

데이터 과학 기초

# 04

## 로지스틱 회귀와 분류

경북대학교 배준현 교수  
(joonion@knu.ac.kr)



## 04. 로지스틱 회귀와 분류

### ■ 선형모델의 일반화:

- 선형회귀분석을 위한 조건:
  - 결과변수가 연속형 변수이면서 정규분포를 따라야 한다.
- 선형회귀분석을 위한 조건에 맞지 않는 경우:
  - 결과변수가 범주형 변수일 때: 로지스틱 회귀분석
  - 결과변수가 어떤 사건이 발생하는 횟수일 때: 포아송 회귀분석



## 04. 로지스틱 회귀와 분류

- 일반화 선형모델: *generalized* linear model
  - 선형회귀모델을 확장: 정규분포를 따르지 않는 결과변수에 대한 회귀모델 생성
    - 표준 선형회귀모델:  $\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m$ 
      - $\mu_y$ : 결과변수의 조건부 평균,  $x_m$ : 예측변수,  $\beta_m$ : 회귀계수,  $m$ : 변수의 개수
    - 일반 선형회귀모델:  $f(\mu_y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m$ 
      - $f(\mu_y)$ : 결과변수의 조건부 평균의 함수 (*link function*)
  - 표준 선형회귀모델은 일반선형모델의 한 특수한 경우
    - 링크함수가 항등함수:  $f(\mu_y) = \mu_y$
    - 확률분포는 정규분포를 따름
  - 회귀계수의 추정: 최대우도법(*MLE, Maximum Likelihood Estimation*)



## 04. 로지스틱 회귀와 분류

- 일반화 선형모델: *generalized linear model*
  - 로지스틱 회귀분석: *logistic regression* analysis
    - 결과변수가 범주형 변수일 때: 정규분포를 따르지 않음
      - 이분 변수(*binary variable*): 예/아니오, 성공/실패, 생존/사망 등
      - 다중 변수(*multicategory variable*): 우수/보통/미흡, A/B/AB/O 등
  - 포아송 회귀분석: *Poisson regression* analysis
    - 결과변수가 어떤 사건이 발생하는 횟수일 때: 포아송 분포를 따름
      - 연간 철도사고횟수, 월간 빈집털이횟수, 일간 상담횟수 등
      - 횟수변수는 포아송 분포를 따르고, 평균과 분산은 종종 상관관계를 가짐



## 04. 로지스틱 회귀와 분류

- 포아송 회귀분석: *Poisson* regression analysis
  - 결과변수가 특정 기간 동안의 **사건발생횟수**(또는 개수)인 경우에 적용
    - 한 시간 동안 걸려오는 상담전화 횟수
    - 하루 동안 발생하는 범죄 횟수
    - 한 달 동안 발생하는 교통사고 횟수 등
  - 포아송 회귀모델: *Poisson* regression model
    - 링크함수는  $\ln(\lambda)$  이며, 확률분포는 포아송 분포를 따름
    - $\ln(\lambda) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m$ 
      - $\lambda$ : 결과변수  $y$ 의 평균



## 04. 로지스틱 회귀와 분류

- 이항 로지스틱 회귀분석: *binomial* logistic regression analysis
  - 결과변수가 이분형 범주일 때 특정 사건이 발생할 확률을 직접 추정
    - 결과변수의 예측값이 항상 1(사건발생)과 0(미발생) 사이의 확률값
    - 확률값이 0.5보다 크면 사건이 발생, 0.5보다 작으면 발생하지 않음
    - 예) 기업부도가 발생할 확률
  - 로지스틱 변환: *logistic transformation*
    - 예측변수의 선형결합을 로그 변환한 결과변수로 나타냄
  - 이항 로지스틱 회귀모델: binomial logistic regression model
    - $\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m$ 
      - $p$ : 이항 사건의 성공 확률(사건발생),  $1 - p$ : 이항 사건의 실패 확률(미발생)



## 04. 로지스틱 회귀와 분류

### ■ 이항 로지스틱 회귀분석:

- 오즈: *odds*

- $odds = \frac{p}{1-p}$  : 사건 발생확률 대 사건 미발생 확률의 비율

- 로짓(*logit*): 오즈에 로그를 취한 값 =  $\ln\left(\frac{p}{1-p}\right)$

- 로지스틱 회귀모델:

- 로그오즈(log odds=logit)에 대한 선형모델

- 링크함수가 로그오즈이며, 확률분포는 이항분포

- 사건발생확률  $p$ 에 대해서 정리:

- $p = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}}, z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m$

- 회귀계수를 알면 결과변수의 사건발생확률을 구할 수 있음



## 04. 로지스틱 회귀와 분류

### ■ 오즈비: *odds ratio*

- 다른 독립변수가 동일하다는 가정하에서
  - 특정 독립변수 한 단위 증가에 따른
  - 사건 발생확률 대 미발생확률 비율의 **변화율**
- 오즈비는 오즈의 정의로부터 도출 가능
  - $\ln(odds) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m$
  - $odds = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m}$
  - $x_1$  변수의 오즈비  $= \frac{e^{\beta_0 + \beta_1 \times 1 + \beta_2 x_2 + \cdots + \beta_m x_m}}{e^{\beta_0 + \beta_1 \times 0 + \beta_2 x_2 + \cdots + \beta_m x_m}} = \frac{e^{\beta_1 \times 1}}{e^{\beta_1 \times 0}} = e^{\beta_1}$





## 04. 로지스틱 회귀와 분류

### ■ 로지스틱 회귀분석과 예측:

- 로지스틱 회귀계수를 알면 사건발생의 확률을 계산 가능
  - $P(\text{사건발생}) = \frac{1}{1+e^{-z}}$ ,  $z = \beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_mx_m$
- 예측: *prediction*
  - 새로운 예측변수의 값을 입력하면 결과변수값이 1인 확률을 예측가능
  - 훈련 데이터로부터 회귀계수를 학습한 결과로
    - 시험 데이터의 고객이 이탈할 확률을 예측할 수 있음



## 04. 로지스틱 회귀와 분류

- 분류와 군집화: *Classification* .vs. *Clustering*
  - 분류: 지도 학습: 정답이 있는 데이터셋을 분류하는 것
    - 예) iris 데이터셋에서 품종의 분류.
    - 예) titanic 데이터셋에서 생존 여부를 예측.
  - 군집화: 비지도 학습: 정답이 없는 데이터셋을 분류하는 것
    - 예) iris 데이터셋에서 모양이 유사한 꽃들의 군집 찾기
    - 예) titanic 데이터셋에서 서로 가까운 사람들의 군집 찾기



## 04. 로지스틱 회귀와 분류

### ■ 분류기의 종류:

- 로지스틱 회귀분석: Logistic Regression
- 의사결정 트리: Decision Tree
- 랜덤 포리스트: Random Forest
- k-최근접 이웃: kNN, k-Nearest-Neighbor
- 나이브 베이지안: Naïve Bayesian
- 서포트 벡터 머신: SVM, Support Vector Machine



## 04. 로지스틱 회귀와 분류

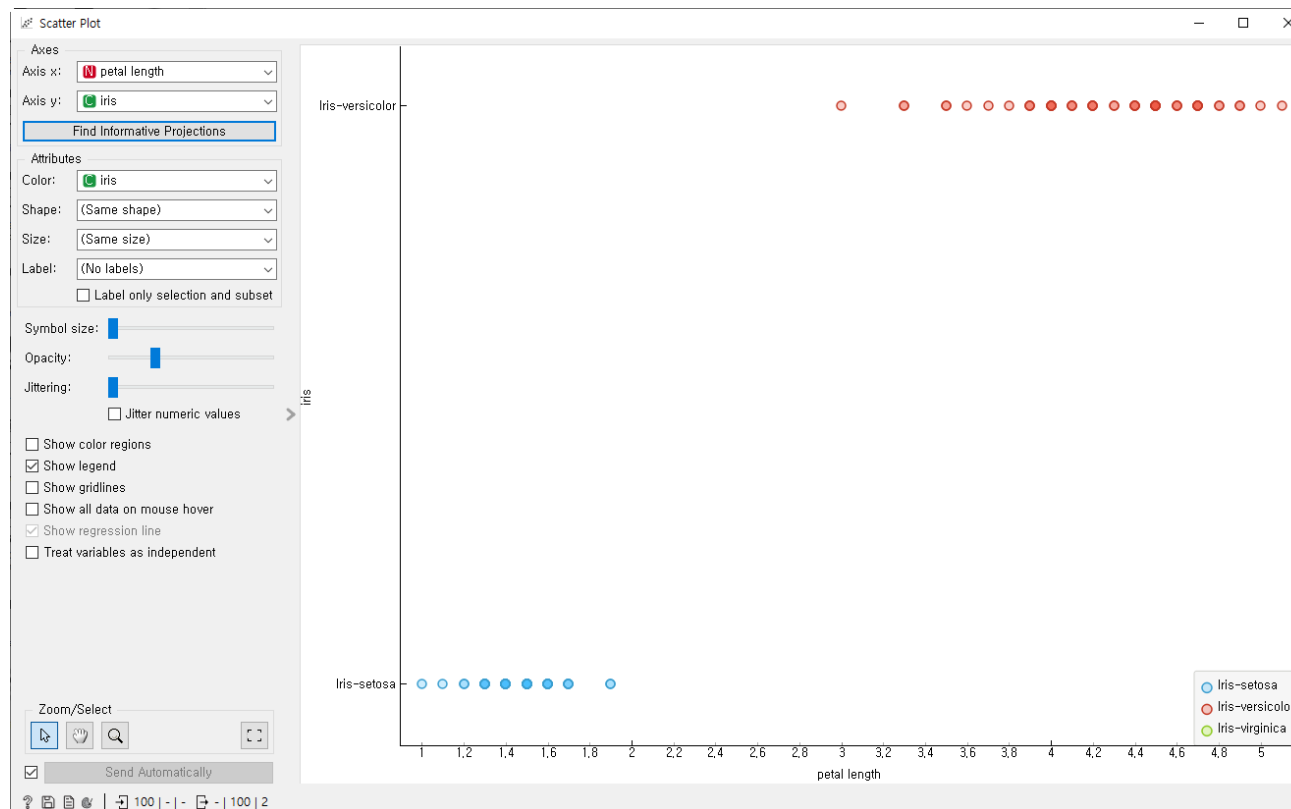
- 이진 분류: *Binary* Classification
  - 분류도 예측의 일종이지만, 종속변수가 범주형 변수
    - 이진 분류: 종속변수 값의 범위가 두 개일 때
    - 예) titanic 데이터셋: survival 변수는 (생존, 사망) 둘 중의 하나
    - 예) 암 진단: 종속 변수가 암에 (걸렸거나, 걸리지 않았거나) 둘 중의 하나



## 04. 로지스틱 회귀와 분류

### ■ 로지스틱 회귀: *Logistic Regression*

- 종속변수의 값이 바이너리 형태인 경우에 적용하기 좋은 회귀 분석 모델
  - 직선으로는 이런 데이터를 잘 설명할 수 없으므로, **적절한 곡선**을 찾아야 함.

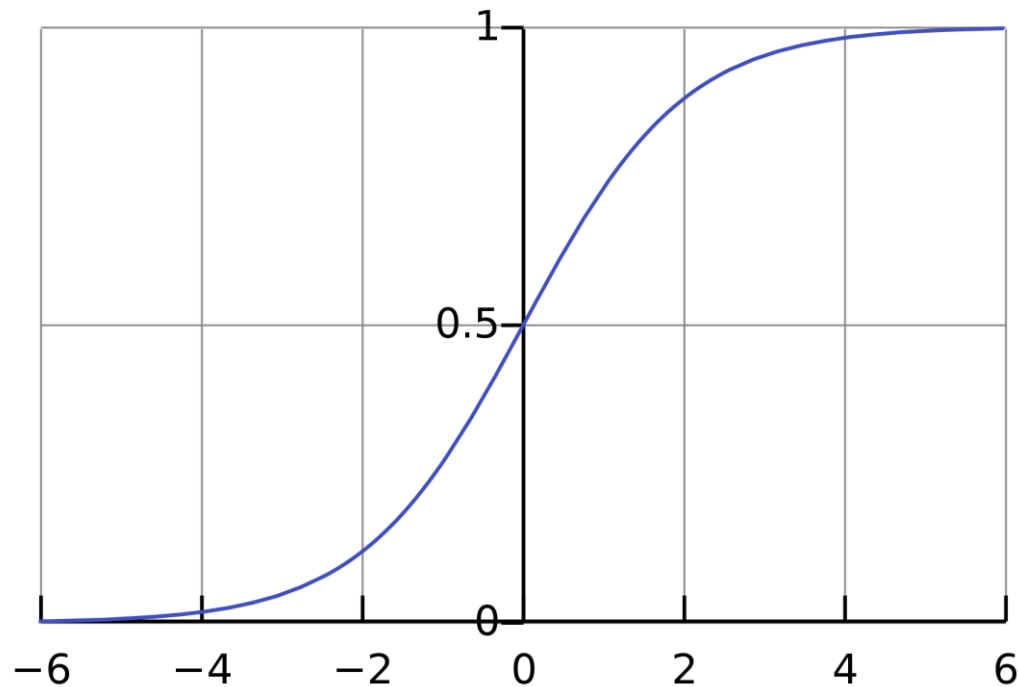




## 04. 로지스틱 회귀와 분류

### ■ 로지스틱 함수: *Logistic* Function

- 바이너리 값을 가지는 범주형 데이터를 잘 설명해 주는 지수 함수
- $y = \frac{1}{1+e^{-x}}$ ,  $e$ 는 자연상수(오일러의 수, 네이피어의 수).





## 04. 로지스틱 회귀와 분류

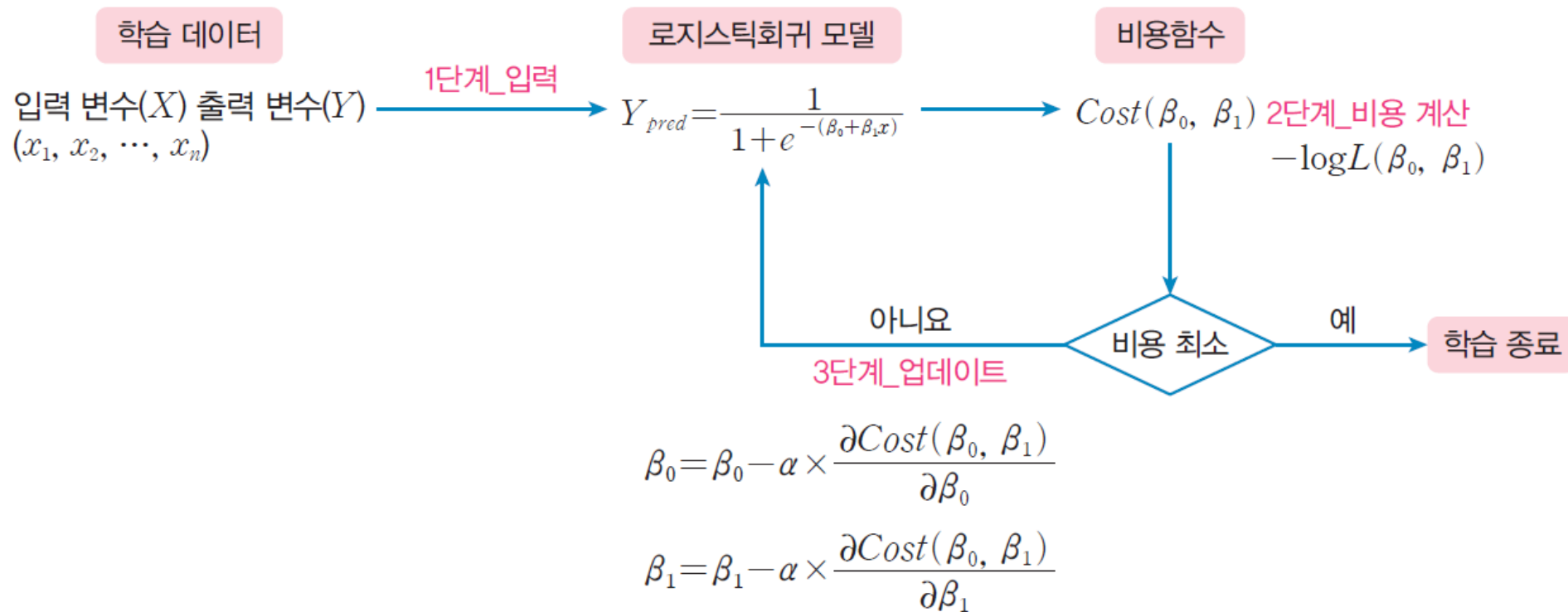
### ■ 로지스틱 함수의 성질과 활용:

- 시그모이드 함수: *Sigmoid* Function
  - *bounded*: 유한한 구간 (a, b) 사이의 **한정된** 값을 갖는다.
  - *monotonic*: 항상 양의 기울기를 가지는 **단조증가** 함수다.
- 로지스틱 함수를 **분류의 기준을 충족할 확률**로 해석
  - $y = \alpha x + \beta, f(x) = \frac{1}{1+e^{-(\alpha x + \beta)}}$
  - $f(x) > 0.5$ :  $y = 1$ 이라고 분류
  - $f(x) < 0.5$ :  $y = 0$ 이라고 분류



## 04. 로지스틱 회귀와 분류

### ■ 로지스틱 회귀식을 학습하기 위한 과정:



출처: 수학과 함께 하는 AI 기초, EBS

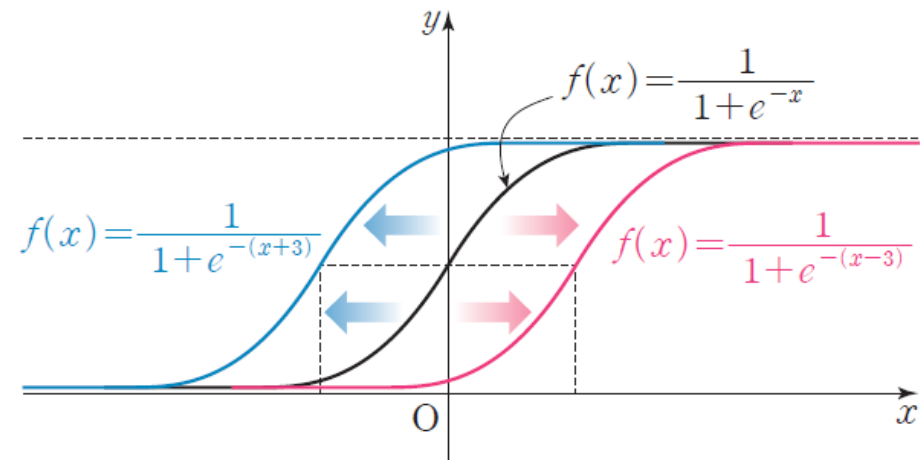
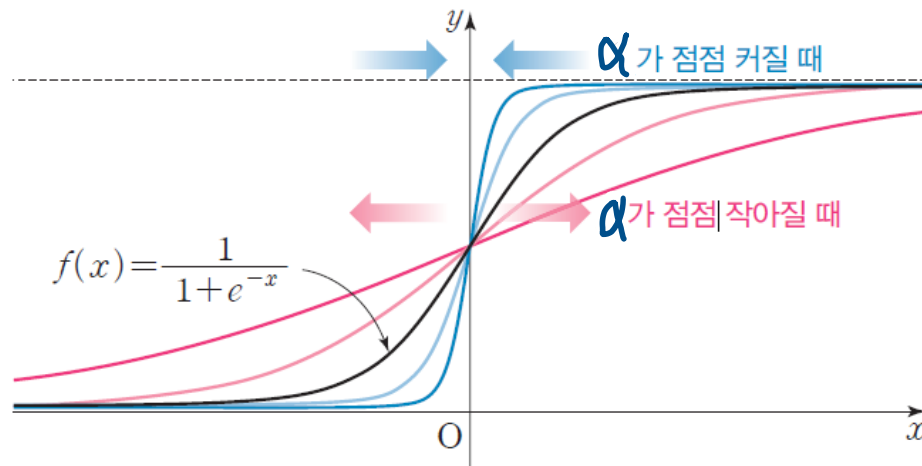




## 04. 로지스틱 회귀와 분류

### ■ 로지스틱 회귀식을 찾는 방법:

- 로지스틱 함수  $f(x)$ 에서 가장 적절한  $\alpha$ 와  $\beta$  찾기
  - 최대우도추정법: MLE, *Maximum Likelihood Estimation*
- 우도 함수: *Likelihood* Function
  - $L(\alpha, \beta) = \prod_{i=1}^n \{f(x_i)\}^{y_i} \{1 - f(x_i)\}^{1-y_i}$





## 04. 로지스틱 회귀와 분류

### ■ R: glm()

```
data(iris)
df <- iris[1:100, ]
df$Species <- as.numeric(df$Species)
head(df)
```

```
model <- glm(Species ~ ., data=df)
model
```

```
> model
```

```
Call: glm(formula = Species ~ ., data = df)
```

```
Coefficients:
```

(Intercept)	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
1.36970	-0.02849	-0.16820	0.20313	0.28785



## 04. 로지스틱 회귀와 분류

```
# Prepare test data
v1 <- c(2.7, 2.4, 1.65, 0.67)
v2 <- c(5.84, 5.48, 3, 2.16)
v3 <- c(3.97, 4.01, 1.7, 0.67)
mat <- matrix(c(v1, v2, v3), nrow=3, ncol=4, byrow=TRUE)
df.test <- data.frame(mat)
colnames(df.test) <- colnames(iris[1:4])

pred <- predict(model, df.test)
df.test$pred = round(pred, 0)
df.test
```

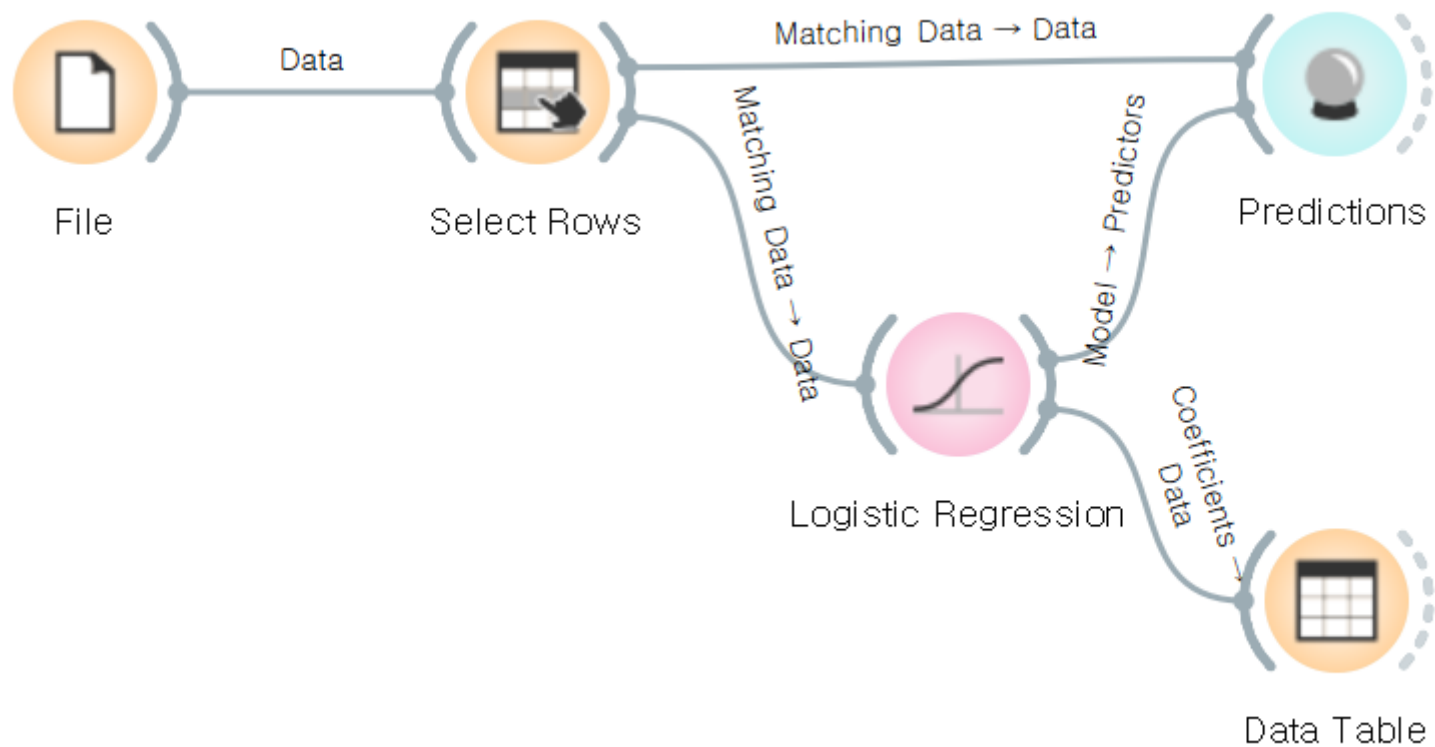
```
> df.test
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	pred
1	2.70	2.40	1.65	0.67	1
2	5.84	5.48	3.00	2.16	2
3	3.97	4.01	1.70	0.67	1



## 04. 로지스틱 회귀와 분류

### ■ Orange: Logistic Regression





## 04. 로지스틱 회귀와 분류

**Logistic Regression**

Name  
Logistic Regression

Regularization type: Lasso (L1) ▾

Strength:  
Weak  Strong  
C=16

☐ Balance class distribution

☒ Apply Automatically

? | 100 | - | 5 |

**Data Table**

Info  
5 instances (no missing data)  
1 feature  
No target variable,  
1 meta attribute

Variables  
☒ Show variable labels (if present)  
☐ Visualize numeric values  
☒ Color by instance classes

Selection  
☒ Select full rows

Restore Original Order

☒ Send Automatically

	name	Iris-versicolor
1	intercept	0
2	sepal length	-0.0466621
3	sepal width	-4.37791
4	petal length	4.99613
5	petal width	0

? | 5 | 5 | 5



## 04. 로지스틱 회귀와 분류

Predictions

Show probabilities for

- Iris-setosa
- Iris-versicolor
- Iris-virginica

Logistic Regression

	iris	sepal length	sepal width	petal length	petal width
1	Iris-setosa	5.1	3.5	1.4	0.2
2	Iris-setosa	4.9	3.0	1.4	0.2
3	Iris-setosa	4.7	3.2	1.3	0.2
4	Iris-setosa	4.6	3.1	1.5	0.2
5	Iris-setosa	5.0	3.6	1.4	0.2
6	Iris-setosa	5.4	3.9	1.7	0.4
7	Iris-setosa	4.6	3.4	1.4	0.3
8	Iris-setosa	5.0	3.4	1.5	0.2
9	Iris-setosa	4.4	2.9	1.4	0.2
10	Iris-setosa	4.9	3.1	1.5	0.1
11	Iris-setosa	5.4	4.1	1.5	0.2
12	Iris-setosa	5.5	4.2	1.6	0.2
13	Iris-setosa	4.9	3.6	1.4	0.1
14	Iris-setosa	4.7	3.6	1.1	0.1
15	Iris-setosa	5.2	3.5	1.2	0.2

1.00 : 0.00 : 0.00 → Iris-setosa

Model AUC CA F1 Precision Recall

Logistic Regression 1.000 1.000 1.000 1.000 1.000

Restore Original Order

Logistic Regression

Name

Logistic Regression

Regularization type: Lasso (L1)

Strength:

Weak Strong

C=16

☐ Balance class distribution

? | 100 | 100 | 1x100



## 04. 로지스틱 회귀와 분류





### ■ 이진 분류의 결과 표현:





- 혼동 행렬: *Confusion Matrix*
  - 이진 분류기의 분류 결과를  $2 \times 2$  행렬로 표시한 행렬
  - 이진 분류기가 분류(예측)할 때, 얼마나 많이 헛갈렸는가를 나타냄

		예측값	
		Positive	Negative
실제값	Positive	TP	FN
	Negative	FP	TN



## 04. 로지스틱 회귀와 분류

		PREDICTED VALUES	
		Positive (CAT)	Negative (DOG)
ACTUAL VALUES	Positive (CAT)	 <p>TRUE POSITIVE</p> <p>6</p> <p>YOU ARE A CAT</p>	 <p>FALSE NEGATIVE</p> <p>1</p> <p>TYPE II ERROR</p> <p>YOU ARE A DOG</p>
	Negative (DOG)	 <p>FALSE POSITIVE</p> <p>2</p> <p>TYPE I ERROR</p> <p>YOU ARE A CAT</p>	 <p>TRUE NEGATIVE</p> <p>11</p> <p>YOU ARE NOT A CAT</p>

		Actual Values	
		1	0
Predicted Values	1	 <p>TRUE POSITIVE</p> <p>You're pregnant</p>	 <p>FALSE POSITIVE</p> <p>You're pregnant</p> <p>TYPE 1 ERROR</p>
	0	 <p>FALSE NEGATIVE</p> <p>You're not pregnant</p> <p>TYPE 2 ERROR</p>	 <p>TRUE NEGATIVE</p> <p>You're not pregnant</p>





## 04. 로지스틱 회귀와 분류

### ■ 분류 모델의 성능 평가 지표: *Evaluation Metric*

- 정확도: *CA*, Classification *Accuracy*

- $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$

- 정밀도: *Precision*

- $Precision = \frac{TP}{TP+FP}$ , 분류기가 양성으로 판정한 것이 얼마나 정확한가?

- 재현율: *Recall*

- $Recall = \frac{TP}{TP+FN}$ , 분류기가 양성으로 판정한 것의 비율은 얼마인가?

- *F1*-Score

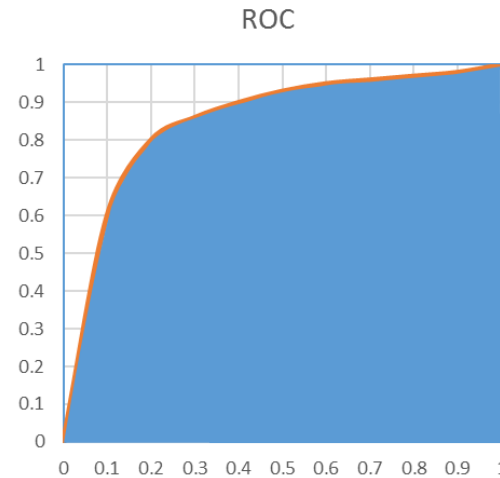
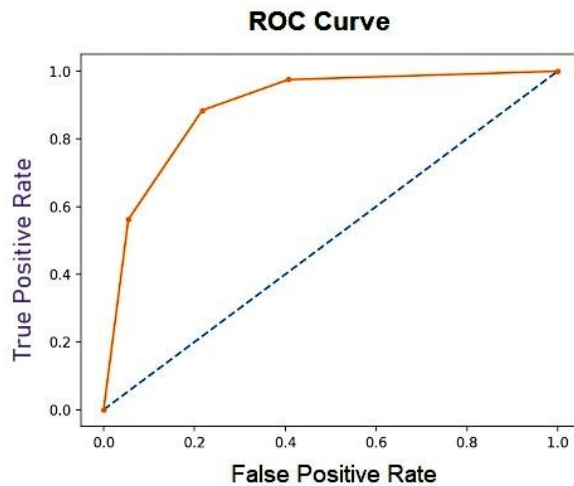
- $F1 = \frac{2 \times precision \times recall}{precision + recall}$ , 정밀도와 재현율의 조화평균



## 04. 로지스틱 회귀와 분류

### ■ 분류 모델의 성능 평가 지표:

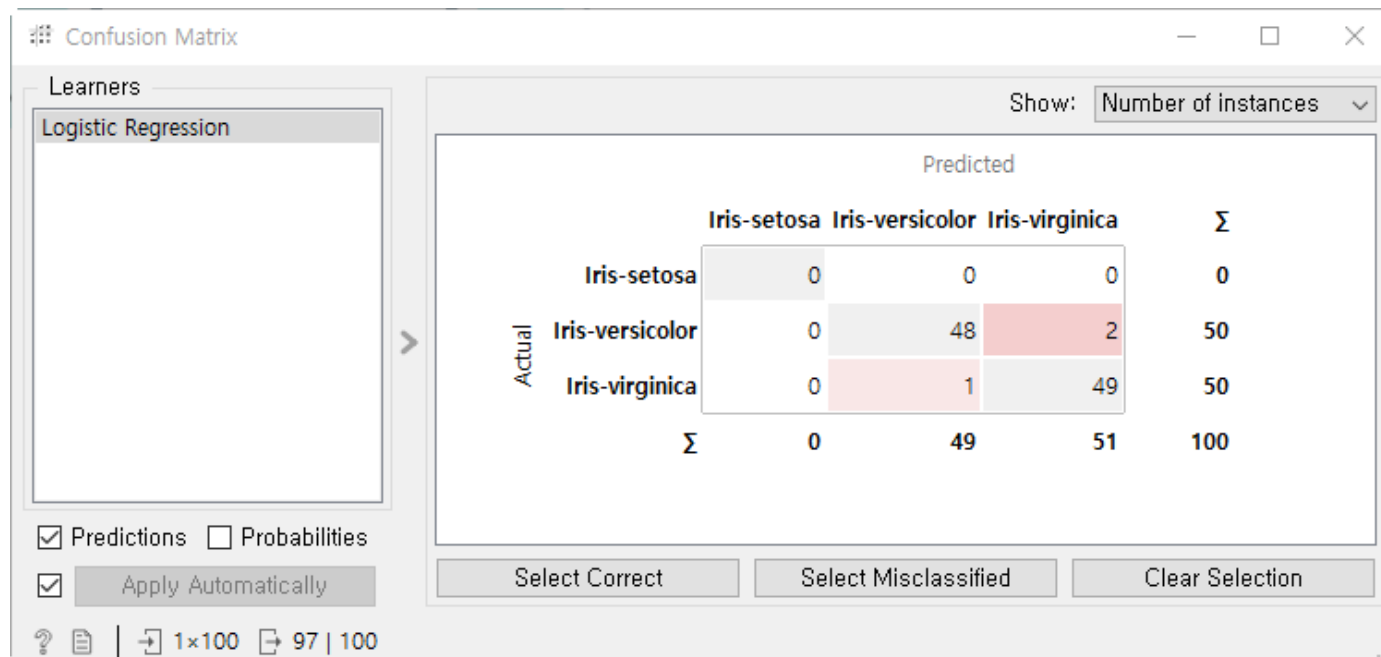
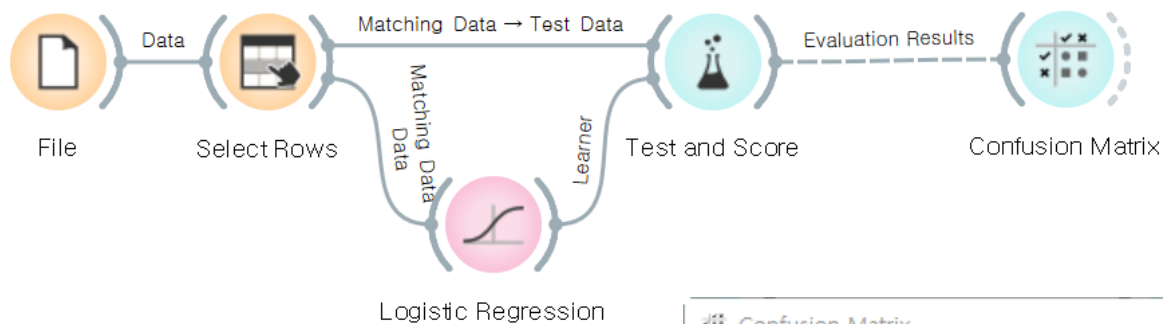
- ROC 곡선: Receiver Operation Characteristic Curve
  - 이진 분류의 결과에서 FP 비율과 TP 비율의 관계를 그린 곡선
- **AUC**: Area Under Curve
  - ROC 곡선의 하부 면적으로 표현하는 성능 평가 지표





## 04. 로지스틱 회귀와 분류

### ■ Orange: Confusion Matrix





## 04. 로지스틱 회귀와 분류

### ■ Orange: Test and Score

**Test and Score**

**Sampling**

- ☒ Cross validation
  - Number of folds: 10
  - ☐ Stratified
- ☐ Cross validation by feature
- ☐ Random sampling
  - Repeat train/test: 10
  - Training set size: 66 %
  - ☒ Stratified
- ☐ Leave one out
- ☐ Test on train data
- ☐ Test on test data

**Target Class**

(Average over classes)

**Model Comparison**

Area under ROC curve

☐ Negligible difference: 0,1

**Evaluation Results**

Model	AUC	CA	F1	Precision	Recall
Logistic Regression	0.993	0.970	0.970	0.970	0.970

**Model Comparison by AUC**

	Logistic Re...
Logistic Regression	

Table shows probabilities that the score for the model in the row is higher than that of the model in the column. Small numbers show the probability that the difference is negligible.

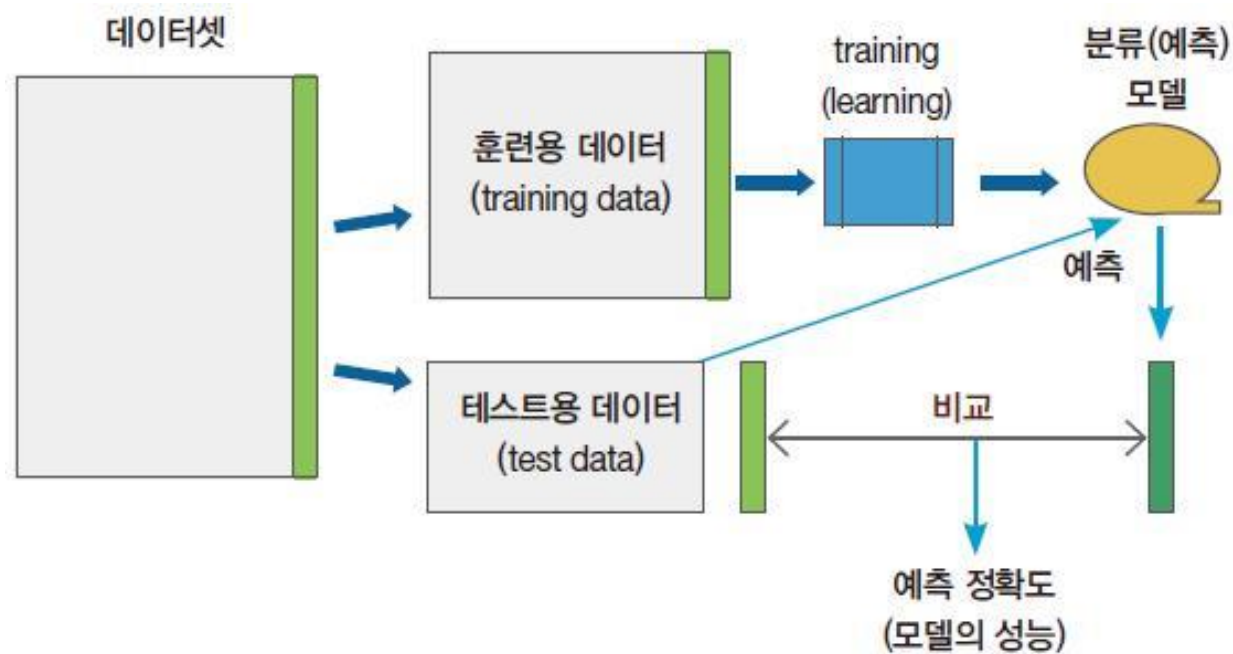
100 | - | 100 | 1x100



## 04. 로지스틱 회귀와 분류

### ■ 교차 검증: *Cross Validation*

- 훈련용 데이터셋을 가지고 분류기의 정확성을 검증하는 방법은?
- 훈련용 데이터와 시험용 데이터로 나누어서 성능 평가
  - 훈련용/시험용 데이터로 나누는 좋은 방법은?





## 04. 로지스틱 회귀와 분류

### ■ 교차 검증 방법들:

- k-폴드 교차 검증: *k-Fold Cross Validation*
  - 데이터셋을 k개의 폴드로 분리하여 하나만 시험용으로 사용
  - *Stratified*: 원본 데이터의 라벨 분포를 먼저 고려
- 랜덤 샘플링: *Random Sampling*
  - 임의로 학습용/시험용 데이터셋을 추출하여 교차 검증
- 리브-원-아웃: *Leave-One-Out* Cross Validation (LOOCV)
  - 폴드 하나에 샘플 하나만 남겨두는 k-폴드 교차 검증



## 04. 로지스틱 회귀와 분류

- k-폴드 교차 검증: k-Fold Cross Validation
  - 주어진 데이터셋을 k개로 나누어서 훈련용/검증용 데이터로 번갈아 사용
  - k개의 검증용 데이터셋으로 분류 모델의 성능을 테스트한 평균값을 지표로 사용





## 04. 로지스틱 회귀와 분류

- 피마 인디언 당뇨병 데이터셋: Pima Indians Diabetes Database
  - 캐글: <https://www.Kaggle.com/uciml/pima-Indians-diabetes-database>
  - 북아메리카 피마 지역 원주민의 당뇨병 데이터
    - 9개의 변수, 768개의 관측값
    - numeric feature 8개, categorical target 1개
  - 목적 변수:
    - Outcome (categorical): 0 또는 1





## 04. 로지스틱 회귀와 분류

### ■ 피마 인디언 당뇨병 데이터셋:

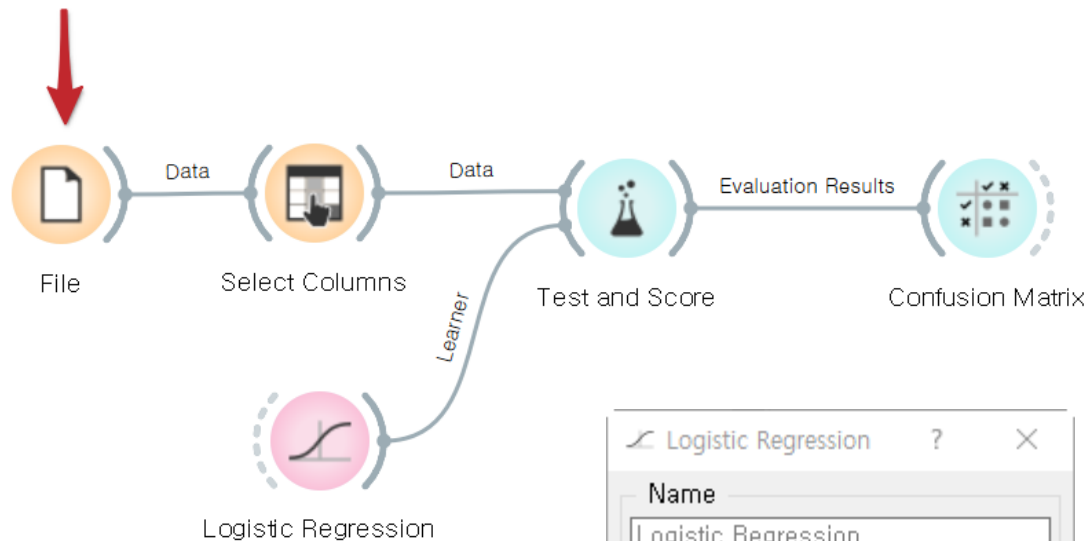
#### • 독립 변수:

- Pregnancies: 임신 횟수
- Glucose: 포도당 검사 수치
- BloodPressure: 혈압 수치 (mm Hg)
- SkinThickness: 삼두근 뒤쪽의 피하지방 측정값 (mm)
- Insulin: 혈청 인슐린 (mu U/ml)
- BMI: 체질량 지수( $weight\ in\ kg / (height\ in\ m)^2$ )
- DiabetesPedigreeFunction: 당뇨 내력 가중치 값
- Age: 나이



## 04. 로지스틱 회귀와 분류

diabetes.csv



**Logistic Regression** ? X

Name  
Logistic Regression

Regularization type: Lasso (L1) v

Strength:  
Weak  Strong  
C=1000

☐ Balance class distribution

☒ Apply Automatically

? | ? | ? | ? | ? | ?

**Select Columns** - □ X

Ignored  
Filter

Features  
Filter

- ☒ Age
- ☒ Pregnancies
- ☒ Glucose
- ☒ BloodPressure
- ☒ SkinThickness
- ☒ Insulin
- ☒ BMI
- ☒ DiabetesPedigreeFunction

Target  
☒ Outcome

Metas

Reset ☒ Ignore new variables by default ☒ Send Automatically

? | ? | ? | 768 | - | 768 | 8



## 04. 로지스틱 회귀와 분류

**Test and Score**

**Sampling**

- ☒ Cross validation
  - Number of folds: 10
  - ☐ Stratified
- ☐ Cross validation by feature
- ☐ Random sampling
  - Repeat train/test: 10
  - Training set size: 66 %
  - ☒ Stratified
- ☐ Leave one out
- ☐ Test on train data
- ☐ Test on test data

**Target Class**

(Average over classes)

**Model Comparison**

Area under ROC curve

☐ Negligible difference: 0,1

**Evaluation Results**

Model	AUC	CA	F1	Precision	Recall
Logistic Regression	0.827	0.771	0.765	0.765	0.771

**Model Comparison by AUC**

Model	Logistic Re...
Logistic Regression	

Table shows probabilities that the score for the model in the row is higher than that of the model in the column. Small numbers show the probability that the difference is negligible.

? | 768 | - | 768 | 1x768

**Confusion Matrix**

**Learners**

Logistic Regression

Show: Number of instances

		Predicted		$\Sigma$
		0	1	
Actual	0	437	63	500
	1	113	155	268
$\Sigma$		550	218	768

☒ Predictions ☐ Probabilities

☒ Apply Automatically

Select Correct Select Misclassified Clear Selection

? | 1x768 | - | 768



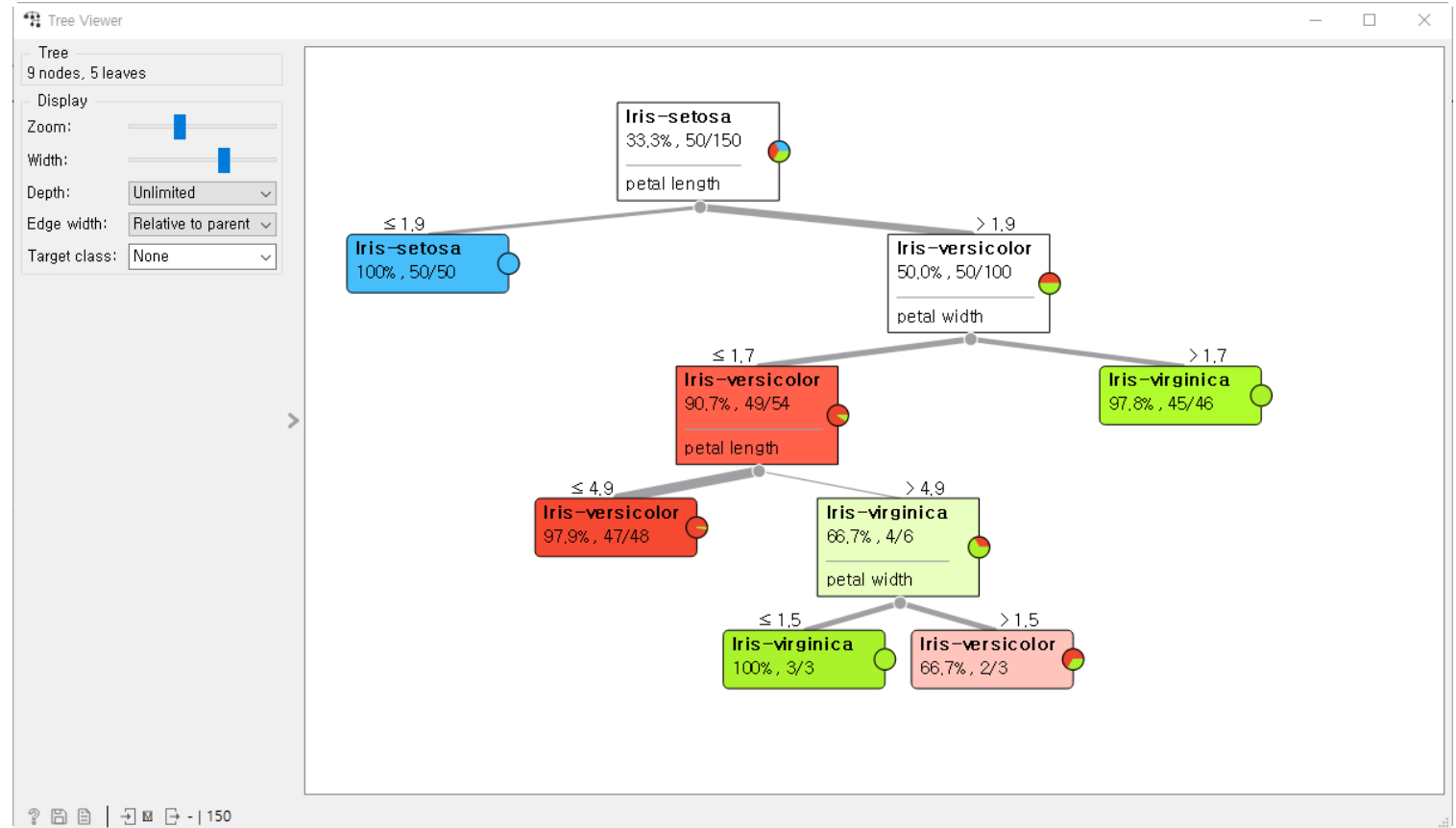
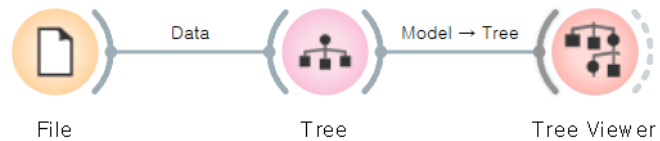
## 04. 로지스틱 회귀와 분류

- 결정 트리: *Decision Tree*
  - 데이터를 학습해서 트리 기반의 분류 규칙을 만드는 방법
  - 일종의 스무고개 방식: 분할 정복 (*Divide-and-Conquer*)
    - 루트 노드: 모든 데이터를 포함
    - 중간 노드: 특정 조건에 따른 데이터의 분할
    - 리프 노드: 분류 조건에 맞는 데이터의 집합



## 04. 로지스틱 회귀와 분류

### ■ Orange: Tree, Tree Viewer





## 04. 로지스틱 회귀와 분류

- 결정 트리를 학습하는 방법:
  - 분할 조건이 되는 특징을 어떻게 식별할 것인가?
    - 순도(*purity*): 단일 분류 데이터를 포함하는 정도
    - 분류의 기준: 각 단계별로 순도가 높아지는 방향으로 분류
  - 정보 이득: *information gain*
    - 데이터를 나누기 전과 데이터를 나눈 후의 정보량의 변화



## 04. 로지스틱 회귀와 분류

### ■ 정보 엔트로피: *Information Entropy*

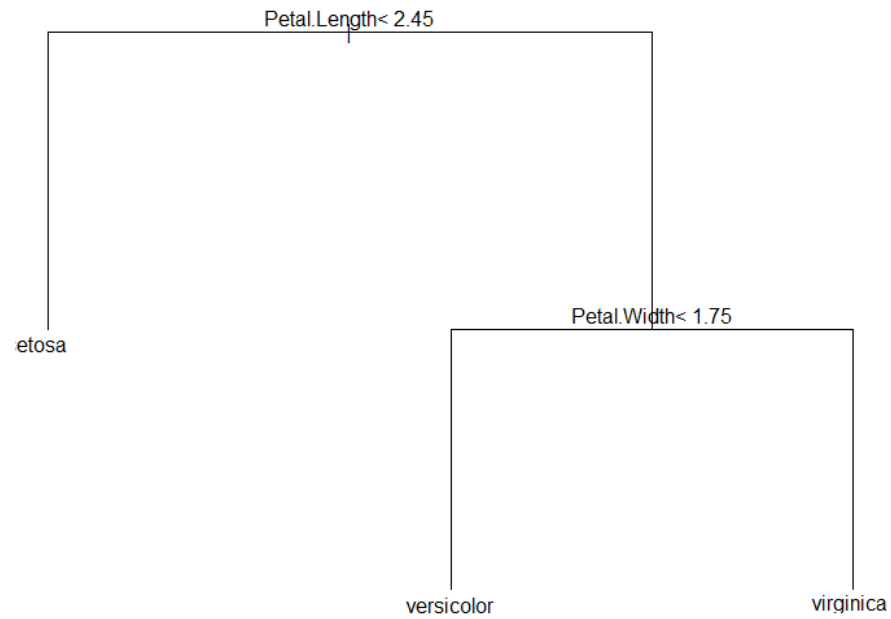
- 정보량: 어떤 데이터가 포함하고 있는 정보의 총 기대치
  - 어떤 사건이 발생할 확률이  $p$ 이면 그 사건의 정보량은  $\log_b \frac{1}{p}$
  - 어떤 사건의 정보량의 기대치:  $p \log_b \frac{1}{p} = -p \log_b p$
- 정보 엔트로피: 전체 사건의 기대치
  - 각 사건의 기대치의 합:  $H = -\sum_{i=1}^n p_i \log p_i$
  - 사건의 확률이  $1/2$ 이라면:  $\log$ 의 밑은 2:  $H = -\sum_{i=1}^n p_i \log_2 p_i$
- 데이터의 순도가 높아지는 방향: 정보 엔트로피가 낮아지는 방향



## 04. 로지스틱 회귀와 분류

### ■ R: rpart()

```
library(rpart)
df <- iris
model <- rpart(Species ~ ., data = df, method="class")
model
plot(model)
text(model)
```

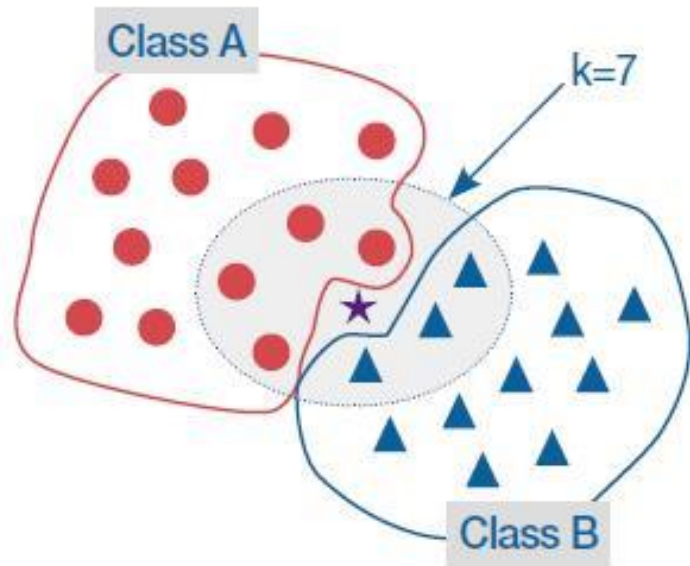






## 04. 로지스틱 회귀와 분류

- k-최근접-이웃: *k-Nearest-Neighbor*
  - 어떤 데이터와 가장 가까운  $k$ 개의 이웃을 보고, 가장 적절한 분류를 결정
  - 협업 필터링: *Collaboration Filtering*
    - 예) 유튜브/넷플릭스/페이스북의 추천 알고리즘





## 04. 로지스틱 회귀와 분류

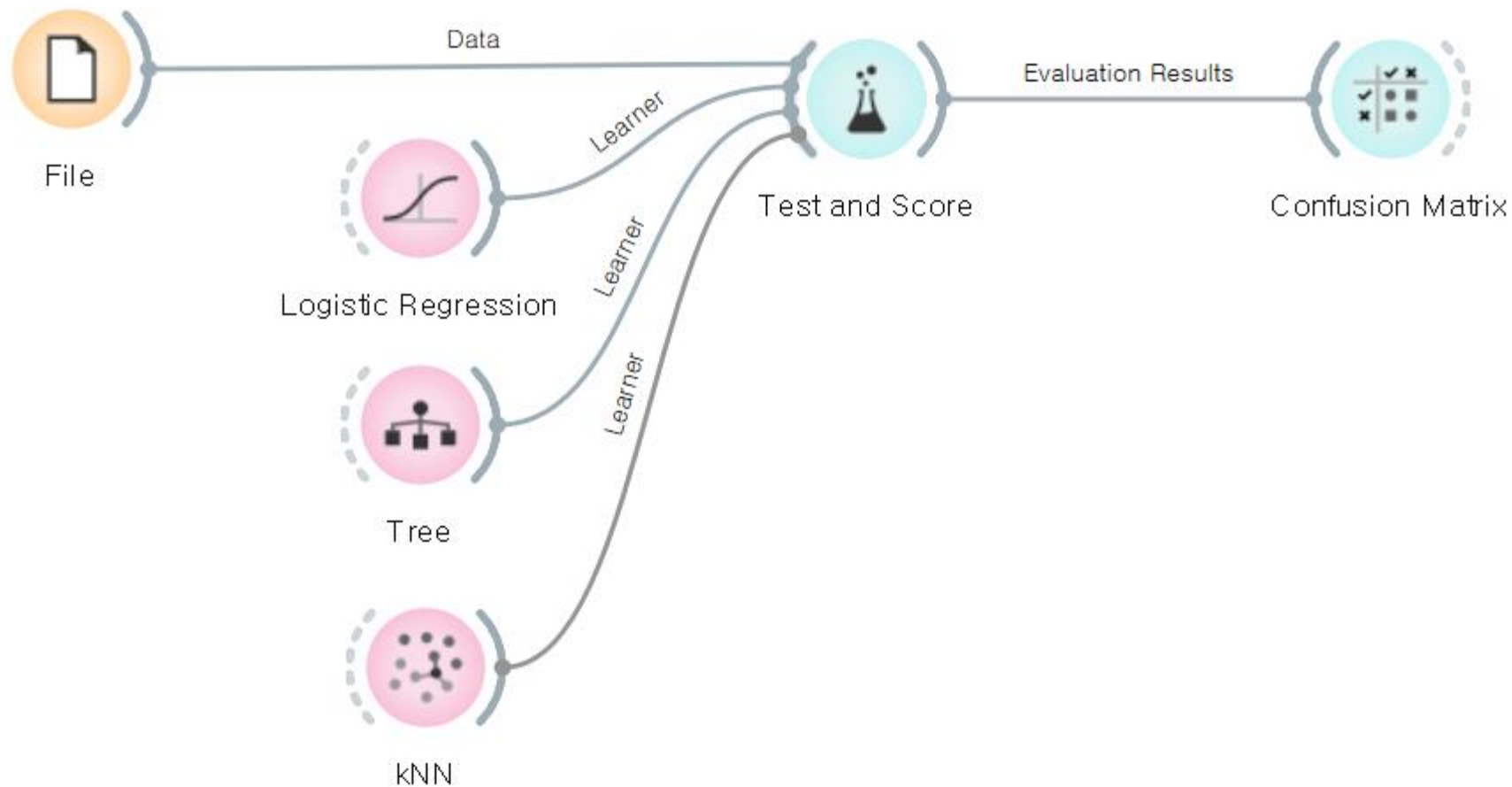
### ■ kNN의 적용:

- k의 값을 어떻게 결정할 수 있는가?
  - 데이터셋의 관측값의 개수가 100이라면? k는 10보다 작은 것이 좋음
  - 보통은 1~10의 값을 차례로 실험해 보면서 예측의 정확도를 평가
- 이웃의 군집이 동수가 나오면?
  - 일반적으로는 둘 중 하나를 임의로 선택 (random choice)
  - k의 값을 줄이거나 늘려서 선택하는 방법도 있음
- k개의 이웃을 찾는 것은 쉬운가?
  - 새로운 데이터 P가 도착했을 때 k개의 이웃을 찾으려면
  - 기존의 모든 데이터들과 거리를 비교해야 함



## 04. 로지스틱 회귀와 분류

### ■ 분류 모델의 성능 비교:





## 04. 로지스틱 회귀와 분류

**Test and Score**

**Sampling**

- ☒ Cross validation
  - Number of folds: 10
  - ☒ Stratified
- ☐ Cross validation by feature
- ☐ Random sampling
  - Repeat train/test: 10
  - Training set size: 25 %
  - ☒ Stratified
- ☐ Leave one out
- ☐ Test on train data
- ☐ Test on test data

**Target Class**

(Average over classes)

**Model Comparison**

Area under ROC curve

☐ Negligible difference: 0.1

**Evaluation Results**

Model	AUC	CA	F1	Precision	Recall
kNN	0.988	0.953	0.953	0.953	0.953
Tree	0.965	0.953	0.953	0.953	0.953
Logistic Regression	0.995	0.947	0.947	0.947	0.947

**Model Comparison by AUC**

	kNN	Tree	Logistic Regression
kNN		0.889	0.165
Tree	0.111		0.023
Logistic Regression	0.835	0.977	

Table shows probabilities that the score for the model in the row is higher than that of the model in the column. Small numbers show the probability that the difference is negligible.

150 | 3x150

*Any Questions?*

