IEOR E4650 Business Analytics

# Session 8: Logistic Regression and Linear Discriminant Analysis

Spring 2018



Prof. Adam Elmachtoub

## Outline

- Classification
- Logistic Regression
- Linear Discriminant Analysis

## Regression and Classification

Explanatory/Input variables vs. Response/Output variables

1. Obtain some kind of model based on observations, or training data $\{(x_i, y_i), i = 1, 2, \ldots, n\}$, through a process called learning (or estimation).

2. Use that model to predict something about data you haven't seen before, but that comes from the same distribution as the training data, called test data.

Types of Supervised Learning

- Regression: predict a quantitative output variable

- Classification: predict a qualitative/categorical output variable

Today we begin to consider classification methods.



## Classification

In many situations we are trying to predict outcomes that are not numeric, for example:

- patient outcomes: recovery, stroke, internal bleeding
- credit card default: yes or no
- vote in election: each of the possible candidates
- car purchases: each of the offered models or no purchase

In all these cases we are trying to predict an outcome from a discrete set of options. We can represent each outcome by a number, say

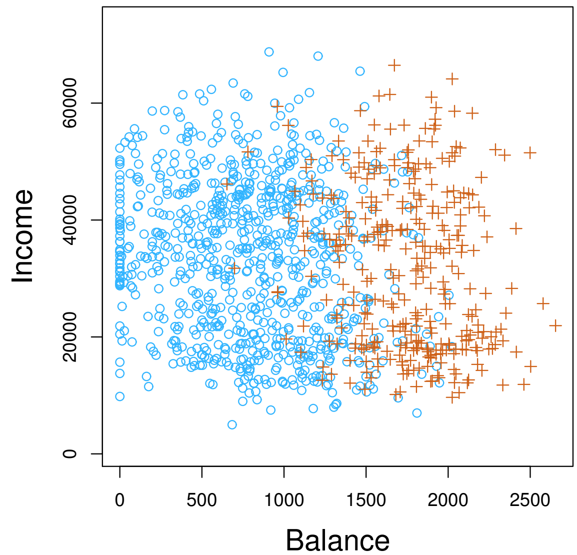
`1 = recovery, 2 = stroke, 3 = internal bleeding,`

but the numbers will be arbitrarily assigned. There is no natural ordering of outcomes (as opposed to numerical values).

# Example: Credit Card Default

A data set contains information on individuals and whether they defaulted on their credit card.

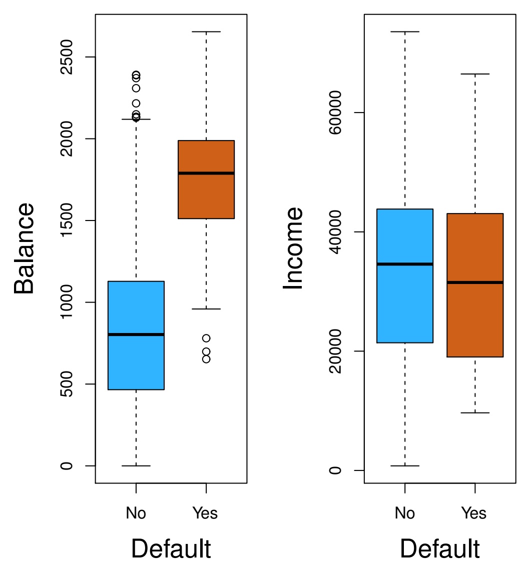
- Individuals who defaulted are shown in orange.
- Individuals who did not default are shown in blue.






# Example: Credit Card Default

Using `Boxplot` we can compare the distribution of the explanatory variables `Balance` and `Income` in the two groups.



## Binary Response and Log-Odds

When there are only two possible outcomes, we are trying to predict a *binary response*. If we label `1 = Yes` and `0 = No` the conditional mean is

$$E[Y|\mathbf{X}] = 1 \times \Pr(Y = 1|\mathbf{X}) + 0 \times \Pr(Y = 0|\mathbf{X}) = \Pr(Y = 1|\mathbf{X}).$$

We might want to use a linear regression, but this expectation is a probability and therefore between 0 and 1. Linear regression is not appropriate as it can give values outside $[0, 1]$.

Let $q(\mathbf{X}) = \Pr(Y = 1|\mathbf{X})$ be the probability of success given $\mathbf{X}$

The *odds of success* $q(\mathbf{X})/(1 - q(\mathbf{X}))$ can take any value in $(0, \infty)$.

The *log-odds* $\log[q(\mathbf{X})/(1 - q(\mathbf{X}))]$ can take any value in $(-\infty, \infty)$.



## Logistic Regression for Binary Response

A logistic regression estimates the log-odds using a linear combination of the explanatory variables:

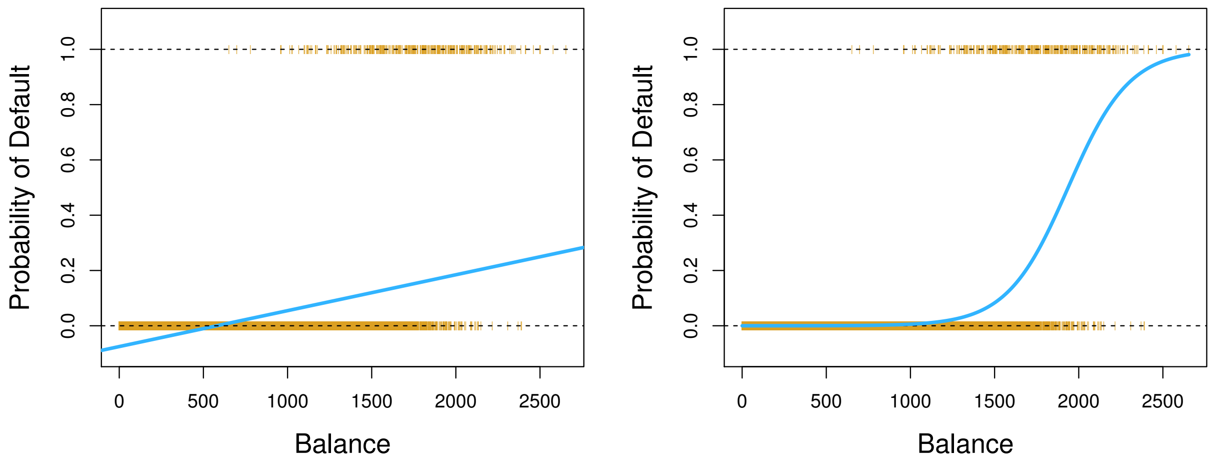
$$\ln\left(\frac{q(\mathbf{X})}{1 - q(\mathbf{X})}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p.$$

Which we can equivalently write as:

$$q(\mathbf{X}) = \Pr(Y = 1|\mathbf{X}) = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)}.$$

# Logistic Regression vs. Linear Regression

Figure: Classification using the `Default` data




- Left: estimated probability of default using linear regression. Some estimated probabilities are negative.
- Right: predicted probability of default using logistic regression. All probabilities lie between 0 and 1.



# Maximum Likelihood Estimation

- The logistic regression model gives us the probability of each outcome given the covariates.
- Assume our data contains $n$ observations containing covariates $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})$ and binary outcome indicator $y_i \in \{0, 1\}$.
- We can ask what is the likelihood of observing this data if outcomes were indeed drawn according the probabilities given by the logistic model. The best logistic model will maximize the likelihood of observing the data we observe.
- We let $f(y_1, \dots, y_n | x_1, \dots, x_n)$ be the joint distribution of the responses given the features.
- We will assume the data points are independent, and thus $f(y_1, \dots, y_n | x_1, \dots, x_n) = \prod_{i=1}^{n} f(y_i | x_i)$
- $f(y_i|x_i)$ is the probability of seeing $y_i$ given $x_i$, thus $f(1|x_i) = q(x_i)$

# Maximum Likelihood Estimation, Cont.

Maximum Likelihood Estimation (MLE) is a general estimation method the chooses the parameters of the model to maximizes the likelihood (probability) of the realized observation.

Under our binary logistic model the likelihood function is

$$
\begin{aligned}
& \prod_{i=1}^{n} f(y_i|x_i) \\
= \quad & \prod_{i=1}^{n} (q(x_i))^{y_i} (1 - q(x_i))^{1-y_i} \\
= \quad & \prod_{i=1}^{n} \left[ \frac{\exp(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip})} \right]^{y_i} \cdot \left[ \frac{1}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip})} \right]^{1-y_i}
\end{aligned}
$$



# Maximum Likelihood Estimation, Cont.

Maximizing the likelihood function w.r.t. $(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)$ is equivalent to minimizing the deviance (the negative logarithm of the likelihood)

$$D = -\left[ \sum_{i=1}^{n} y_i \log(q(x_i)) + \sum_{i=1}^{n} (1 - y_i) \log(1 - q(x_i)) \right],$$

where $q_i$ is

$$q(x_i) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip})}{(1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}))}.$$

## Logistic Regression in R

- The R command `glm()` fits generalized linear models, a class of models that includes logistic regression.
- For logistic regression, the family of error distribution is the binomial, i.e., `glm(formula, family = binomial, data)`
- Use function `summary()` to see the logistic regression results.



## Example: Default Data

- The `Default` data is part of the ISLR library. This is simulated data of 10,000 credit card accounts. The outcome variable we are interested in is whether there was a default.
- The predictors include the balance and income of the card holder, and an indicator for students.

```
> head(Default)
  default student   balance    income
1      No      No  729.5265 44361.625
2      No     Yes  817.1804 12106.135
3      No      No 1073.5492 31767.139
4      No      No  529.2506 35704.494
5      No      No  785.6559 38463.496
6      No     Yes  919.5885  7491.559
```

# Example: Default Data, Cont.

Using a Logistic regression, we can try to predict default probabilities.

- Can we predict who will default?
- What determines defaults?

Sample R code:

```
>library(ISLR)
>attach(Default)
>lgrf = glm(default ~ balance + income + student ,
+ data = Default, family = binomial)
```



# Summary of Logistic Regression

```
> summary(lgrf)
Call:
glm(formula = default ~ balance + income + student, family = binomial,
    data = Default)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4691  -0.1418  -0.0557  -0.0203   3.7383

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.087e+01  4.923e-01 -22.080  < 2e-16 ***
balance      5.737e-03  2.319e-04  24.738  < 2e-16 ***
income       3.033e-06  8.203e-06   0.370  0.71152
studentYes  -6.468e-01  2.363e-01  -2.738  0.00619 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2920.6  on 9999  degrees of freedom
Residual deviance: 1571.5  on 9996  degrees of freedom
AIC: 1579.5
```

## Understanding the Regression Summary

`call`: we start off with a repeat of the model specification.

`Deviance statistics`: Min and Max; first quartile (1Q) and third quartile (3Q); Median to the deviance of the model

`Coefficients`

- This is the table of primary interest. Here, we get estimates of the regression coefficients, standard errors, and tests of whether each regression coefficient is statistically different from 0. The layout is nearly identical to the corresponding part of the `lm` output.
- The `z-statistic` plays the same role of `t-statistic` in the linear regression output.
- A large value of the z-stat is evidence against the null hypothesis $H_0 : \beta_i = 0$, and is equal to $\widehat{\beta}_i / SE(\widehat{\beta}_i)$.
- A small p-value indicates the coefficient is statistically significant.



## Understanding the Regression Summary, Cont.

- `Null deviance`: is the deviance of a model that contains only the intercept. Serves as a benchmark.
- `Residual deviance`: is the deviance the full model with the estimated coefficients. Corresponds to the residual sum of squares of ordinary regression analyses, and can be compared with the null deviance.
- `AIC`: (Akaike information criterion) a measure of goodness of fit that takes the number of fitted parameters into account.
- `Number of iterations`: technical information about the fitting procedure. A large number may indicate the software failed to find the optimal coefficients.

## Default Data (continued)

The $p$-value on `Income` was large, suggesting that we should remove it.

```
> lgrf = glm(default ~ balance  + student , data = Default,
+family = binomial)
> summary(lgrf)
...
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.075e+01  3.692e-01 -29.116  < 2e-16 ***
balance      5.738e-03  2.318e-04  24.750  < 2e-16 ***
studentYes  -7.149e-01  1.475e-01  -4.846 1.26e-06 ***
...
Residual deviance: 1571.7  on 9997  degrees of freedom
```

We can use model selection tools here, for example LASSO.



## Turning Logistic Regression into a Classifier

- We have estimates of the coefficients $\hat{\beta}_0, \ldots, \hat{\beta}_p$.
- We can use this to estimate success probability $\hat{q}(x)$ for $X = x$.
- We can classify $X = x$ as a success if $\hat{q}(x) \geq \tau$, and as a failure otherwise.
- As $\tau$ decreases, more instances are classified as successful, increasing both the true positive and the false positive rate.
- As $\tau$ increases, fewer instances are classified as successful, decreasing the true positive and the false positive rate.
- $\tau$ is a parameter that is chosen by the user. For example, in fraud detection $\tau$ may be significantly smaller than $1/2$ to be useful!
- We will talk more about this in following classes.

## Estimation Accuracy

- Out of sample observations: $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, where $y_1, \ldots, y_n$ are qualitative.
- Estimation accuracy: *error rate*

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

where $\widehat{y}_i$ is the predicted class for the $i$th observation; $I(y_i \neq \widehat{y}_i)$ is indicator variable, i.e.,

$$I(y_i \neq \widehat{y}_i) = \begin{cases} 1, & y_i \neq \widehat{y}_i \\ 0, & \text{otherwise.} \end{cases}$$

- When false positives are just as bad as false negatives, you use this to do Model Selection and Model Assesment
- Note that the choice of $\tau$ doe note influence the coefficients determined by the logistic regression



## Logistic Regression with Shrinkage methods

Minimize the deviance $D$ plus the shrinkage penalty on $\beta$

- Logistic Regression plus Ridge Penalty

$$D + \lambda \sum_{j=1}^{p} \hat{\beta}_j^2$$

- Logistic Regression plus LASSO Penalty

$$D + \lambda \sum_{j=1}^{p} |\hat{\beta}_j|$$

where $D$ is

$$D = -\left[\sum_{i=1}^{n} y_i \log(q_i) + \sum_{i=1}^{n}(1 - y_i)\log(1 - q_i)\right],$$

and $q_i$ is

$$q_i = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p)}{(1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p))}$$

## Logistic Regression with Shrinkage methods cont.

- Can also do Elastic Net with Logistic Regression
- Can use package glmnet with 'binomial' family
- Almost exact same as doing linear regression with shrinkage penalties!
- Can apply model selection ideas to choose $\lambda$ and $\tau$



## Multinomial Logistic Regression

Often we have more than 2 possible outcomes, as noted in previous slides. Assume that we have $K \geq 2$ possible (unordered) outcomes $\{0, 1, 2, \ldots, K-1\}$, and denote the probabilities of each option $k = 1, \ldots, K-1$ by

$$q_k(\mathbf{X}) = \Pr(Y = k|\mathbf{X}) = \frac{\exp(\beta^{(k)\top}\mathbf{X})}{1 + \sum_{\ell=1}^{K} \exp(\beta^{(\ell)\top}\mathbf{X})}$$

and for option 0 by

$$q_0(\mathbf{X}) = \Pr(Y = 0|\mathbf{X}) = \frac{1}{1 + \sum_{\ell=1}^{K} \exp(\beta^{(\ell)\top}\mathbf{X})}$$

Option 0 is usually the default or "outside option" and serves as a normalization (i.e., we set $\beta^{(0)} = 0$). For any option $k$ the coeficients $\beta^{(k)}$ capture how important are the covariates for option $k$. Note this is exactly the binary logistic model when the options are $\{0, 1\}$.

## Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis uses Bayes' law to classify points.

Assume again that the outcome is $Y \in \{1, \ldots, K\}$. We are interested in the probability of outcome $k$ given covariates $x$, which we denote

$$p_k(x) = P(Y = k|X = x)$$

Consider separating the data by looking only at observations with $Y = k$. We assume that:

- We can get the probabilities $\pi_k = P(Y = k)$ of getting an observation from the distribution of $Y = k$.
- The conditional distribution of $X$ given $Y = k$, denoted $f_k(x)$, is normally distributed with mean $\mu_k$ and covariance matrix $\Sigma_k$
- $\Sigma_k = \Sigma \quad k = 1, \ldots, K$, that is, all covariance matrices are equal.



## Classification Using Bayes' Law

Under the assumptions above

- The prior probabilities are $\pi_k = P(Y = k)$.
- The conditional distribution of $X|Y = k$ is $f_k(x)$.

By Bayes' law, the posterior probability of $Y|x$ is given by:

$$p_k(x) = P(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}.$$

To apply this formula we need to know or estimate $\pi_k, \mu_k$'s and $\Sigma$.

## Estimating the LDA

Given data $(x_i, y_i), i = 1, \ldots, n$, we can estimate $\pi_k, \mu_k$ and $\Sigma$ as follows:

$$\hat{\pi}_k = \frac{\sum_{i=1}^n \delta(y_i = k)}{n},$$

$$\hat{\mu}_k = \frac{\sum_{i=1}^n \delta(y_i = k) x_i}{\sum_{i=1}^n \delta(y_i = k)},$$

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n (x_i - \sum_{k=1}^K \delta(y_i = k)\hat{\mu}_k)(x_i - \sum_{k=1}^K \delta(y_i = k)\hat{\mu}_k)'.$$

Here $\delta(y_i = k) = 1$ if $y_i = k$ and zero otherwise.



## Linear Discriminant Analysis as a Classifier

- Classify an observation $x$ to the label $k$ with highest $p_k(x)$.
- This is equivalent to assigning observation $x$ to the class for which

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

  is the largest.
- This expression is linear in $x$, for each $k$, and this is why it is called linear discriminant analysis.
- Use $\hat{\mu}_k, \hat{\pi}_k$ and $\hat{\Sigma}$ instead in the formula if $\mu_k, \pi_k$ and $\Sigma$ are unknown.

# Linear Discriminant Analysis: Graphical Example

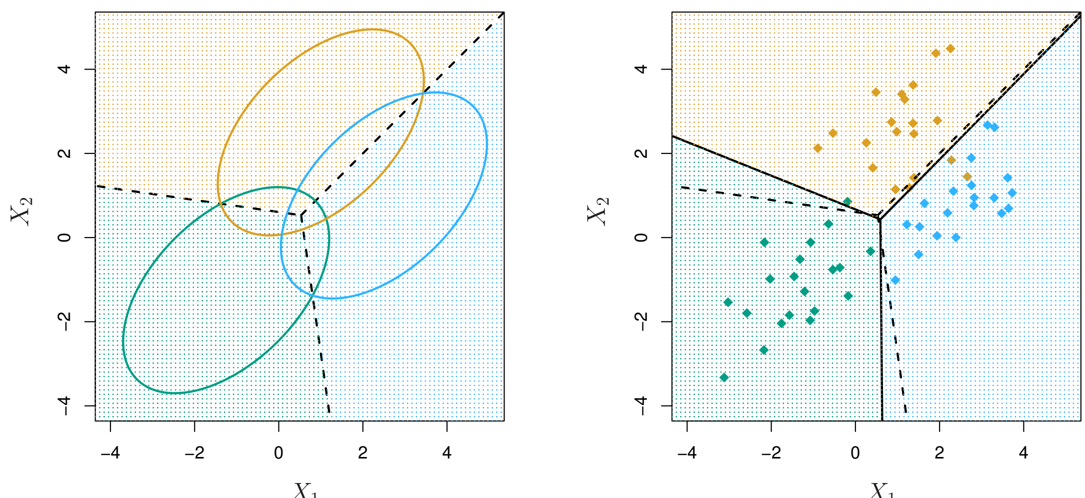



Figure: Three classes, each drawn from multivariate normal with two predictors. The ellipses contain 95% of the probability for each of class. Dashed lines are the Bayes' boundaries. Right: 20 observations from each class, and dashed lines are LDA boundaries.



# Linear Discriminant Analysis in R

We will apply LDA to the data set Default from the ISLR library.

```
> library(MASS)
> ldaf = lda(default ~ balance + income, data = Default)
> ldaf
Call:
lda(default ~ balance + income, data = Default)

Prior probabilities of groups:
    No    Yes
0.9667 0.0333

Group means:
      balance   income
No   803.9438 33566.17
Yes 1747.8217 32089.15

Coefficients of linear discriminants:
                 LD1
balance 2.230835e-03
income  7.793355e-06
```

# Summary

- Logistic Regression and Linear Discriminant Analysis are two classification methods
- They are also two forms of regression, give exact probabilities!
- In both cases, we take the option with the highest probability to get a classifier.
- But actually we can modify the classification rule depending on the cost of misclassification
- Next time we talk about Pandora (Read case) and $k$-Nearest Neighbors