Linear & Logistic Regression and Classification

Kishan G. Mehrotra

Department of EECS, Syracuse University

October, 23

We consider three important problems in this presentations:

1. Linear Regression
2. Linear Classification
3. Logistics Regression (classification)

LINEAR REGRESSION

Linear regression is a useful tool for predicting a quantitative response and is a widely used statistical learning method.

1. Businesses often use linear regression to understand the relationship between advertising spending and revenue.
2. Medical researchers often use linear regression to understand the relationship between drug dosage and blood pressure of patients.
3. Agricultural scientists often use linear regression to measure the effect of fertilizer and water on crop yields.

For the above examples; the regression model would take the following form:

1. Businesses: revenue $= \beta_0 + \beta_1$ (ad spending)
   The coefficient β_0 would represent total expected revenue when ad spending is zero and the coefficient β_1 would represent the average change in total revenue when ad spending is increased by one unit (e.g. one dollar).
2. Medical: blood pressure $= \beta_0 + \beta_1$(dosage)
3. Agricultural:
   crop yield $= \beta_0 + \beta_1$(amount of fertilizer) $+ \beta_2$ (amount of water)

Simple Linear Regression

The model is:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

where ϵ represents potential error in the linear relationship.

Assumption

ϵ is a random variable with mean 0 and standard deviation σ.

Business example

The exact relation between revenue and ad. spending is rarely exact and linear The difference is called 'error' resulting in the model
revenue $= \beta_0 + \beta_1$ (ad spending) $+$ error

Estimating the Coefficients

In practice, β_0, β_1 and σ are unknown. So, we use data to estimate the coefficients.

Dataset

$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$

Least Square Estimation

Minimize Residual Sum of Squares (RSS) defined as:

$$\text{RSS} = (y_1 - \beta_0 - \beta_1 x_1)^2 + (y_2 - \beta_0 - \beta_1 x_2)^2), \cdots, (y_n - \beta_0 - \beta_1 x_n)^2$$

Estimates

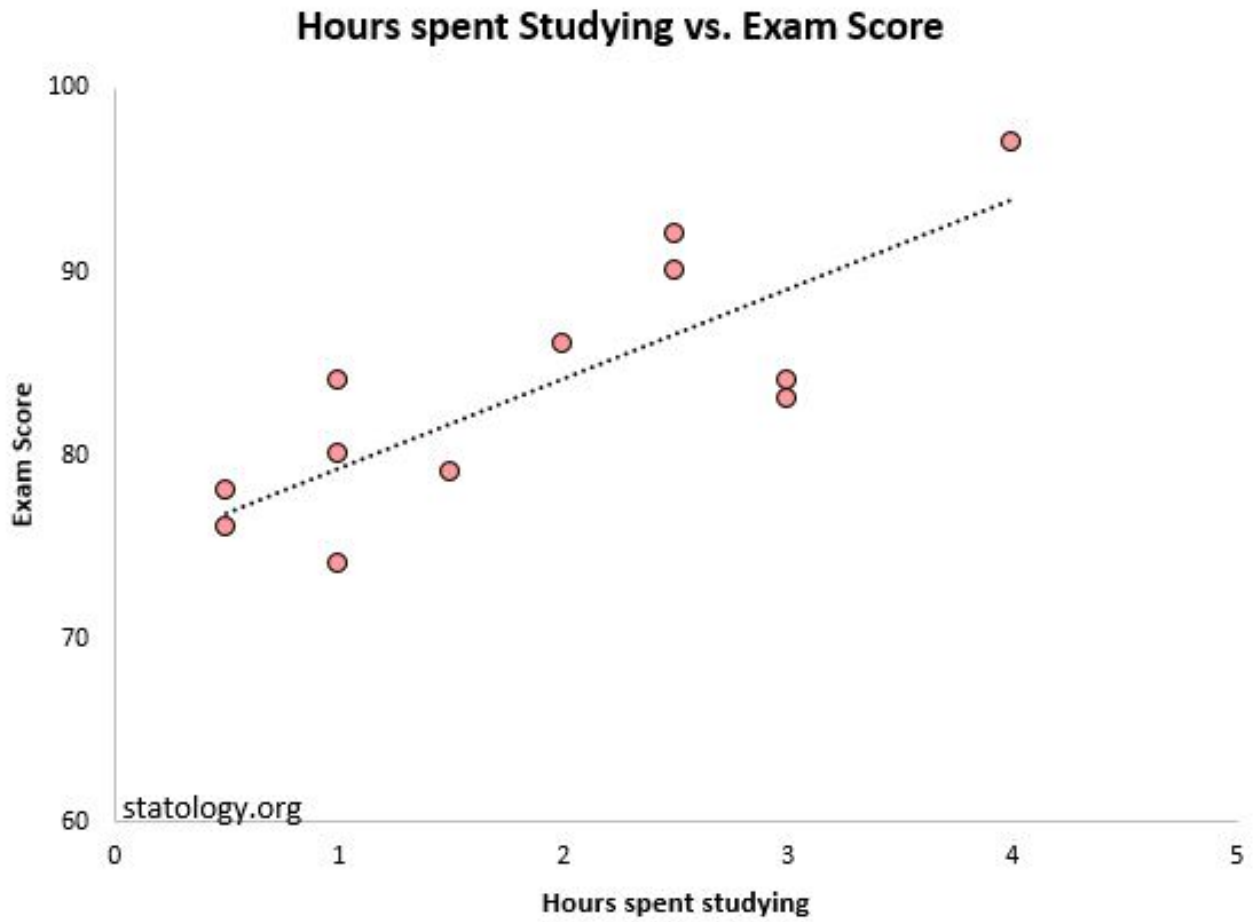
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

and

$$\hat{\sigma}^2 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

Best Fit and Scatter Plot of Errors



How Good are the estimates?

Estimate of β_0

$$\mathsf{E}(\hat{\beta}_0) = \beta_0 \text{ and } \mathsf{SE}(\hat{\beta}_0^2) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right]$$

Estimate of β_1

$$\mathsf{E}(\hat{\beta}_1) = \beta_1 \text{ and } \mathsf{SE}(\hat{\beta}_1^2) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

These values, in addition to the assumption that the distribution of ϵ is Gaussian, allows us to test hypotheses about these parameter estimates.

Assessing the Accuracy of the Model

The quality of a linear regression fit is typically assessed using two related quantities: the residual standard error (RSE) and the R^2 statistic.

The residual standard error (RSE)

$$\text{RSE} = \sqrt{\frac{1}{n-2}\text{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

The R^2 Statistics

$$R^2 = \frac{\text{Total Sum of Squares} - \text{RSS}}{\text{Total sum of Squares}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

where $\text{TSS} = \sum_{i=1}^{n}(y_i - \bar{y}_i)^2$

Typically, a large value of R^2 implies the fit is good. In current setting $R^2 = \text{cov}^2(X, Y)$

Predicting value of Y

Given a new x_0, the predicted value $y_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$.

Summary of Regression Analysis

	coefficients	Standard error	t-value	p-value
Intercept				
β_1				

	df	SS	Mean SS	F	p-value
Regression	1	RSS	RMS	$\frac{\text{RMS}}{\text{Res.MS}}$	
Residual	$n-2$	Residual SS	Res. MS		
Total	$n-1$	Total SS			

Summary for the advertising data; number of units sold on TV advertising budget

	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

Quantity	Value
Residual standard error	3.26
R^2	0.612
F-statistic	312.1

Multiple Linear Regression

In practice we often have more than one predictor. Agricultural scientists often use linear regression to measure the effect of fertilizer and water on crop yields.

The Model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{X} + \epsilon$$

As before, estimates of unknown parameters are obtained using the least square approach, that is: Minimize Residual Sum of Squares (RSS) defined as:

$$\text{RSS} = (y_1 - \beta_0 - \boldsymbol{\beta}^T \boldsymbol{X}_1)^2 + (y_2 - \beta_0 - \boldsymbol{\beta}^T \boldsymbol{X}_2)^2) + \cdots (y_n - \beta_0 - \boldsymbol{\beta}^T \boldsymbol{X}_n)^2$$

Prediction

As before

$$\hat{y} = \hat{\beta}_0 + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}$$

Some Important Questions

When we perform multiple linear regression, we usually are interested in answering a few important questions.

1. Is at least one of the predictors $X_1, X_2, \cdots, X_p$ useful in predicting the response?
2. Do all the predictors help to explain Y , or is only a subset of the predictors useful?
3. How well does the model fit the data?
4. Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

LINEAR DISCRIMINANT ANALYSIS

Typical Dataset

response variable Y is Yes/No

Age	Sex	ChestPain	RestBP	Chol	Fbs	RestECG	MaxHR	ExAng	Oldpeak	Slope	Ca	Thal	AHD
63	1	typical	145	233	1	2	150	0	2.3	3	0.0	fixed	No
67	1	asymptomatic	160	286	0	2	108	1	1.5	2	3.0	normal	Yes
67	1	asymptomatic	120	229	0	2	129	1	2.6	2	2.0	reversable	Yes
37	1	nonanginal	130	250	0	0	187	0	3.5	3	0.0	normal	No
41	0	nontypical	130	204	0	2	172	0	1.4	1	0.0	normal	No

What is Classification?

In a classification problem (also known as discriminant analysis) the goal is to classify a new observation to Class 1, Class 2, ..., or Class K.

This determination depends on the investigation of a set of *training data*; *i.e.*, data objects whose class label is known.

Examples:

1. To predict whether an email is a spam and should be moved to the Junk folder.
2. Handwritten Digit Recognition; *i.e.*, to identify images of single digits 0 - 9 correctly.
3. Given the EKG of a patient find the associated heart disease.

The Classification problem

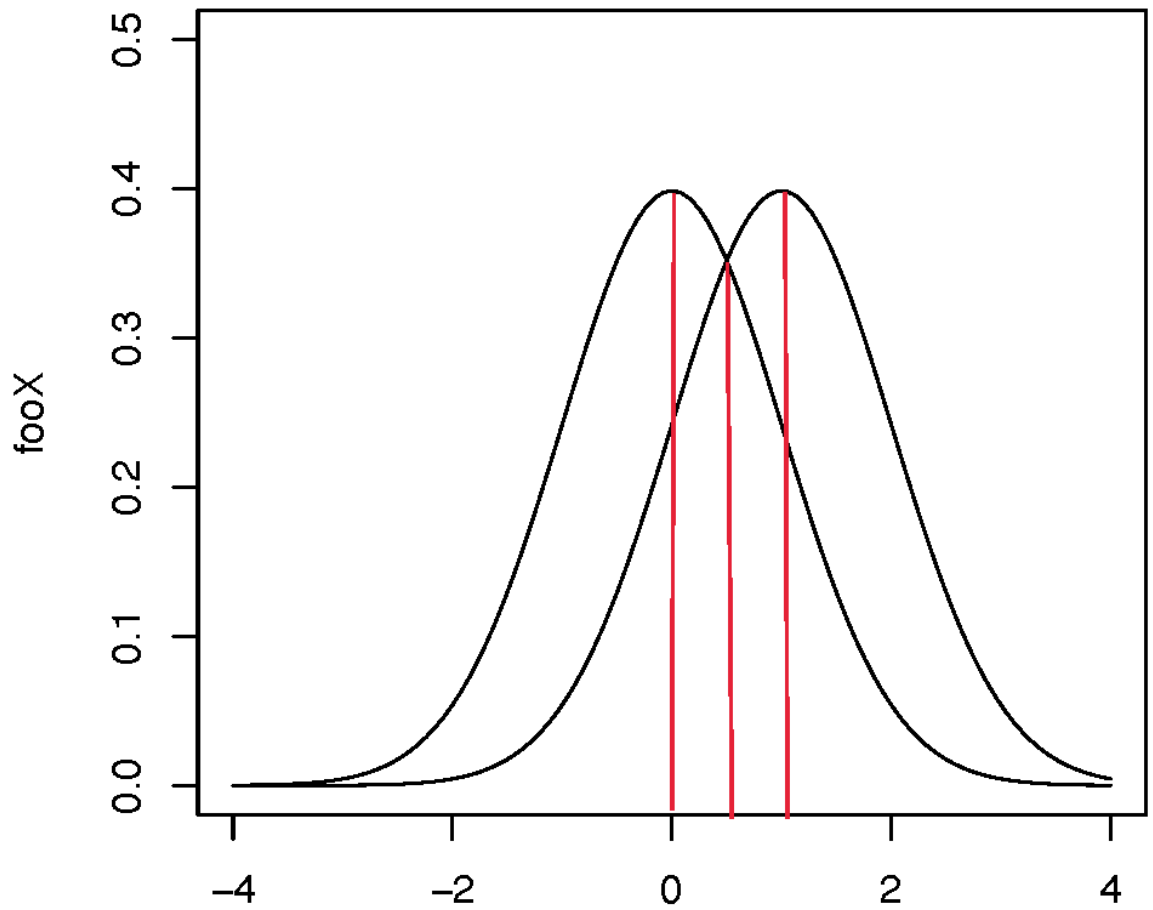
We assume that we know the distribution of the predictors $\boldsymbol{X}$ given the different categories that Y can takes, i.e., we know

$$P(\boldsymbol{X} = \boldsymbol{x}|Y = k).$$

Our Goal is : Find the category of Y given an observation $X = x_0$
In other words, for example, we know how the EKG of a patient suffering from heart disease k should look like for $k = 1, \cdots, K$ and our goal is: *Given the EKG of a NEW patient find the associated heart disease.*

A Simple Illustration

Consider two-class problem with one-dimensional predictor, following Gaussian distribution.



We consider three possible Approach

1. Distance Based Approach
2. Fisher's Approach
3. Bayes Approach

1. Distance Based Approach

In this, a geometric intuitive, we calculate the Mahalanobis distance of a new observation $\boldsymbol{x}_0$ from means of distributions $\boldsymbol{\mu}_k$ for $k = 1, 2, \cdots, K$; i.e.,

$$D_k(\boldsymbol{x}_0; \boldsymbol{\mu}_k) = (\boldsymbol{x}_0 - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_0 - \boldsymbol{\mu}_k)$$

The classification rule is:

Assign $\boldsymbol{x}_0$ to class with closest distance.

Gaussian Distribution

Suppose that the observations are from two Gaussian distributions with common population covariance $\boldsymbol{\Sigma}$, then the rule will be: If

$$(\boldsymbol{x}_0 - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_0 - \boldsymbol{\mu}_1) \leq (\boldsymbol{x}_0 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_0 - \boldsymbol{\mu}_2)$$

then assign $\boldsymbol{x}_0$ to class 1, else to class 2. However, the means and covariance are not known and are estimated by using the training set:

$$\hat{\boldsymbol{\mu}}_k = \overline{\boldsymbol{x}}_k \text{ and } \hat{\boldsymbol{\Sigma}} = \boldsymbol{S}^2$$

where

$$\boldsymbol{S}^2 = \frac{1}{(n_1 + n_2 - 2)} \left((n_1 - 1)\boldsymbol{S}_1^2 + (n_2 - 1)\boldsymbol{S}_2^2 \right),$$

and

$$\boldsymbol{S}_k^2 = \frac{1}{n_k} \sum_{i=1}^{n_k} (\boldsymbol{x}_i - \overline{\boldsymbol{x}}_k)(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_k)^T$$

2. Fisher's Linear Discriminant Rule

If

$$(\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_1)^T \boldsymbol{S}^{-1} (\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_1) \leq (\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_2)^T \boldsymbol{S}^{-1} (\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_2)$$

then assign $\boldsymbol{x}_0$ to class 1, else to class 2.

Straightforward simplification results in the following rule.

If

$$(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)^T \boldsymbol{S}^{-1} \left(\boldsymbol{x}_0 - \frac{1}{2}(\overline{\boldsymbol{x}}_1 + \overline{\boldsymbol{x}}_2) \right) \leq 0$$

then assign $\boldsymbol{x}_0$ to class 1, else to class 2.

3. A brief Introduction to Bayesian Approach to classification

Since, our goal is to find the value $Y = k$, given an observation $\boldsymbol{X} = \boldsymbol{x}_0$ using the Bayes' theorem we flip these distributions around to model

$$P(Y = k|\boldsymbol{X} = \boldsymbol{x}_0)$$

to get:

$$P(Y = k|\boldsymbol{X} = \boldsymbol{x}_0) = \frac{f_k(\boldsymbol{x}_0)\pi_k}{\sum_{J=1}^{K} f_j(\boldsymbol{x}_0)\pi_j}$$

where π_k is the a priori probability and $f_k(\boldsymbol{x}_0)$ is the density function for class k.

The Bayes' classifier

Assign $\boldsymbol{x}_0$ to the group which has the largest posterior probability.

This is easy – Assign $\boldsymbol{x}_0$ to the group which has the largest numerator.

Two Class problem: Gaussian Distribution

Suppose $P(\boldsymbol{X} = \boldsymbol{x}|Y = k)$ is Gaussian $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, $k = 1, 2$ and $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$. Then, the Bayes' theorem gives:

$$f(Y = k|\boldsymbol{X} = \boldsymbol{x}) = \frac{(2\pi\boldsymbol{\Sigma})^{-\frac{p}{2}} \exp(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k))\pi_k}{\sum_{J=1}^{2}(2\pi\boldsymbol{\Sigma})^{-\frac{p}{2}} \exp(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_j))\pi_j}$$

Here p denotes the dimensionality of the $\boldsymbol{X}$ random variable.
The Bayes classifier will be the one that maximizes this ratio for $\boldsymbol{x}_0$.

Two Class problem: Gaussian Distribution-cont.

As before, we focus on the numerator only and obtain the rule:
If

$$\exp(-\frac{1}{2}(\boldsymbol{x}_0-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_0-\boldsymbol{\mu}_1))\pi_1 \geq \exp(-\frac{1}{2}(\boldsymbol{x}_0-\boldsymbol{\mu}_2^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_0-\boldsymbol{\mu}_2))\pi_2$$

then $\boldsymbol{x}_0$ belongs to Class 1, otherwise to class 2.

After taking the log and simplifying the expression, the rule, in its simplest form is:
If

$$(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}\left((\boldsymbol{x}_0-\frac{1}{2}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)\right) \geq (\log\pi_2-\log\pi_1)$$

then $\boldsymbol{x}_0$ belongs to Class 1, otherwise to class 2.

A priori probability

In most cases, we don't have a prior knowledge and therefore $\pi_1 = \pi_2 = \frac{1}{2}$. In the rest of the discussion we make this assumption.

Cost

In some cases, cost of misclassifying an observation from one class is much higher than the other class. In that case we may want to bring cost into consideration in the above model.

However, in the rest of the discussion, we keep the cost of misclassification equal for all classes.

Linear Discriminant - Two Class problem

In general, we know the distributional form of $P(\boldsymbol{X} = \boldsymbol{x}|Y = k)$ but do not know the associated parameters, such as $\boldsymbol{\mu}_k$, $k = 1, 2$ and $\boldsymbol{\Sigma}$. However, we have samples from these Gaussian populations:

$$\boldsymbol{x}_{1,1}, \boldsymbol{x}_{1,2}, \cdots, \boldsymbol{x}_{1,n_1} \text{ from } \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$$

and

$$\boldsymbol{x}_{2,1}, \boldsymbol{x}_{2,2}, \cdots, \boldsymbol{x}_{2,n_2} \text{ from } \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$$

Using these samples, as before, we estimate the means as $\overline{\boldsymbol{x}}_k$, $k = 1, 2$ and the common variance as

$$\boldsymbol{S}^2 = \frac{1}{(n_1 + n_2 - 2)} \left((n_1 - 1)\boldsymbol{S}_1^2 + (n_2 - 1)\boldsymbol{S}_2^2 \right)$$

where $\boldsymbol{S}_1^2$ and $\boldsymbol{S}_2^2$ are sample covariance matrices from sample 1 and sample 2, respectively.

Thus, the decision rules will be: If

$$(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)^T \boldsymbol{S}^{-2} \left((\boldsymbol{x}_0 - \frac{1}{2}(\overline{\boldsymbol{x}}_1 + \overline{\boldsymbol{x}}_2) \right) \geq 0$$

then $\boldsymbol{x}_0$ belongs to Class 1, otherwise to class 2.

Where is the boundary?

When

$$-\frac{1}{2}\left[(\boldsymbol{x}_0 - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_0 - \boldsymbol{\mu}_1)\right] = -\frac{1}{2}\left[(\boldsymbol{x}_0 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_0 - \boldsymbol{\mu}_2)\right]$$

This simplifies to

$$\boldsymbol{x}_0^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \frac{1}{2}\left[\boldsymbol{\mu}_1^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_2\right]$$

Case of $p = 1$

$$\boldsymbol{x}_0^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \frac{1}{2}\left[\boldsymbol{\mu}_1 \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_2\right]$$

reduces to

$$\frac{x_0(\mu_1 - \mu_2)}{\sigma^2} = \frac{1}{2}\frac{\mu_1^2 - \mu_2^2}{\sigma^2}$$

or

$$x_0 = \frac{1}{2}(\mu_1 + \mu_2)$$

Case when $K > 2$

Above concepts are easily extended when the number of classes is more than 2. For example, in general, the decision rule will be:

If

$$(\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_k)^T \boldsymbol{S}^{-2} (\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_k)$$

is smallest then assign $\boldsymbol{x}_0$ to class k.

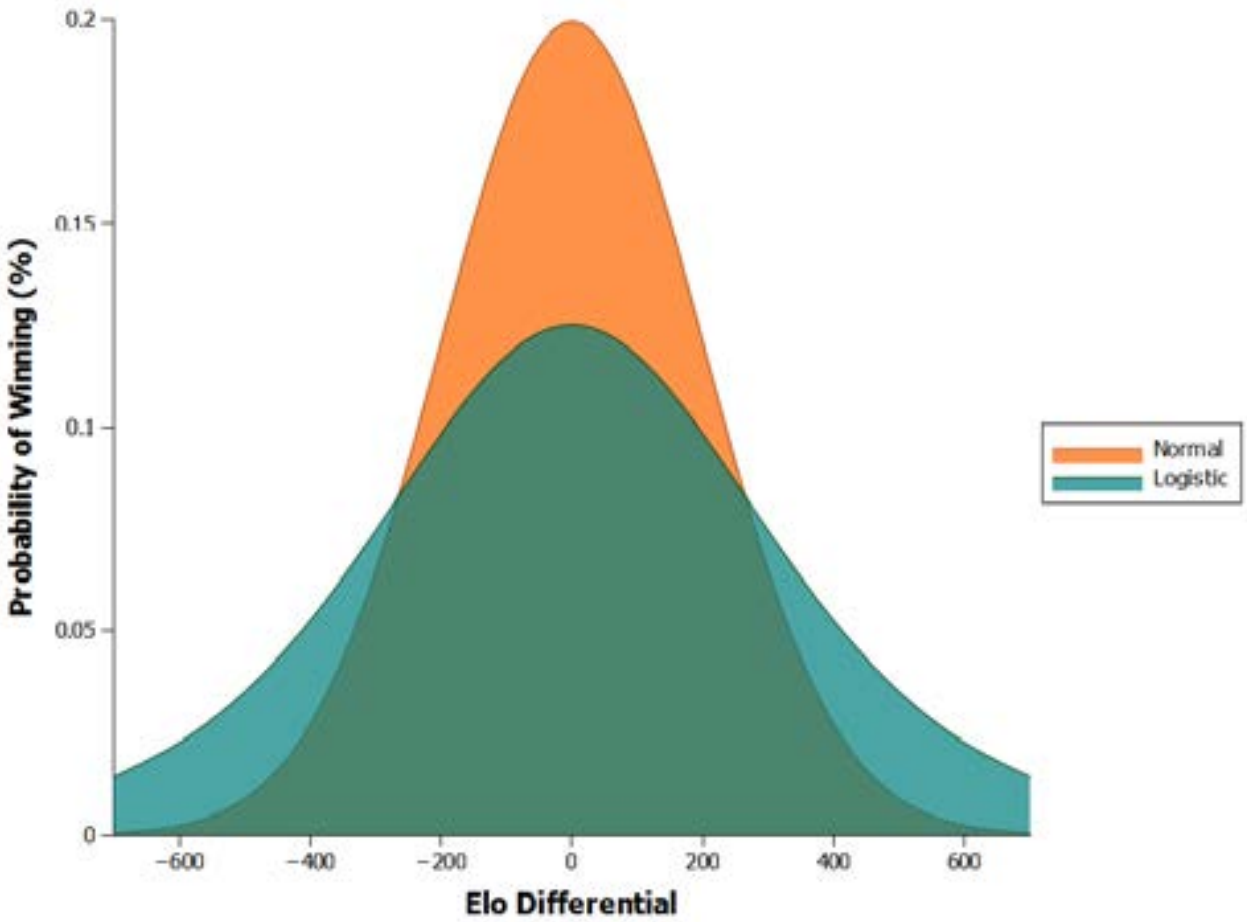
In this case the boundaries are given by:

$$(\boldsymbol{x}_0^T \boldsymbol{S}^{-2} \overline{\boldsymbol{x}}_j) - \frac{1}{2}(\overline{\boldsymbol{x}}_j^T \boldsymbol{S}^{-2} \overline{\boldsymbol{x}}_j) = (\boldsymbol{x}_0^T \boldsymbol{S}^{-2} \overline{\boldsymbol{x}}_k) - \frac{1}{2}(\overline{\boldsymbol{x}}_k^T \boldsymbol{S}^{-2} \overline{\boldsymbol{x}}_k), \quad k \neq j$$

For example, when $K = 3$, there will be three boundaries separating the 3 classes.

LOGISTIC REGRESSION

Logistic Distribution



Logistic Regression:Two class Problem

In linear regression, the outcome Y is continuous, and we set

$$Y = \beta_0 + \boldsymbol{\beta}_1^T \boldsymbol{X} + \epsilon$$

However, this does not work for classification since Y can only be 0 or 1. In other words, Y is a Binomial random variable with probability of success $\theta = P(Y = 1)$ and probability of failure $1 - \theta = P(Y = 0)$. But θ depends on $\boldsymbol{X}$. So we model it as:

$$\theta(\boldsymbol{x}) = P(Y = 1 | \boldsymbol{X} = \boldsymbol{x}) = \frac{\exp(\beta_0 + \boldsymbol{\beta}_1^T \boldsymbol{X})}{1 + \exp(\beta_0 + \boldsymbol{\beta}_1^T \boldsymbol{X})}$$

Note that the above ratio will always take a value between 0 and 1, hence suitable to model θ.

Decision Rule

If

$$\frac{\exp(\beta_0 + \boldsymbol{\beta}_1^T \boldsymbol{X})}{1 + \exp(\beta_0 + \boldsymbol{\beta}_1^T \boldsymbol{X})} > 0.75$$

then $\boldsymbol{x}_0$ belongs to class 1, else class 0.

Note that the relationship is still represented using a linear function of $\boldsymbol{x}$ as $\beta_0 + \boldsymbol{\beta}_1^T \boldsymbol{X}$. This is an example of Generalized Linear Models.

Odds Ratio

The odds ratio

$$odds = \frac{P(Y = 1|\boldsymbol{X} = \boldsymbol{x})}{P(Y = 0|\boldsymbol{X} = \boldsymbol{x})} = \exp(\beta_0 + \boldsymbol{\beta}_1^T \boldsymbol{X})$$

Equivalently,

$$\log(odds) = (\beta_0 + \boldsymbol{\beta}_1^T \boldsymbol{X})$$

Hence, in a logistic regression, the log odds is a linear function of $\boldsymbol{X}$.

Logistic Decision Rule: $K > 2$

In this case, Y takes more than two values. Therefore, instead of binomial distribution, we can use multinomial distribution. In other words, we model $\theta_k = P(Y = k)$ as:

$$P(Y = k|\boldsymbol{X} = \boldsymbol{x}) = \frac{\exp(\beta_{0,k} + \boldsymbol{\beta}_{1,k}^T \boldsymbol{x})}{1 + \sum_{j=1}^{K-1} \exp(\beta_{0,j} + \boldsymbol{\beta}_{1,j}^T \boldsymbol{x})}$$

for $k = 1, \cdots, K-1$ and

$$P(Y = K|\boldsymbol{X} = \boldsymbol{x}) = \frac{1}{1 + \sum_{j=1}^{K-1} \exp(\beta_{0,j} + \boldsymbol{\beta}_{1,j}^T \boldsymbol{x})}$$

Parameter Estimation

In most cases, unknown parameters are estimated using the *Maximum Likelihood Method*

In the context of Gaussian distribution, it was almost obvious that population means should be estimated by the sample means and population covariance matrices by sample covariance matrices. But it is not obvious how to estimate the parameters $\beta_{0,j}, \boldsymbol{\beta}_{1,j};\ j = 1, \cdots, K-1$, of the logistic distribution. They are estimated by using the maximum likelihood approach.

Details ignored here.

Example:LDA, Iris Data

The Iris dataset contains 150 samples of iris flowers from three different species: Setosa, Versicolor, and Virginica. Each sample includes four features: sepal length, sepal width, petal length, and petal width. The goal is to build a classifier that can predict the species of an iris flower based on these four measurements.

Examples: Using **R**: IRIS data

R-code:

```
data(iris)
# partition IRIS data in two sets – training set (70%) and test set (30%)
set.seed(1)
sample <- sample(c(TRUE,FALSE), nrow(iris),replace = TRUE, prob =
c(0.7,0.3))
train <- iris[sample, ]
test <-iris[!sample,]
# modeling the LDA
library(MASS)
model <- lda(Species ., data = train)
```

Example (cont.)

RESULTS:
model
Call:
lda(Species ., data = train)
Prior probabilities of groups: (Although there were 50 observations for each group, sampling makes minor difference)
setosa versicolor virginica
0.3207547 0.3207547 0.3584906
Group means: (The average of each predictor within each class)

Species	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
setosa	4.982353	3.411765	1.482353	0.2411765
versicolor	5.994118	2.794118	4.358824	1.3676471
virginica	6.636842	2.973684	5.592105	2.0552632

These values suggest that the influence on each class. For example, Sepal.Length is most influential in identifying a species. To identify vs setosa Sepal.Width plays an important role.

Example (cont.)

Coefficients of linear discriminants:

predictors	LD1	LD2
Sepal.Length	0.9567859	0.6393462
Sepal.Width	1.3101692	1.6359934
Petal.Length	-2.3090751	-1.5467715
Petal.Width	-2.7028553	3.4538084

In other words, the first linear line (actually a hyper plane) that separates three classes is:

$$0.956 \times \text{Sepal.Length} + 1.310 \times \text{Sepal.Width} - 2.309 \times \text{Petal.Length} - 2.702 \times \text{Petal.Width}$$

Similarly, the second line is defined as:

$$0.639 \times \text{Sepal.Length} + 1.635 \times \text{Sepal.Width} - 1.546 \times \text{Petal.Length} + 3.4538084 \times \text{Petal.Width}$$

Proportion of trace:
LD1 LD2
0.9921 0.0079

Performance on Test Set

Homework : find the performance of this model on the test set, as described in the following example.

Logistics Regression; dataset 'Default'

This dataset has three predictor variables:

1. *student*: Indicates whether or not an individual is a student.
2. *balance*: Average balance carried by an individual.
3. *income*: Income of the individual.

and the dependent variable *default* indicates whether or not an individual defaulted.

We will use student status, bank balance, and income to build a logistic regression model that predicts the probability that a given individual defaults.

The dataset, consisting of 1000 observations is divided in two pars – 70% in the training set and the remaining 300 in the test set.

R-Code and Results

```
set.seed(1)
#Use 70% of dataset as training set and remaining 30% as testing set
sample <- sample(c(TRUE, FALSE), nrow(data), replace=TRUE,
                 prob=c(0.7,0.3))
train <- data[sample, ]
test <- data[!sample, ]
#fit logistic regression model
model <- glm(default student+balance+income, family="binomial",
             data=train)
```

R-Code and Results

Coefficients:

	Estimate	Std. Error	z value	$\Pr(>\|z\|)$
Intercept	-11.478	0.623	-18.412	2×10^{-14}
studentYes	-0.493	0.286	-1.726	0.084
balance	0.006	0.0003	20.384	2×10^{-14}
income	7.8×10^{-6}	9.9×10^{-5}	0.788	0.4304

Number of Fisher Scoring iterations: 8

The coefficients in the output indicate the average change in log odds of defaulting. For example, a one unit increase in balance is associated with an average increase of 0.006 in the log odds of defaulting.

R-Code and Results cont.

The p-values in the output also give us an idea of how effective each predictor variable is at predicting the probability of default:

- P-value of student status: 0.0843
- P-value of balance: is almost 0.0
- P-value of income: 0.43

We can see that balance and student status seem to be important predictors since they have low p-values while income is not nearly as important.

R-Code and Results cont.

We can also compute the importance of each predictor variable in the model

predictor	**Importance**
studentYes	1.726393
balance	20.383812
income	0.788449

Higher values indicate more importance. These results match up nicely with the p-values from the model. Balance is by far the most important predictor variable, followed by student status and then income.

Using the Model to Make Predictions

As an illustration – We use the fitted model to make predictions about whether or not an individual will default based on its (student status, balance, and income).
Consider two individuals:

1. (balance, income, student) = (1400, 2000, Yes). Then the predicted prob. of default = 0.0273
2. (balance, income, student) = (1400, 2000, No). Then the predicted prob. of default = 0.0439

R-Code to calculate probability of default for each individual in test dataset is:
predicted $<-$ predict(model, test, type="response")

Model Diagnostics: Model's Performs on the Test Dataset

Default Probability

By convention, an individual with Prob.('default') > 0.5 will be predicted to default.

However, we can find the optimal probability to use to maximize the accuracy of our model by using the *optimalCutoff() function*. For this dataset,*optimal cut-off = 0.5451712*.

Thus, an individual with a Prob(default) ≥ 0.5451712 will be predicted to default, while any individual with a probability less than this number will be predicted to not default.

Confusion matrix and other measures

Using this threshold, we can create a confusion matrix which shows our predictions compared to the actual defaults. Recall Default $= 1$ (True positive) and Default $= 0$ (True Negative).

	0	1	Total
0	2912	64	2976
1	21	39	60
Total	2933	103	3036

Sensitivity (also known as the true positive rate) $= \frac{39}{39+64} = 0.3786408$
Specificity (also known as the true negative rate) $= \frac{2912}{2912+21} = 0.9928401$
The total misclassification error rate is $\frac{64+21}{3036} = 0.02799$; that is, it is 2.7% for this model.