Seoul National University
Dept. of Civil and Environmental Engineering

Instructor: Junho Song
junhosong@snu.ac.kr

M1586.002500 Information Engineering for CE Engineers

In-Class Material: Class 09

Classification (ISL Chapter 4)

Given: Sample data set $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ where y_i is **qualitative** or **categorical**

Question: How to assign proper category y to a new observation x?

→ "Classification"

1. Classification

(a) **Linear regression**: assumes that the response Y has "**q_____**" value

→ The linear relationship between X and Y, $Y = \beta_0 + \beta_1 X$
→ However, the response Y could be "**qualitative**" or "**categorical**"

e.g.) Eye colors (brown, blue, green…), Diagnoses (seizure, stroke, overdose…)

(b) **Classification**: predicts the qualitative responses

e.g.) Logistic Regression, Linear Discriminant Analysis, K-Nearest Neighbors…

The classification method often predicts the "**p_____**" of each of the categories of a qualitative response
→ Why not using the linear regression for classification purpose?

Example 1 (Classification by Linear Regression): Consider three diagnoses in an emergency room

$$Y = \begin{cases} 1 & \text{if stroke} \\ 2 & \text{if drug overdose} \\ 3 & \text{if eplieptic seizure} \end{cases}$$

Assume the *linear relationship* between X and Y: $Y = \beta_0 + \beta_1 X$

Issues:

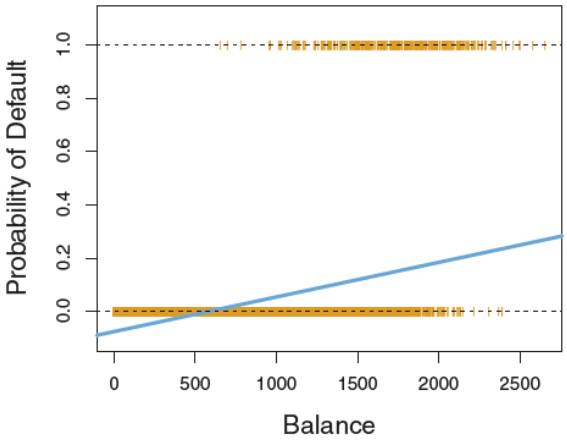
This model insists that the difference between stroke and drug overdose is the same as the difference between drug overdose and epileptic seizure

One can choose the different coding which is equally reasonable
→ 1, 2 and 3 for epileptic seizure, stroke and drug overdose respectively
→ Totally different relationship, and thus different prediction results

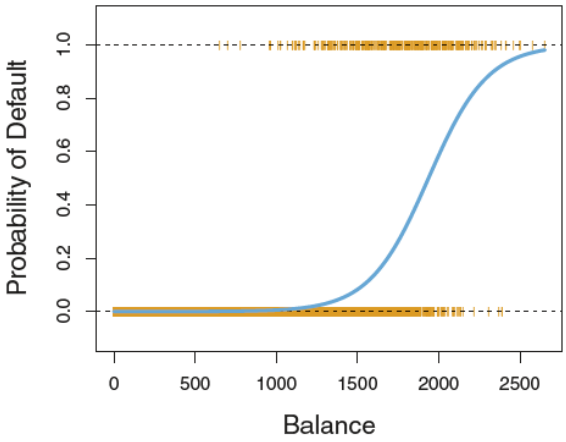
Note: the situation is better when the response is binary → 1) Can use 0/1 coding (the same classification results even if flipped), and 2) classify by $\hat{Y} > 0.5$

Challenge in p_____ classification: the linear regression can have the value outside of $[0, 1]$ (see figure below). Therefore, a different regression function (nonlinear) is desired → **"Logistic Regression"**

Instructor: Junho Song
junhosong@snu.ac.kr



<Linear Regression>



<Logistic Regression>

2. Logistic Regression for 2-Class Classification

(a) Logistic regression model: use the **"Logistic function"**, an S-shaped curve that gives the output in the range of $[0, 1]$, to represent $p(X) = p(Y = y|X)$

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

After a bit of manipulation,

$$\underbrace{\frac{p(X)}{1 - p(X)}}_{\text{"odds"}} = e^{\beta_0+\beta_1 X}$$

This is termed **"odds"**, which can take the value in the range of $[0, \infty]$. Values of the odds close to 0 and ∞ respectively represent low and high probability of $Y = y$, i.e., 0 and 1.

By taking logarithm,

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

The left-hand side is called the **"log-odds"** or **"logit"**
→ The linear regression between X and the logit $\in (-\infty, \infty)$

(b) Nonlinearity of $p(X)$

Same with the linear regression, increasing X by one unit changes the logit by β_1.

However, increasing X by one unit *does not* change $p(X)$ by β_1 because the relationship between $p(X)$ and X is nonlinear (S-shaped curve). The *rate* of change in $p(X)$ depends on the current value of X.
- $\beta_1 > 0$: increasing X is *associated* with increasing $p(X)$
- $\beta_1 < 0$: increasing X is *associated* with decreasing $p(X)$

Instructor: Junho Song
junhosong@snu.ac.kr

(c) Estimation of β_0 and β_1 : Least Square Estimation (LSE) cannot be used due to the nonlinearity → use **"Maximum Likelihood Estimation (MLE)"**

Basic intuition: Trying to find $\hat{\beta}_0$ and $\hat{\beta}_1$ such that plugging these estimates into the logistic function, yields a number close to response data in the train dataset.

In case of binary response, the **"L_____ function"** can be written as:

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i\prime:y_{i\prime}=0} (1 - p(x_{i\prime}))$$

Maximum likelihood estimation is general approach that can be used to fit the nonlinear models (LSE is the special case of MLE)

(d) Hypothesis test

Same hypothesis test as the linear regression is used
- Null Hypothesis H_0: There is no relationship between X and $p(X)$, $\beta_1 = 0$
- Alternative Hypothesis H_1: There is some relationship between X and $p(X)$, $\beta_1 \neq 0$

"Z-statistics" plays the same role as the *t*-statistic in the linear regression output
e.g.) Z-statistics associated with β_1 is equal to $\hat{\beta}_1 / SE(\hat{\beta}_1)$

	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

(e) Logistic regression for more than 2 classes: Multiple-class extensions of logistic regression is *possible* but not often used in practice

3. Multiple Logistic Regression

(a) Generalized equation of logit for multiple linear equation

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

The logistic function of **"Multiple predictors"** are be derived as

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

where $\beta_0, \beta_1, \cdots, \beta_p$ can be estimated by maximum likelihood estimation

Instructor: Junho Song
junhosong@snu.ac.kr

(b) **Confounding**

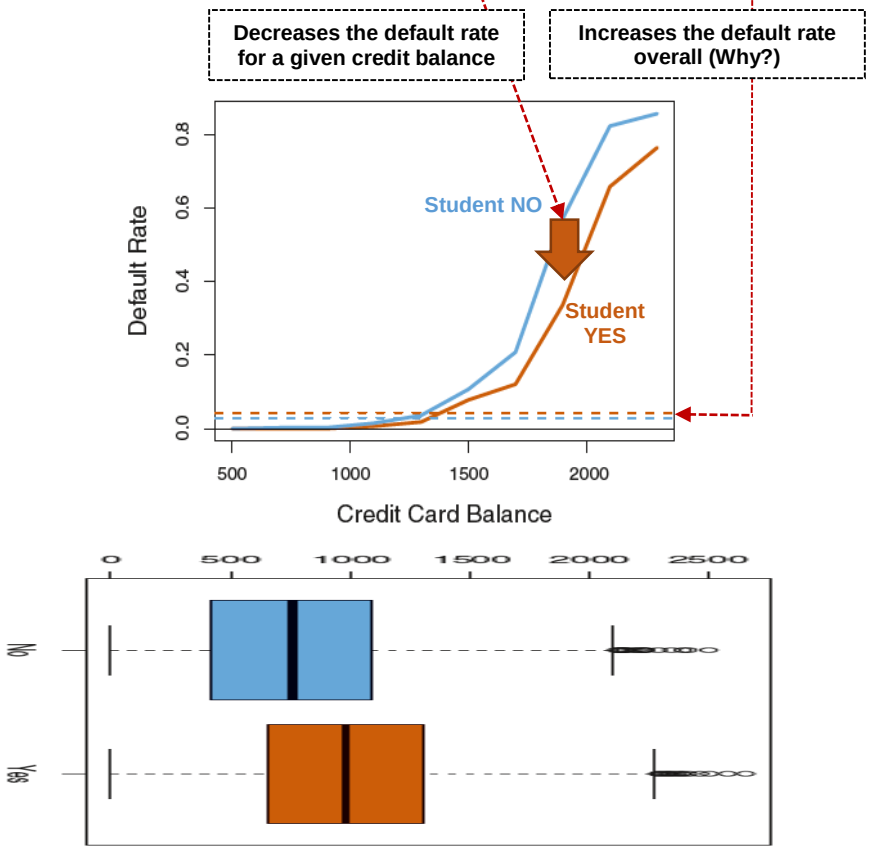
In many cases, one predictor is correlated with the other predictors.

As in the linear regression setting, the results obtained using one predictor may be different from those obtained using multiple predictors with correlation.

The phenomenon is known as **"Confounding"**

	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

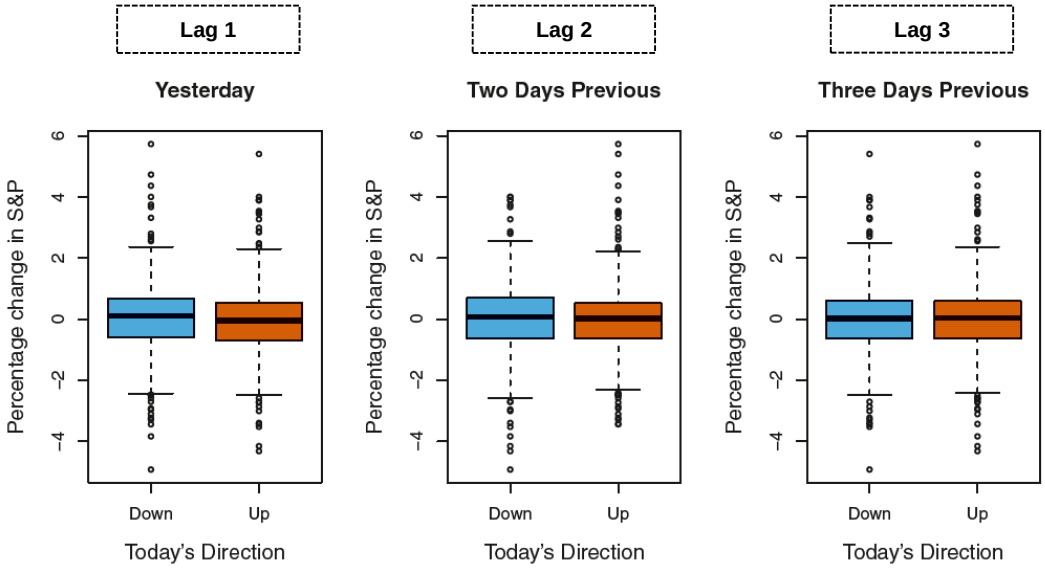
	Coefficient	Std. error	Z-statistic	P-value
Intercept	−10.8690	0.4923	−22.08	<0.0001
balance	0.0057	0.0002	24.74	<0.0001
income	0.0030	0.0082	0.37	0.7115
student[Yes]	−0.6468	0.2362	−2.74	0.0062

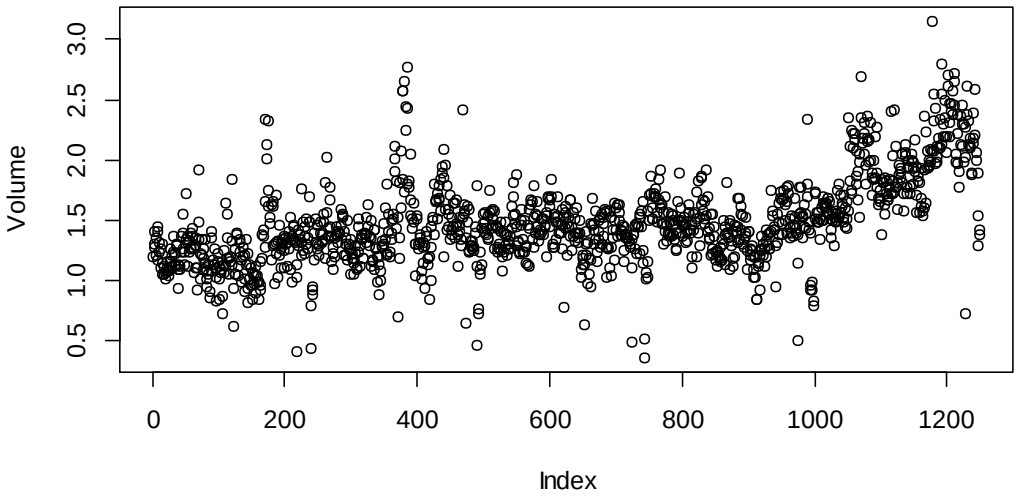


Instructor: Junho Song
junhosong@snu.ac.kr

```
library(ISLR) # ISLR library
names(Smarket) # Variable names in 'Smarket' data
dim(Smarket) # Dimension of 'Smarket' data
summary(Smarket)
pairs(Smarket) # plot the scatterplot matrices

cor(Smarket) # Error message due to the categorical variable "Direction"
cor(Smarket[, -9])
# Correlation between variables
# only pair showing significant correlation: year and volume

attach(Smarket)
plot(Volume)
# plot the 'Volume' in 'Smarket' data
```

Smarket in ISLR package: daily movements in the S&P 500 stock index from 2001 to 2005





Instructor: Junho Song
junhosong@snu.ac.kr

- Training data: 2001~2005 (entire set)
- Predictors: Lag1, Lag2, Lag3, Lag4, Lag5, Volume
- Response: Direction (“Up” = 1)

```
# train the model with the whole dataset
glm.fit = glm(Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 + Volume,
              data = Smarket, family = binomial)
contrasts(Direction) # # 1 for 'Up' category
summary(glm.fit)
# glm: Generalized Linear Models (includes logistic regression)
# 'binomial' option makes glm use logistic regression
# fit data to glm
# Lag1 (return on the previous day) shows smallest p (but weak)

coef(glm.fit) # get the coefficients of fitted model
summary(glm.fit)$coef
summary(glm.fit)$coef[, 4] #p-values only

glm.probs = predict(glm.fit, type = "response")
# predict the probability that the market will go up
# type = "response" option tells R to output probabilities of the form
   P(Y=1|X)
glm.probs[1:10] # show the first ten probabilities only

glm.pred = rep("Down", 1250) # create a vector of 1,250 'Down' elements
glm.pred[glm.probs > 0.5] = "Up"
# replace by "Up" if the predicted probability is greater than 0.5
glm.pred[1:10]

table(glm.pred, Direction) # create confusion matrix
(507 + 145)/1250
mean(glm.pred == Direction)
# calculate the rate of correctly predicted the movement of market
```

- Training data: 2001~2004
- Test data: 2005
- Predictors: Lag1, Lag2, Lag3, Lag4, Lag5, Volume
- Response: Direction (“Up”= 1)

```
# train the model with the dataset before 2005
# repeat the same process with new dataset
train = (Year < 2005)
Smarket.2005 = Smarket[!train, ] # use only dataset from 2005
dim(Smarket.2005) # dimension of new dataset
Direction.2005 = Direction[!train] # new 'Direction' variable

glm.fit = glm(Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 + Volume,
              data = Smarket, family = binomial, subset = train)
# training using the data before 2005 (using the 'subset' option)

glm.probs = predict(glm.fit, Smarket.2005, type = "response")
# prediction for dates in 2005

glm.pred = rep("Down", 252)
glm.pred[glm.probs > 0.5] = "Up"
table(glm.pred, Direction.2005)
mean(glm.pred == Direction.2005)
mean(glm.pred != Direction.2005)
```

Instructor: Junho Song
junhosong@snu.ac.kr

- Training data: 2001~2004
- Test data: 2005
- Predictors: Lag1, Lag2

```
# train the model with the part of predictors ('Lag1' and 'Lag2')
# repeat the same process with new dataset
glm.fit = glm(Direction ~ Lag1 + Lag2, data = Smarket,
              family = binomial, subset = train)
glm.probs = predict(glm.fit, Smarket.2005, type = "response")
glm.pred = rep("Down", 252)
glm.pred[glm.probs > 0.5] = "Up"
table(glm.pred, Direction.2005)
mean(glm.pred == Direction.2005)
106/(106 + 76)
# calculate the accuracy rate when the model predicts an increase in the
   market

# predict the 'Direction' associated with particular values of 'Lag1' and
   'Lag2'
# ('Lag1', 'Lag2') = (1.2, 1.1) & (1.5, -0.8)
predict(glm.fit,
        newdata = data.frame(Lag1 = c(1.2, 1.5), Lag2 = c(1.1, -0.8)),
        type = "response")
```

Seoul National University
Dept. of Civil and Environmental Engineering

Instructor: Junho Song
junhosong@snu.ac.kr

M1586.002500 Information Engineering for CE Engineers
In-Class Material: Class 10
Classification (ISL Chapter 4)

1 (Review) Bayes classifier

(a) Bayes classifier assigns each observation x to most likely class
→ Assigns x to class k when the conditional probability $\Pr(Y = k|X = x)$ is the largest

From the Bayes' theorem,

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

- $\Pr(Y = k) = \pi_k$: The *prior* probability of kth class:
- $\Pr(X = x|Y = k) = f_k(x)$: The conditional density function of x if observation comes from the kth class

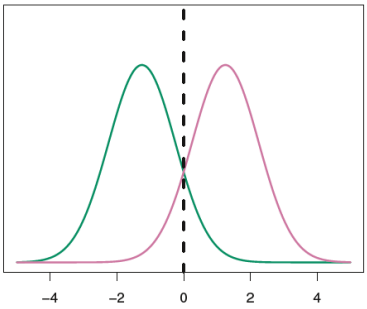
(b) "Linear discriminant analysis (LDA)" : an approximate of the Bayes classifier, i.e. $\pi_k \to \hat{\pi}_k$, $f_k(x) \to \hat{f}_k(x)$

Assumptions: The data within each class are normally distributed and share common covariance $\boldsymbol{\Sigma_k = \Sigma}$.

(c) LDA performs better than logistic regression when:

- the response classes are well-separated (logistic regression is often unstable)
- the predictors X are approximately normal in each of the classes
- more than two response classes are of interest

2 Linear Discriminant Analysis (LDA) for $p = 1$



(a) Bayes' theorem:

Density function $f_k(x)$ for univariate case ($p = 1$):

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu_k)^2}{2\sigma^2}\right)$$

Conditional probability of k, $p_k(x) = \Pr(Y = k|X = x)$

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

$$\propto \pi_k \exp\left(x\frac{\mu_k}{\sigma^2} - \frac{\mu_k{}^2}{2\sigma^2}\right)$$

Instructor: Junho Song
junhosong@snu.ac.kr

(b) Bayes classifier: Find the k with largest $p_k(x)$ (or equivalently, $\log p_k(x)$)

"Discriminant function" can be derived as

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

→ Find the k with largest discriminant value $\delta_k(x)$

(c) "Decision boundary" is a set of values x for which $\delta_l(x) = \delta_k(x)$

Example 1 (Bayes classifier for 2-class problem): Identify the Bayes decision boundary with $K = 2$ and $\pi_1 = \pi_2 = 0.5$. Suppose that the probability densities for each class are $N(-1.25,1^2)$ and $N(1.25,1^2)$

From $\delta_1(x) = \delta_2(x)$,

$$x = \frac{\mu_1^2 - \mu_2^2}{2(\mu_1 - \mu_2)} = \qquad\qquad =$$

See Figure on Page 1

(d) Estimates of model parameters μ_k, σ^2, and π_k using the training data set $\{(x_i, y_i)\}_{i=1\ldots n}$

Means: average of training observations in the kth class

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

Common variance: weighted average of sample variances in each class

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

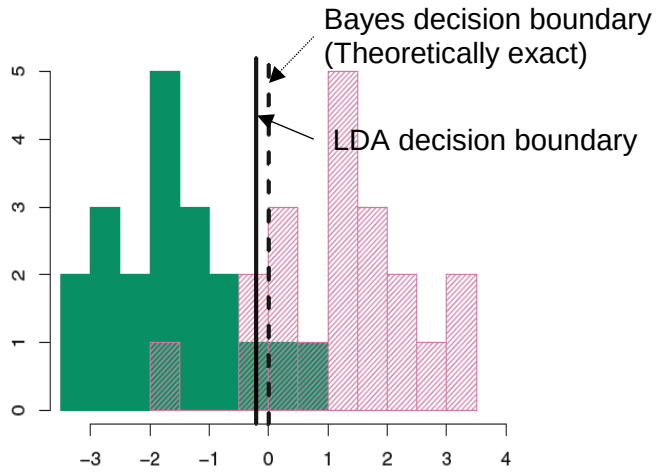
Class membership probabilities: proportions of samples that belong to the k-th class

$$\text{the prior knowlege (if available)} \quad \text{OR} \quad \hat{\pi}_k = n_k/n$$

→ Discriminant functions:

$$\hat{\delta}_k(x) = x\frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{{\hat{\mu}_k}^2}{2\hat{\sigma}^2} + log(\hat{\pi}_k)$$

Note: "*Discriminant function*" is a "*l*____" function of x → *Linear Discriminant Analysis*

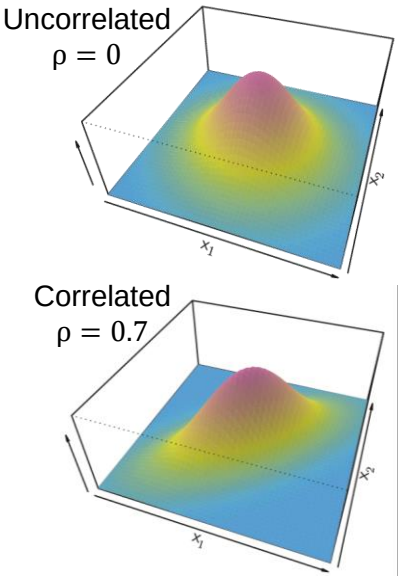


Instructor: Junho Song
junhosong@snu.ac.kr

(e) Procedures of LDA:

Estimate the parameters $\hat{\mu}_k$, $\hat{\sigma}^2$, and $\hat{\pi}_k$ → (Find the decision boundary) → Assign new observation x to the class k for which $\hat{\delta}_k$ is the largest

3 Linear Discriminant Analysis (LDA) for multiple predictors ($p > 1$)



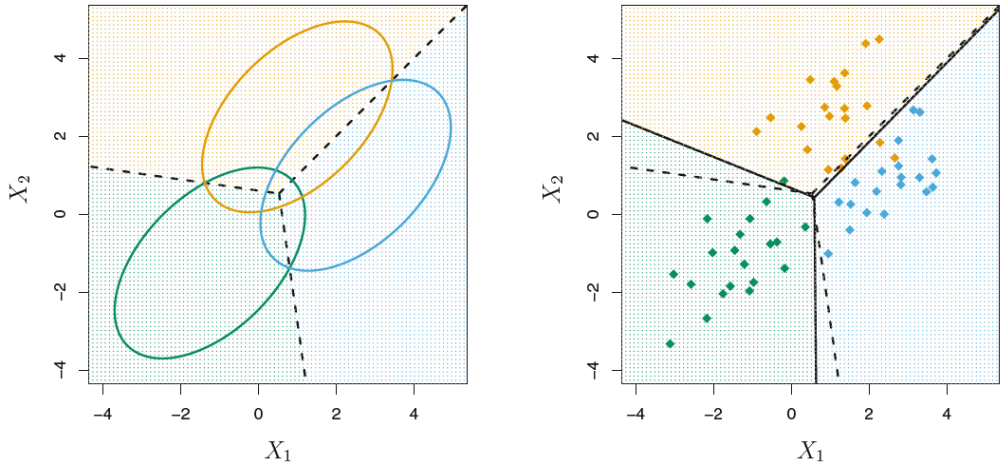
(a) Distribution function of x in kth class:
Multivariate Gaussian (multivariate normal) distribution

$$f_k(x) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\mu_k)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\mu_k)\right)$$

(b) Bayes classifier for multiple predictors

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

(c) "Decision boundaries" are a set of values x for which $\delta_l(x) = \delta_k(x)$ (→ Boundaries are "linear" functions of x)




```
library(ISLR) # ISLR library for stock market data
attach(Smarket) # add variables in 'Smarket' to the search path

train=(Year<2005) # Index of the train data set

Smarket.2005=Smarket[!train,] # Train data set: variables
Direction.2005=Direction[!train] # Train data set: responses (UP/DOWN)

library(MASS) # MASS library for LDA
lda.fit=lda(Direction~Lag1+Lag2,data=Smarket,subset=train)
# Fit the Gaussian parameters and find the LDA boundaries

lda.fit # Show information about the model
plot(lda.fit, type="both") # plot the linear part of discriminant for each
  group
```

Instructor: Junho Song
junhosong@snu.ac.kr

```
Call:
lda(Direction ~ Lag1 + Lag2, data = Smarket, subset = train)

Prior probabilities of groups:
    Down       Up
0.491984 0.508016
```

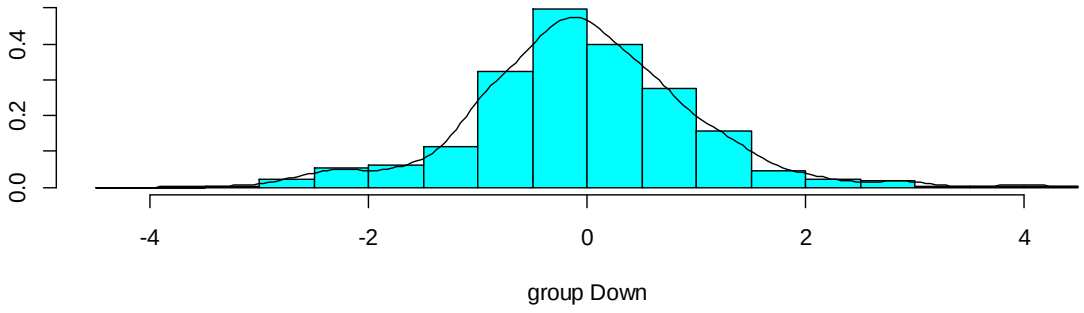
π_1 π_2

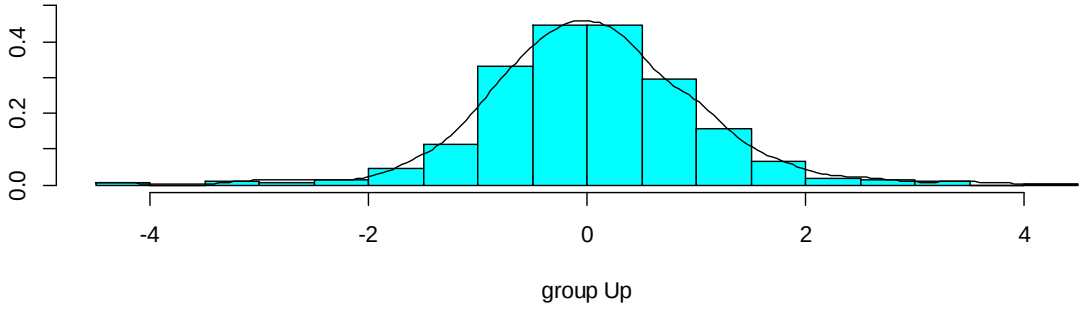
```
Group means:
            Lag1         Lag2
Down  0.04279022  0.03389409
Up   -0.03954635 -0.03132544


Coefficients of linear discriminants:
             LD1
Lag1 -0.6420190
Lag2 -0.5135293
```

μ_1 μ_2

$$\delta_2(x) = -0.642\text{Lag1} - 0.514\text{Lag2} + \cdots \propto p_2(x)$$

`plot(lda.fit)` provides histograms of the linear part of $\delta_2 = \delta_{up}$ for each training group – the larger δ_2 is, the more likely it is for the classifier predicts the group 2, i.e. “Up”





Instructor: Junho Song
junhosong@snu.ac.kr

Example 2 (Type of errors): Predict whether or not an individual will default on the basis of credit card balance and student status.

Suppose the training error rate of LDA result is 2.5%. Is this accuracy sufficient?

Followings should be noted:

1) Test error rate is usually higher than the training error (beware overfitting)

2) Consider the case in which the proportion of defaulted individuals is extremely small. For example, when only 3% among the samples are default, then simple *null* classifier (i.e. assigning to "no default" every time) will achieve 97% accuracy. Is it a desired result?

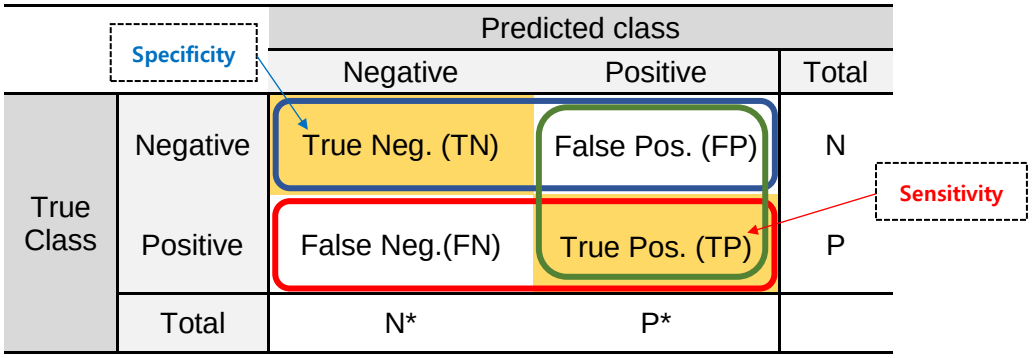
Example Result: a *confusion matrix*

		Predicted default state		
		No	Yes	Total
True default state	No	9,644	23	9,667
	Yes	252	81	333
	Total	9,896	104	10,000

Overall error rate=2.75%
Error rate among whom did not default: 23/9,667=0.2%
Error rate among whom defaulted: 252/333=75.7% (NG)

4 Analysis of Classification Results:

(a) "Confusion matrix" gives the important measures for classification testing:



- False positive rate (FP/N) : Type I error, or (1–**specificity**)
- True positive rate (TP/P) : (1–Type II error), **sensitivity**, power, recall, hit rate
- Positive predictive value (TP/P*) : precision
- Negative predictive value (TN/N*)
- Accuracy ((TN+TP)/total): (1–overall error rate)

Instructor: Junho Song
junhosong@snu.ac.kr

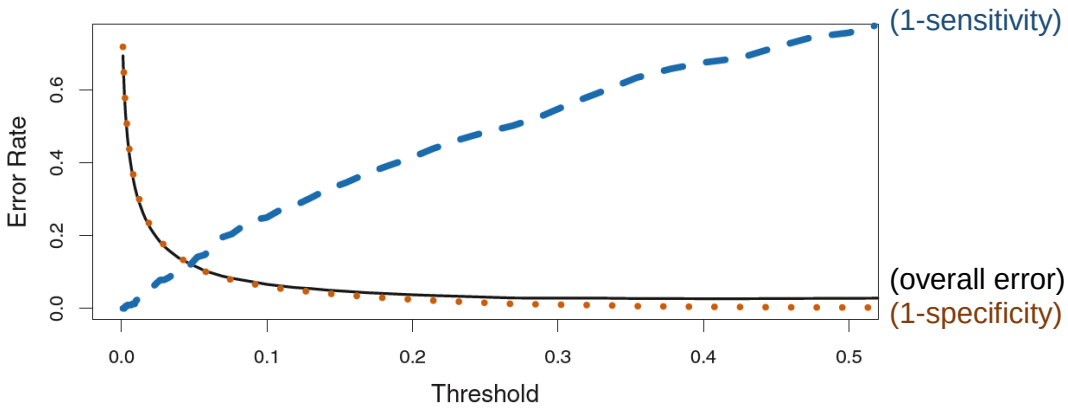
(b) Modification of LDA

LDA aims to minimize the "overall error rate" regardless of the type of errors. However, in practical case in which both a high "sensitivity" and "specificity" are desired, classifier could be modified.

For example, in the two-class case, class 1 is selected if $p_1(x) > p_2(x)$ or equivalently, $p_1(x) > 0.5$. Threshold other than 0.5 could be selected to have better sensitivity, such as $p_1(x) > 0.2$

(c) Trade-off between "sensitivity" and "overall accuracy" (Example 2 & 3)

Note: $N \gg P$ → minimizing overall error leads to a small value of (1-specificity)



→ Threshold should be decided based on *domain knowledge*

```
lda.pred=predict(lda.fit, Smarket.2005)
# Prediction of class for the Train set
lda.class=lda.pred$class # Predicted results for the train set (UP/DOWN)
table(lda.class,Direction.2005) # Confusion matrix
mean(lda.class==Direction.2005) # Overall accuracy
```

Example 3 (Modified LDA): Predict whether or not an individual will default on the basis of credit card balance and student status.

In Example 2, an observation was assigned to default class if $\Pr(default = Yes|X = x) > 0.5$. Consider the case $\Pr(default = Yes|X = x) > 0.2$ is used instead.

Example Result: a *confusion matrix*

		Predicted default state		
		No	Yes	Total
True default state	No	9,432	235	9,667
	Yes	138	195	333
	Total	9,570	430	10,000

Instructor: Junho Song
junhosong@snu.ac.kr

A. Overall error rate=3.7% (0.9% ↑)
B. Error rate among whom did not default: 235/9,667=2.4% (2.2% ↑) , "specificity" worsens
C. Error rate among whom defaulted: 138/333=41.4% (34.3% ↓), "sensitivity" improved

(d) "ROC (Receiver Operating Characteristic) curve" : a display of the two types of the error metrics for different model parameters

The overall performance of a classifier can be quantified by the "area under the ROC curve (AUC)"

→ the larger AUS, the better the classifier