52	0.0004

$$\widehat{Pr}(\texttt{default} = \texttt{Yes}|\ \texttt{student} = \texttt{Yes}) = \frac{e^{-3.5041+0.4049\times 1}}{1+e^{-3.5041+0.4049\times 1}} = 0.0431$$

$$\widehat{Pr}(\texttt{default} = \texttt{Yes}|\ \texttt{student} = \texttt{No}) = \frac{e^{-3.5041+0.4049\times 0}}{1+e^{-3.5041+0.4049\times 0}} = 0.0292$$

Multiple Logistic Regression

Multiple Logistic Regression

- Logistic regression with several variables:

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$$

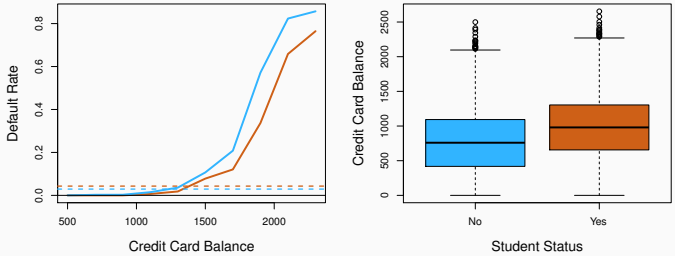
$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+...+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+...+\beta_p X_p}}$$

- Again we use Maximum Likelihood to estimate $\beta_0, \beta_1, ..., \beta_p$

	Coefficient	Std. Error	Z-Statistic	p-Value
`Intercept`	-10.8690	0.4923	-22.08	<0.0001
`balance`	0.0057	0.0002	24.74	<0.0001
`income`	0.0030	0.0082	0.37	0.7115
`student [Yes]`	-0.6468	0.2362	-2.74	0.0062

- Why is coefficient for `student` now negative, while it was positive before?

Multiple Logistic Regression



- Orange and blue lines show default rates of students and non-students as a function of credit card balance
- For a fixed value of balance and income a student is less likely to default than a non-student
- *Dotted line*: across all values of `balance` and `income`, students have an overall higher default rate than non-students. Thus, a positive coefficient in the simple model.

- Student Status and Credit Card Balance are correlated
- Students tend to have higher balances than non-students, which from the dotted line implies a higher probability of default, implying *A student is riskier than a non-student.*
- **BUT**: for each level of balance, students default less than non-students.
- Multiple logistic regression can tease this out. This phenomenon is also called *confounding.*

- Assume a student with a credit card balance of \$1500 and an income of \$40 000, what is his probability of default?

$$\hat{p}(X) = \frac{e^{-10.869+0.00574\times1500+0.003\times40-0.6468\times1}}{1+e^{-10.869+0.00574\times1500+0.003\times40-0.6468\times1}} = 0.058$$

- A non-student with the same balance and income has the following probability of default:

$$\hat{p}(X) = \frac{e^{-10.869+0.00574\times1500+0.003\times40-0.6468\times0}}{1+e^{-10.869+0.00574\times1500+0.003\times40-0.6468\times0}} = 0.105$$

Generative Models for Classification

General Models for Classification

- Logistic Regression: Directly model $Pr(Y = k|X = x)$
- Now: Model the distribution of X in each of the classes separately, and then use Bayes theorem to flip things around and obtain $Pr(Y = k|X = x)$.
- *Recall Bayes Theorem*

$$Pr(Y = k|X = x) = \frac{Pr(X = x|Y = k) \times Pr(Y = k)}{Pr(X = x)}$$

General Models for Classification - Bayes Theorem

- *Recall Bayes Theorem - continued:*
  - Suppose we have K classes and thus Y can take on K distinct and unordered possible values
  - Let π_k be the *prior* probability that a randomly chosen observation comes from the kth class
  - Let $f_k(x) \equiv Pr(X|Y = k)$ denote the *density* function of X for an observation that comes from the kth class
  - Remember: $f_k(x)$ is rel. large, if the probability that an observation in the kth class has $X \approx x$
  - Then *Bayes Theorem* states:

$$Pr(Y = k|X = x) = \frac{Pr(Y = k) \times Pr(X = x|Y = k)}{Pr(X = x)} = \frac{\pi_k \times f_k(x)}{\sum_{l=1}^{K} \pi_l \times f_l(x)},$$

  - It is the probability that the observation belongs to the kth class, given the predictor value for that observation.

General Models for Classification - Bayes Theorem

- Recall Bayes Theorem:

$$Pr(Y = k|X = x) = \frac{\pi_k \times f_k(x)}{\sum_{l=1}^{K} \pi_l \times f_l(x)},$$

- Instead of directly computing the posterior probability $p_k(x) = Pr(Y = k|X = x)$, we can simply plug in estimates of π_k and $f_k(x)$
- Estimating π_k is easy if we have a random sample from the population: we simply compute the fraction of the training observations that belong to the kth class.
- But: Estimating the density function $f_k(x)$ is much more challenging $\Rightarrow$ we need some simplifying assumptions.

General Models for Classification - Bayes Theorem

- Depending on the assumptions we are making we end up with different classifiers:
  - **Linear Discriminant Analysis (LDA)**: $f_k(x)$ is assumed to be normal/ Gaussian; Covariances between the classes are assumed equal
  - **Quadratic Discriminant Analysis (QDA)**: $f_k(x)$ is assumed to be normal/ Gaussian; Covariances between the classes are assumed not to be equal
  - **Naive Bayes**: $f_k(x)$ is unknown; Covariances between the classes are assumed not to be equal
- Our focus is on Linear Discriminant Analysis (LDA) with one ($p = 1$) and several ($p > 1$) predictors.

General Models for Classification

- Why do we need another method, when we have logistic regression?
  - *Stability.* When there is substantial separation between the two classes, the parameter estimates for the logistic regression model are surprisingly unstable. The methods that we consider in this section do not suffer from this problem.
  - *Distribution and Sample Size.* If the distribution of the predictors X is approximately normal in each of the classes and the sample size is small, then the approaches in this section may be more accurate than logistic regression.
  - *Extensions.* The methods in this section can be naturally extended to the case of more than two response classes.

Generative Models for Classification

Linear Discriminant Analysis with $p = 1$

GMC - LDA when $p = 1$

- Assume $p = 1$ - i.e. we only have one predictor.
- Goal: We want an estimate of $f_k(x)$ that we can plug into Bayes Theorem in order to estimate $p_k(x)$.
- Then classify an observation to the class k, for which $p_k(x)$ is highest.
- To estimate $f_k(x)$ we need to make an assumption about its form: Here we assume *gaussianity* or normality - that is $f_k(x)$ has a normal shape.

GMC - LDA when $p = 1$

- Recall the form of the Gaussian density:

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{1}{2}(\frac{x-\mu_k}{\sigma_k})^2} \tag{7}$$

- Here μ_k is the mean, and σ_k^2 the variance (in class k). We will assume that all the $\sigma_1^2 = \sigma_2^2 = ...\sigma_K^2$ (i.e. the variance is the same across all classes).
- Plugging this into Bayes formula, we get a rather complex expression for $p_k(x) = Pr(Y = k|X = x)$:

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu_k}{\sigma})^2}}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu_l)}{\sigma})^2}} \tag{8}$$

- Luckily, there are simplifications and cancellations.

GMC - LDA when $p = 1$

- To classify at the value $X = x$, we need to see which of the $p_k(x)$ is largest.
- Expressing equation (8) up to a proportionality constant:

$$p_k(x) \propto \pi_k \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu_k}{\sigma})^2}$$

- Taking logs, and discarding terms that do not depend on k, we see that this is equivalent to assigning x to the class with the largest discriminant score $\delta_k(x)$:

$$\begin{aligned}
\log(p_k(x)) &\propto \log(\pi_k) - \log(\sqrt{2\pi\sigma}) - \frac{1}{2\sigma^2}(x-\mu_k)^2 \\
&\propto \log(\pi_k) - \frac{1}{2\sigma^2}[x^2 - 2x\mu_k + \mu_k^2] \\
\delta_k(x) &= x \times \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)
\end{aligned}$$

- Note that $\delta_k(x)$ is a linear function of x - i.e. thats why the name **Linear Discriminant Analysis**!

GMC - LDA when $p = 1$

- For example, assume $K = 2$ and $\pi_1 = \pi_2$, then the Bayes classifier assigns an observation to class 1, if:

$$2x(\mu_1 - \mu_2) > \mu_1^2 - \mu_2^2 \tag{9}$$

  and to class 2 otherwise.

- The Bayes decision boundary is the point for which $\delta_1(x) = \delta_2(x)$. One can show that this amounts to:

$$x = \frac{\mu_1 + \mu_2}{2} \tag{10}$$

- Homework: Can you derive them yourself? (Hint: set $\delta_1(x) > \delta_2(x)$ & $\delta_1(x) = \delta_2(x)$, respectively)
- Let's look at an example...

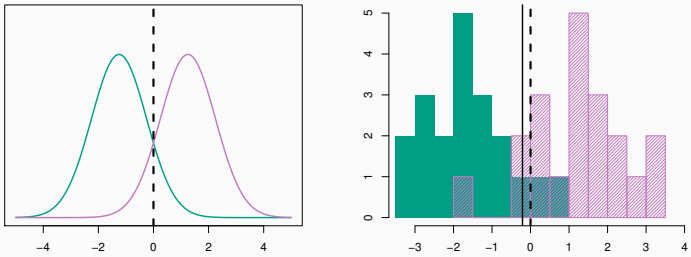


Figure 4: Left: Two one-dimensional normal density functions are shown. The dashed vertical line represents the Bayes decision boundary. Right: 20 observations were drawn from each of the two classes, and are shown as histograms. The Bayes decision boundary is again shown as a dashed vertical line. The solid vertical line represents the LDA decision boundary estimated from the training data.

- Assumption: Everything is known.
- Left: The two normal density functions that are displayed, $f_1(x)$ and $f_2(x)$, represent two distinct classes. The parameters for these classes are respectively: $\mu_1 = 1.5$, $\mu_2 = 1.5$ and $\sigma_1^2 = \sigma_2^2 = 1$.
- Typically we don't know these parameters; we just have the training data. In that case we simply estimate the parameters and plug them into (8).
- Note: There is some overlap between the normal distributions - thus given some $X = x$ values, there is some uncertainty to which class the values belong.
- Further assume that each observation is equally likely to come from either class: $\pi_1 = \pi_2 = 0.5$ - then we can see that the Bayes classifier assigns the observations to class 1 if $x < 0$ and to class 2 otherwise.
- The dashed vertical line is the Bayes decision boundary (Note: we can only compute the Bayes decision boundary as we know that X is drawn from a Gaussian distribution and we know all the parameters involved).

GMC - LDA when $p = 1$

- In practice, to apply the Bayes classifier we need to estimate the parameters: $\hat{\pi}_k$, $\hat{\mu}_k$ and $\hat{\sigma}^2$, using the training data at hand.
- Estimating the parameters:

$$\hat{\pi}_k = \frac{n_k}{n}$$

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i$$

$$\hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=1} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2$$

$$= \sum_{k=1}^{K} \frac{n_k - 1}{n-K} \hat{\sigma}_k^2$$

where $\hat{\sigma}_k^2 = \frac{1}{n_k - 1} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2$ is the usual formula for the estimated variance in the kth class.

GMC - LDA when $p = 1$

- Having obtained our estimates for $\hat{\pi}_k$, $\hat{\mu}_k$ and $\hat{\sigma}^2$, we plug them into the following formula and assign an observation $X = x$ to the class for which:

$$\delta_k(x) = x \times \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k) \tag{11}$$

  is largest.

- So classifying to the largest $\hat{\delta}_k(x)$ amounts to classifying to the class for which $\hat{Pr}(Y = k|X = x)$ is largest.
- The word linear in the classifier's name stems from the fact that the discriminant functions $\delta_k(x)$ are linear functions of x.

- Lets continue with the previous example and recall Figure 4:
  - The right-hand panel of Figure 4 displays a histogram of a random sample of 20 observations from each class.
  - Begin by estimating $\hat{\pi}_k$, $\hat{\mu}_k$ and $\hat{\sigma}^2$.
  - Then computed the decision boundary, shown as a black solid line, that results from assigning an observation to the class for which (11) is largest.
  - All points to the left of this line will be assigned to the green class, while points to the right of this line are assigned to the purple class.
  - Here, given $n_1 = n_2 = 20$, we have $\hat{\pi}_1 = \hat{\pi}_2$ and the decision boundary corresponds to the midpoint between the sample means for the two classes, $\frac{\hat{\mu}_1 - \hat{\mu}_2}{2}$.
  - We can observe that the LDA decision boundary is slightly to the left of the optimal Bayes decision boundary, which instead equals $\frac{\mu_1 - \mu_2}{2} = 0$.

Generative Models for Classification

Linear Discriminant Analysis with $p > 1$

GMC - LDA when $p > 1$

- Let's extend the LDA to the case of multiple predictors ($p > 1$)
- The multivariate Gaussian distribution assumes that each individual predictor follows a one-dimensional normal distribution, with some correlation between each pair of predictors.

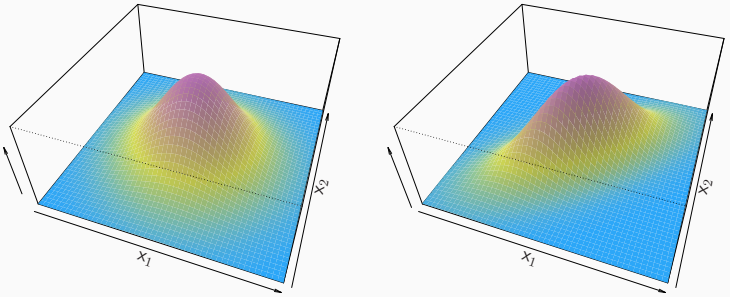


Figure 5: Two multivariate Gaussian density functions are shown, with $p = 2$. Left: The two predictors are uncorrelated. Right: The two variables have a correlation of 0.7.

GMC - LDA when $p > 1$

- To indicate that a p-dimensional random variable X has a multivariate Gaussian distribution, we write $X \sim N(\mu, \mathbf{\Sigma})$,
- $E(X) = \mu$ is the mean of X (a vector with p components)
- $Cov(X) = \mathbf{\Sigma}$ is the $p \times p$ covariance matrix of X.
- Formally the Multivariate Gaussian Density is given by:

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

- In the case of $p > 1$ predictors, the LDA classifier assumes:
  - Observations in the kth class are drawn from a multivariate Gaussian distribution $N(\mu_k, \mathbf{\Sigma})$
  - μ_k is a class-specific mean vector
  - the covariance matrix ($\mathbf{\Sigma}$) that is common to all K classes

GMC - LDA when $p > 1$

- Plugging the density function for the kth class, $f_k(X = x)$, into (11) and rearranging:

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k \tag{12}$$

- Despite its complex form $\delta_k(x)$ is a linear function.
- As before, in practice, we need to estimate $\mu_1, ..\mu_K$, $\pi_1, ..., \pi_K$ and $\mathbf{\Sigma}$.
- Let's look at an example...

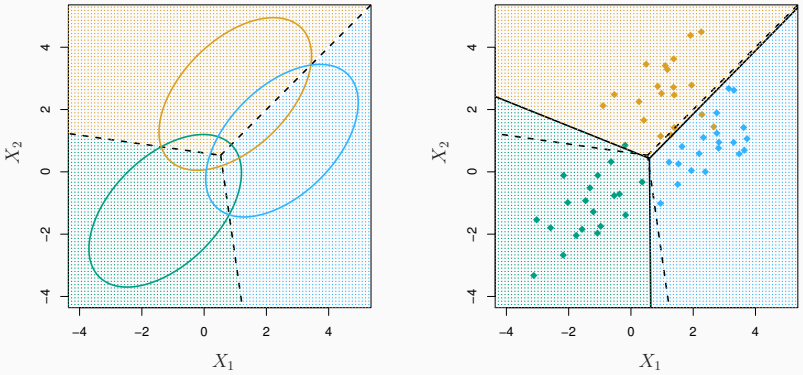


Figure 6: Here $\pi_1 = \pi_2 = \pi_3 = \frac{1}{3}$. The dashed lines are known as the Bayes decision boundaries. Were they known, they would yield the fewest misclassification errors, among all possible classifiers.

- Three equally-sized Gaussian classes are shown with class-specific mean vectors and a common covariance matrix.
- The three ellipses represent regions that contain 95% of the probability for each of the three classes.
- The dashed lines are the Bayes decision boundaries, i.e. they represent the set of values x for which $\delta_k(x) = \delta_l(x)$.
- Note there are three lines representing the Bayes decision boundaries because there are three pairs of classes among the three classes.
- That is, one Bayes decision boundary separates class 1 from class 2, one separates class 1 from class 3, and one separates class 2 from class 3.
- These three Bayes decision boundaries divide the predictor space into three regions.
- Then, the Bayes classifier will classify an observation according to the region in which it is located.

GMC - LDA when $p > 1$

- Once again, we need to estimate the unknown parameters $\mu_1, ...\mu_K$, $\pi_1, ..., \pi_K$ and Σ_i the formulas are similar to those used in the one dimensional case.
- To assign a new observation $X = x$, LDA plugs these estimates into (8) to obtain quantities $\hat{\delta_k}(x)$
- Once we have estimates $\hat{\delta}_k(x)$, we can turn these into estimates for class probabilities

$$\hat{Pr}(Y = k|X = y) = \frac{e^{\hat{\delta}_k(x)}}{\sum_{l=1}^{K} e^{\hat{\delta}_l(x)}}$$

- So classifying to the largest $\hat{\delta}_k(x)$ amounts to classifying to the class for which $\hat{Pr}(Y = k|X = x)$ is largest.
- When $K = 2$, we classify to class 2 if $\hat{Pr}(Y = k|X = x) \leq 0.5$, else to class 1.

Classification - Evaluation

Classification - Evaluation

- How to evaluate a classifier?
- We can use a *confusion matrix* to extract information about the performance of the classifier.

		True Class		
		- or Null	+ or Null	Total
Predicted Class	- or Null	True Neg. (TN)	False Neg. (FN)	N*
	+ or Non-null	False Pos. (FP)	True Pos (TP)	P*
	Total	N	P	

Table 2: Confusion Matrix. Possible results when applying a classifier or diagnostic test to a population.

- N (P) - true number of negative (positive) cases
- N^* (P^*) - number of classified negative (positive) cases

Classification - Evaluation

Name	Definition	Synonyms
False Pos. rate	FP/N	Type I error, 1−Specificity
True Pos. rate	TP/P	1−Type II error, power, sensitivity, recall
Pos. Pred. value	TP/P*	Precision, 1−false discovery proportion
Neg. Pred. value	TN/N*	

Table 3: Important measures for classification and diagnostic testing, derived from the previous table.

- Class specific performance - i.e. how does the classifier perform in identifying those with or without the specified characteristic.
- Sensitivity (True Positive Rate): the proportion of those who have been classified as positive out of those who are actually truly positive.
- Specificity (True Negative Rate): proportion of those who have been classified as negative out of those who are truly negative.

- Let's perform classification on the default data set to predict if an individual will default, based on their income and credit card balance. (A confusion matrix example)...
- Use 10,000 training observations and an Linear Discriminant Analysis algorithm

		True Default Status		
		No	Yes	Total
Predicted	No	9644	252	9896
Default Status	Yes	23	81	104
	Total	9667	333	10000

Table 4: A confusion matrix compares the classifiers predictions to the true default statuses for the 10,000 training observations in the Default data set. Elements on the diagonal of the matrix represent individuals whose default statuses were correctly predicted, while off-diagonal elements represent individuals that were misclassified. The classifier made incorrect predictions for 23 individuals who did not default and for 252 individuals who did default.

Questions

a) What is the sensitivity of the classifier?

b) What is the specificity of the classifier?

c) What is the precision of the classifier?

d) What is the training error rate of the overall model? How does this training error do, in comparison to a null classifier - i.e. one that always predicts that an individual will not default?

e) What can you say about the class specific performance of the classifier - that is the performance of the classifier for identifying true defaulter, and true non-defaulters?

Classification - Evaluation

Confusion Matrix interpretation:

- $(23 + 252)/10,000$ errors - a 2.75% misclassification rate !
- Some caveats: This is training error, and we may be overfitting (the ratio of p vs. n is important)
- If we used a null classifier - i.e. always classify to class `No` in this case — we would make 333/10000 errors, or only 3.33%.
- Of the true `No`'s, we make $23/9667 = 0.2\%$ errors; of the true `Yes`'s, we make $252/333 = 75.7\%$ errors! $\Rightarrow$While overall error rate is low, the error among those who defaulted is pretty high.

Classification - Evaluation

- False positive rate: The fraction of negative examples that are classified as positive — 0.2% in example
- False negative rate: The fraction of positive examples that are classified as negative — 75.7% in example
- The previous table was created by classifying each observation to the class for which the probability is the largest.
- For the Bayes classifier this amounts to assign a value to class `Yes` if:

$$\widehat{Pr}(\texttt{Default} = \texttt{Yes} | \texttt{Balance}, \texttt{Student}) > 0.5$$

- We can change the two error rates by changing the threshold from 0.5 to some other value in $[0, 1]$:

$$\widehat{Pr}(\texttt{Default} = \texttt{Yes} | \texttt{Balance}, \texttt{Student}) \geq \text{threshold}$$

and vary the "threshold".

- Use 10000 training observations and a threshold value of 0.2:

		True Default Status		
		No	Yes	Total
Predicted	No	9432	138	9570
Default Status	Yes	235	195	430
	Total	9667	333	10000

Table 5: A confusion matrix compares the classifiers predictions to the true default statuses for the 10,000 training observations in the Default data set. Elements on the diagonal of the matrix represent individuals whose default statuses were correctly predicted, while off-diagonal elements represent individuals that were misclassified. The classifier made incorrect predictions for 23 individuals who did not default and for 252 individuals who did default.

Questions:

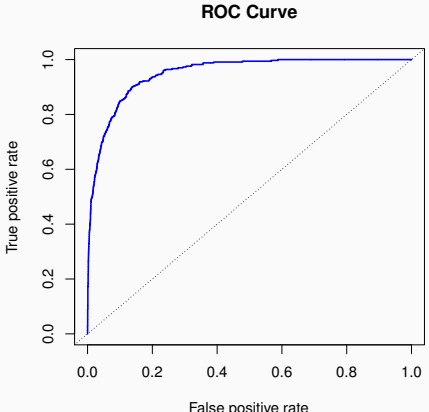
a) What is the performance of the classifier among the individuals that truly defaulted?

b) What is the performance of the classifier among the individuals who do not defaulted?

c) What is the overall error rate of the model?

Classification - Evaluation

- Varying the threshold...
- The *Receiver Operating Characteristic* (ROC) curve is a popular graphic to simultaneously display the False positive and True positive rate.



Classification - Evaluation

The ROC curve continued...

- The ROC curve traces out two types of error as we vary the threshold value for the probability of default (actual thresholds are not shown):
  - The true positive rate (sensitivity): the fraction of defaulters that are correctly identified, using a given threshold value.
  - The false positive rate (1-specificity): the fraction of non-defaulters that we classify incorrectly as defaulters, using that same threshold value.
- The ideal ROC curve hugs the top left corner, indicating a high true positive rate and a low false positive rate.
- The dotted line represents the "no information" classifier; this is what we would expect if student status and credit card balance are not associated with probability of default.
- The performance of the classifier over all possible thresholds is given by the area under the ROC curve (AUC). Higher AUC is better.
- ROC curves are useful for comparing different classifier.

Homework

- **Homework: Please read Section 13.1 on Hypothesis Testing**

Disclaimer: This material has been prepared by Philipp Kremer and Constantin Lisson in 2021 and draws very extensively on James, G., Witten, D., Hastie, T. & Tibshirani, R. (2021). *An introduction to statistical learning* and the corresponding lecture slides available from these authors.

Homework Exercises

Suppose we collect data for a group of students in a statistics class with variables $X_1 =$ hours studied, $X_2 =$ undergrad GPA, and $Y =$ receive an A. We fit a logistic regression and produce estimated coefficient, $\hat{\beta}_0 = -6, \hat{\beta}_1 = 0.05, \hat{\beta}_2 = 1$.
Questions:

a) Estimate the probability that a student who studies for 40h and has an undergrad GPA of 3.5 gets an A in the class.

b) How many hours would the student in part a) need to study to have a 50% chance of getting an A in the class?

Solutions to review and homework questions

Solutions to review and homework questions

Review questions from Slide 11:

a)

$$\frac{p(X)}{(1-p(X))} = 0.37$$
$$p(X) = 0.37(1-p(X))$$
$$1.37p(X) = 0.37$$
$$p(X) = \frac{0.37}{1.37} = 27\%$$

b)

$$odds = \frac{p(X)}{(1-p(X))} = \frac{0.16}{0.84} = 0.19$$

Solutions to review and homework questions

Review questions from Slide 16:

a) Approx. 95% CI: $0.0055 \pm 2 \times 0.0002 = [0.0051, 0.0059]$

b) Yes, as the p-values for both coefficients β_0 and β_1 are smaller than $\alpha = 0.05$.

c) Given that `balance` is zero, the probability for default is equal to:

$$p(X) = \frac{e^{-10.6513}}{1 - e^{-10.6513}} \approx 0 \tag{13}$$

d) A one-unit increase in balance is associated with an increase in the log odds of default by 0.0055 units.

Solutions to review and homework questions

Review questions from Slide 45:

a) Sensitivity: Percentage of true defaulters that are identified

$$\frac{81}{333} = 0.243 \approx 24.3\% \tag{14}$$

b) Specificity: Percentage of non-defaulters that have been correctly classified

$$(1 - \frac{23}{9667}) = 0.998 \approx 99.8\% \tag{15}$$

c) Precision: Percentage of those predicted as `default`, that have truly defaulted

$$\frac{81}{104} = 0.7788 \approx 77.88\% \tag{16}$$

Solutions to review and homework questions

Review questions from Slide 45 - cont.:

d) Training Error Rate: $\frac{(23+252)}{10000} = 0.0275 \approx 2.75\%$ If we classified to the prior — always to class No in this case — we would make $\frac{333}{10000}$ errors, or only 3.33% - thus using a trivial *null* classifier leads to a only slightly higher training set error.

e) Of the true `No's`, we make $\frac{23}{9667} = 0.002 \approx 0.2\%$ errors; of the true `Yes's`, we make $\frac{252}{333} = 0.757 \approx 75.7\%$ errors!

Solutions to review and homework questions

Review questions from Slide 49:

a) $\frac{195}{333} = 0.5856 \approx 58.56\%$

b) $\frac{9432}{9667} = 0.9757 \approx 97.57\%$

c) Training Error Rate: $\frac{(235+138)}{10000} = 0.0373 \approx 3.73\%$

Solutions to review and homework questions

Review questions from Slide 54: Note:

$$p(X) = \frac{e^{(\beta_0+\beta_1 X_1+\beta_2 X_2)}}{(1+e^{(\beta_0+\beta_1 X_1+\beta_2 X_2)})}$$

where: $X_1 = \textit{hours studied}$ and $X_2 = \textit{undergrad GPA}$

a)

$$\begin{aligned} p(X) &= \frac{e^{(-6+0.05X_1+X_2)}}{1+e^{(-6+0.05X_1+X_2)}} \\ &= \frac{e^{(-6+0.05\times 40+3.5)}}{(1+e^{(-6+0.05\times 40+3.5)})} \\ &= \frac{e^{(-0.5)}}{(1+e^{(-0.5)})} \\ &= 37.75\% \end{aligned}$$

Solutions to review and homework questions

Review questions from Slide 54:

b)

$$p(X) = \frac{e^{(-6+0.05X_1+X_2)}}{1+e^{(-6+0.05X_1+X_2)}}$$

$$0.50 = \frac{e^{(-6+0.05X_1+3.5)}}{1+e^{(-6+0.05X_1+3.5)}}$$

$$0.50(1+e^{(-2.5+0.05X_1)}) = e^{(-2.5+0.05X_1)}$$

$$0.50 + 0.50e^{(-2.5+0.05X_1)} = e^{(-2.5+0.05X_1)}$$

$$0.50 = 0.50e^{(-2.5+0.05X_1)}$$

$$\log(1) = -2.5 + 0.05X_1$$

$$X_1 = \frac{2.5}{0.05} = 50 \ \textit{hours}$$

This material draws extensively on James, G., Witten, D., Hastie, T. & Tibshirani, R. (2021). *An introduction to statistical learning* and the lecture slides available from these authors.