Classification

Yun-Nan Chang

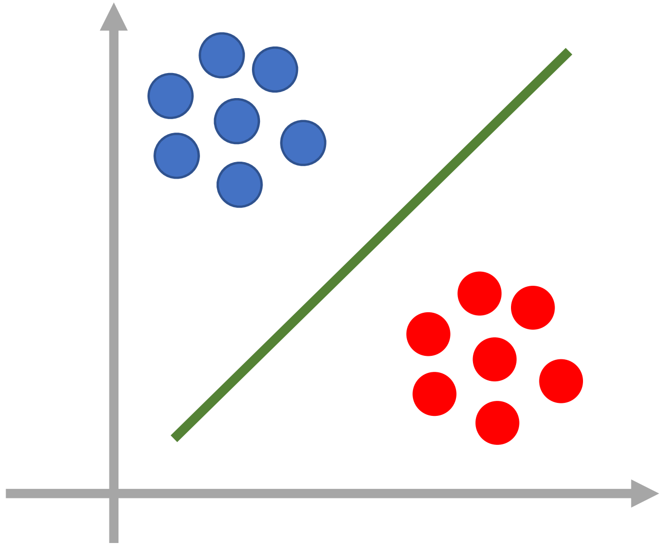
1 Classification

Classification Method

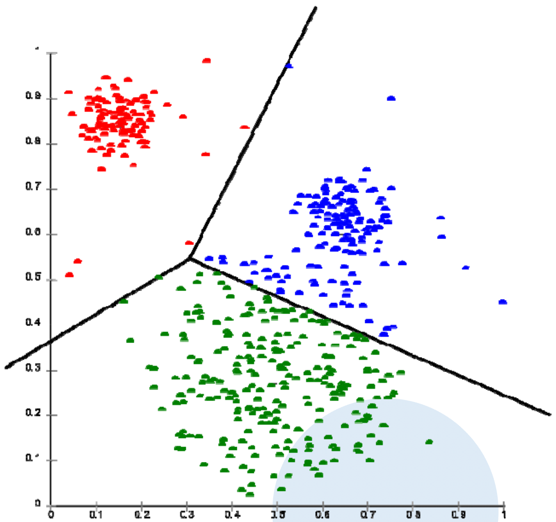
- Regression
- K-Means
- k-NN
- SVM

Classification Method

Linear Regression



K-Means



Decision theory

- In order to make decision based on a given x, we are interested in the probability of $P(C_k|x)$
- Using Bayes' theory $P(C_k|x) = \frac{P(x|C_k)P(C_k)}{P(x)}$
  - For two classes: $P(C_1|x) = \frac{P(x|C_1)P(C_1)}{P(x)} = \frac{P(x|C_1)P(C_1)}{P(x|C_1)P(C_1)+P(x|C_2)P(C_2)}$
  - If $P(C_1|x) > 0.5$=>class 1
- If we can know $p(C_1), p(C_2), p(x|C_1), p(x|C_2)$, we can derive $P(C_k|x)$ and make the decision.
  - It's called ***Probabilistic generative model***.

Minimizing the misclassification rate

- The probability of the misclassification will be
- p(mistake) = p$(x \in R_1, C_2) + $ p$(x \in R_2, C_1)$

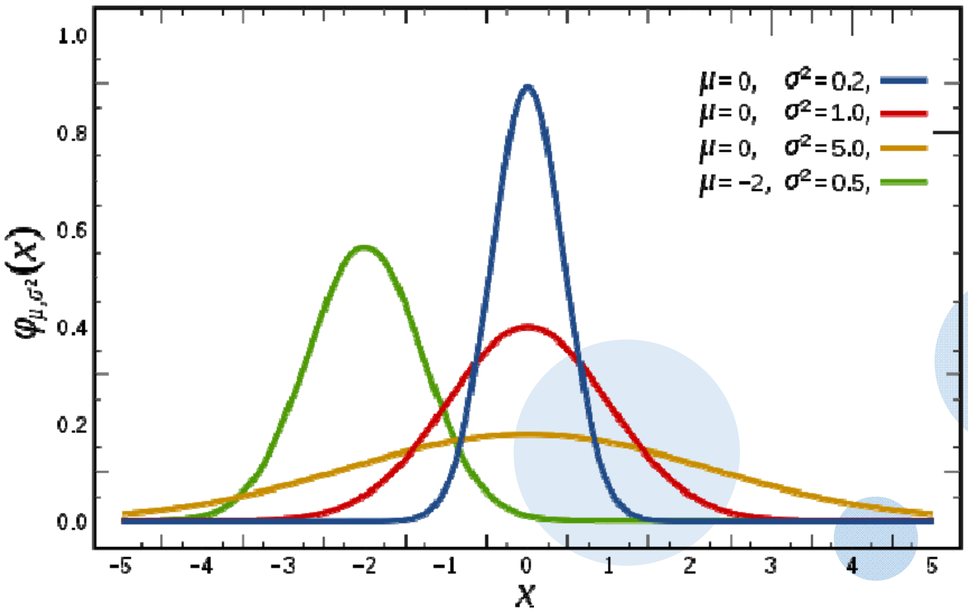
  $$= \int_{R_1} p(x, C_2) + \int_{R_2} p(x, C_1)$$

- The combined area of green an blue regions remain constant, we should try to minimize the red region.
- For multiclasses, p(correct) $= \sum_1^K \int_{R_k} p(x, C_k)$
- Expected loss $\mathrm{E}[L] = \sum_k \sum_j \int_{R_j} L_{kj} p(x, C_k)$

Gaussian Distribution

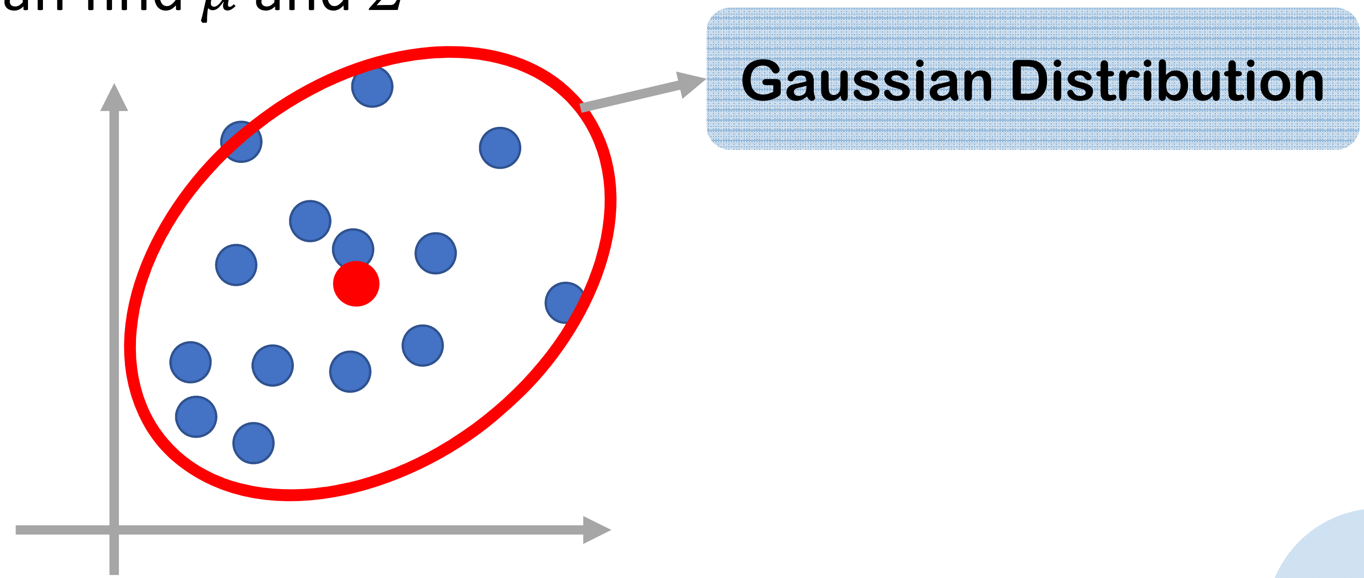
◈Gaussian Distribution

$$f_{\mu,\Sigma}(x) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}$$



Probability

- Assume the points are sampled from a Gaussian distribution.
- We can find μ and Σ

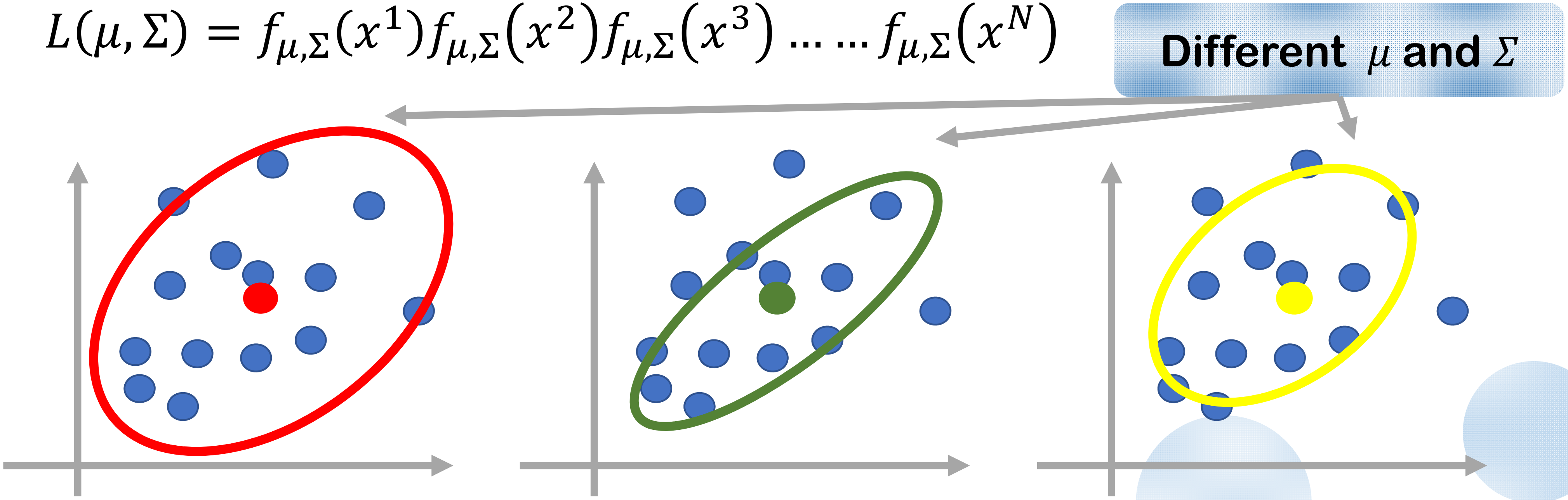


$$f_{\mu,\Sigma}(x) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}$$

Maximum Likelihood

- We can find the 'best' μ and Σ to get the Maximum $L(\mu, \Sigma)$

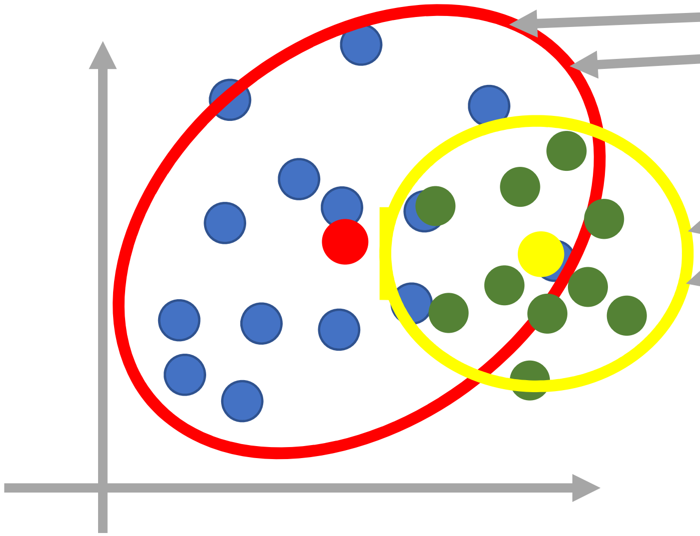
$$L(\mu, \Sigma) = f_{\mu,\Sigma}(x^1) f_{\mu,\Sigma}(x^2) f_{\mu,\Sigma}(x^3) \dots \dots f_{\mu,\Sigma}(x^N)$$



Classification

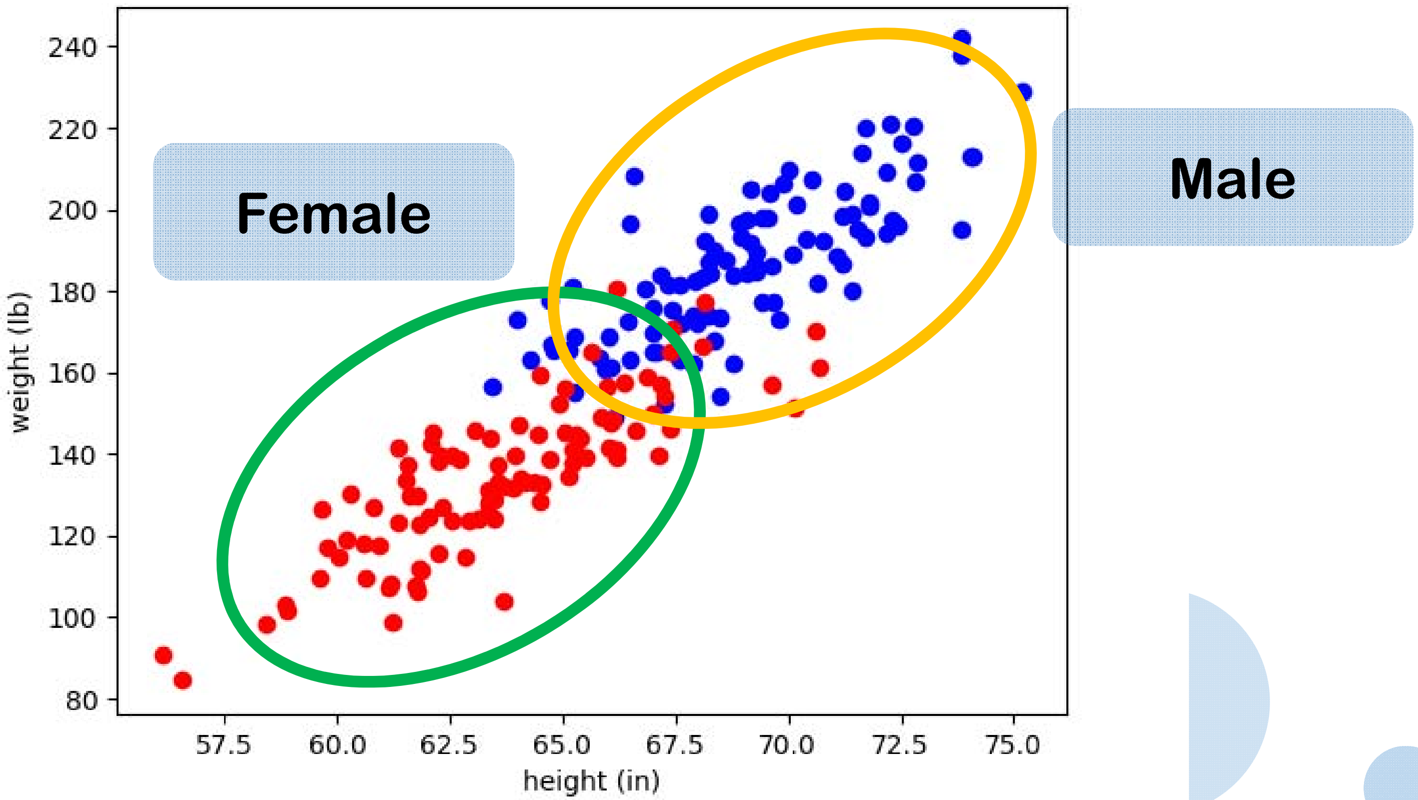
- We can do classification now.

$$P(C_1|x) = \frac{P(x|C_1)P(C_1)}{P(x|C_1)P(C_1) + P(x|C_2)P(C_2)}$$



$$f_{\mu,\Sigma}(x) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}$$

Example



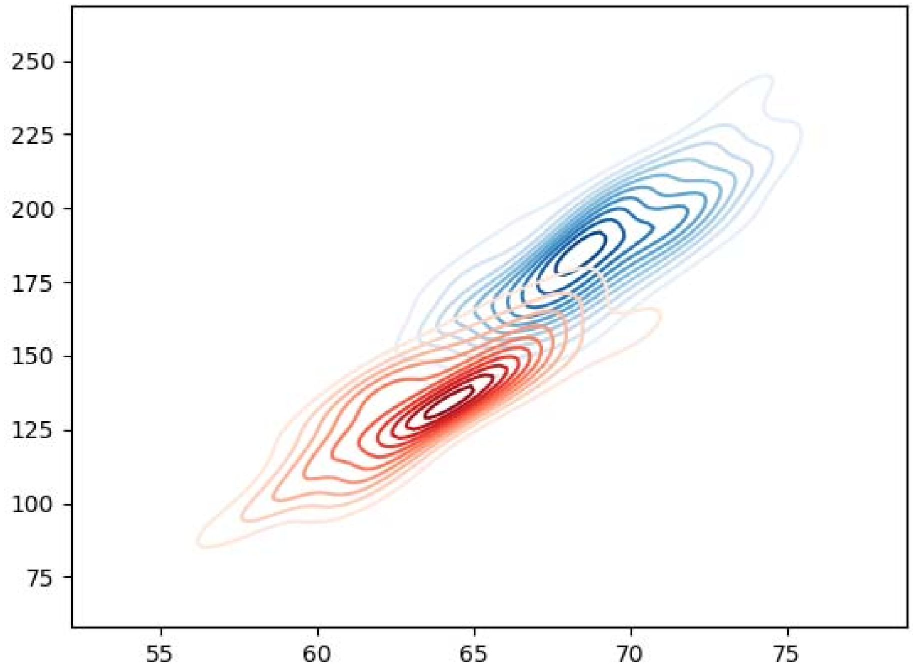
Female
Male
240
220
200
180
160
140
120
100
80
weight (lb)
57.5
60.0
62.5
65.0
67.5
70.0
72.5
75.0
height (in)

Gaussian Distribution

Female

$$\mu = \begin{bmatrix} 63 \\ 134 \end{bmatrix}$$

$$\sum = \begin{bmatrix} 8.24 & 46.37 \\ 46.37 & 373.05 \end{bmatrix}$$

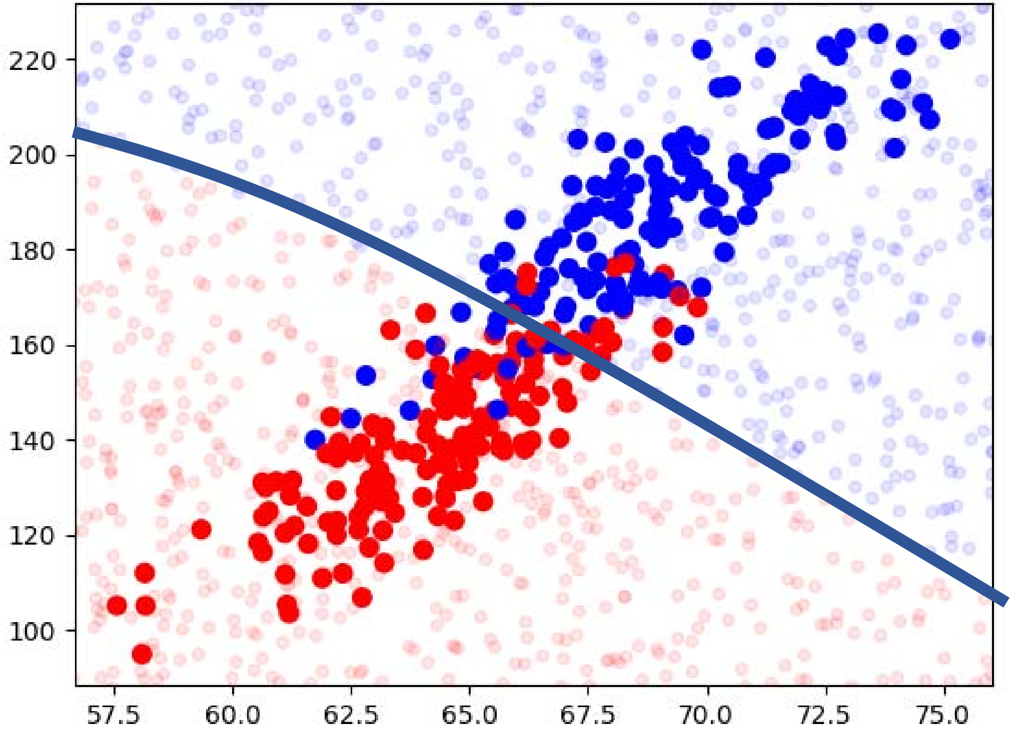


Male

$$\mu = \begin{bmatrix} 69 \\ 186 \end{bmatrix}$$

$$\sum = \begin{bmatrix} 6.76 & 39.57 \\ 39.57 & 372.56 \end{bmatrix}$$

Decision Bounce



Modifying Model

- Find μ^1, μ^2, Σ maximizing the likelihood $L(\mu^1,\mu^2,\Sigma)$

Male: $x^1, x^2, x^3, \ldots\ldots, x^{79}$

Female: $x^{80}, x^{81}, x^{82}, \ldots\ldots, x^{140}$

μ^1 $\quad$ Σ $\quad$ μ^2

Find μ^1, μ^2, Σ maximizing the likelihood $L(\mu^1,\mu^2,\Sigma)$

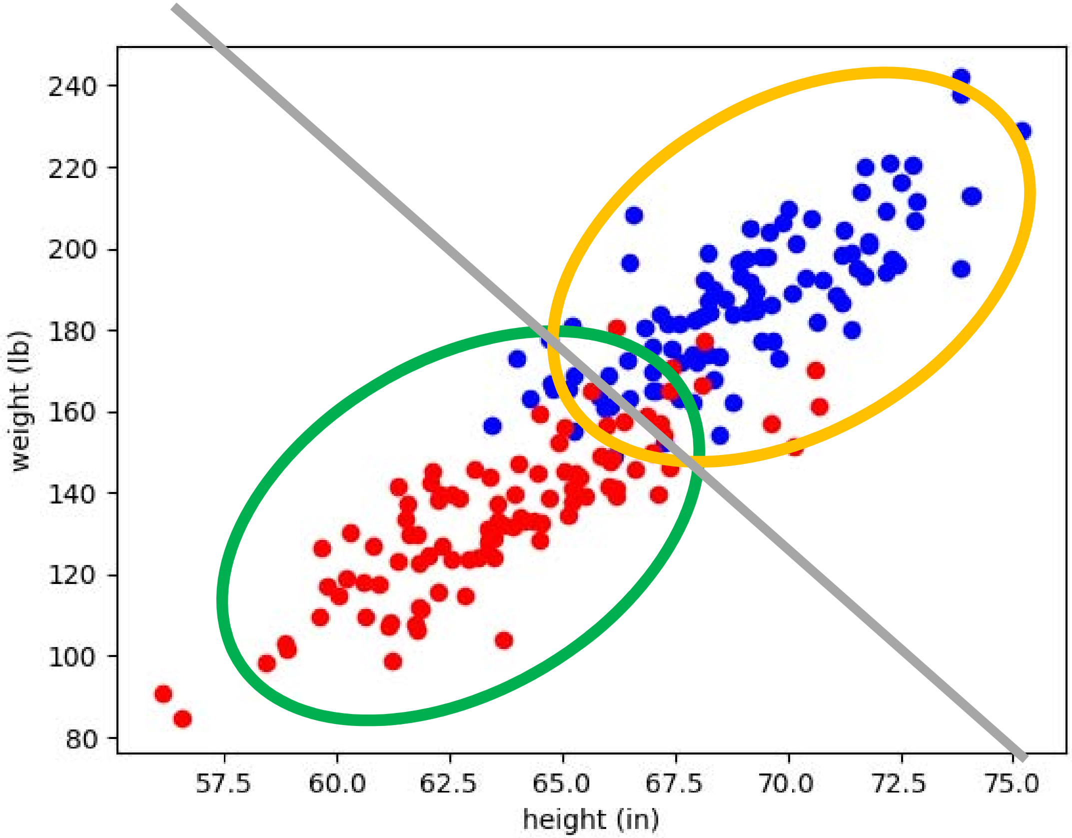
$$L(\mu^1,\mu^2,\Sigma) = f_{\mu^1,\Sigma}(x^1) f_{\mu^1,\Sigma}(x^2) \cdots f_{\mu^1,\Sigma}(x^{79})$$
$$\times f_{\mu^2,\Sigma}(x^{80}) f_{\mu^2,\Sigma}(x^{81}) \cdots f_{\mu^2,\Sigma}(x^{140})$$

μ^1 and μ^2 is the same

$$\Sigma = \frac{79}{140}\Sigma^1 + \frac{61}{140}\Sigma^2$$

From NTU Prof. H-Y. Lee's slide

Example



240
220
200
180
160
140
120
100
80
weight (lb)
57.5
60.0
62.5
65.0
67.5
70.0
72.5
75.0
height (in)

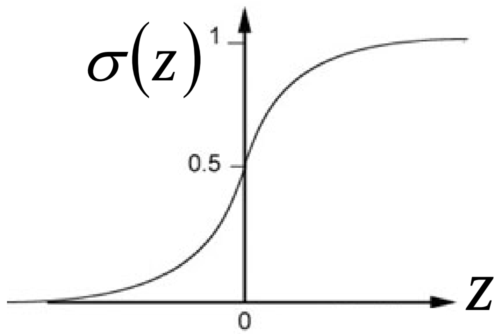
Posterior Probability

$$P(C_1|x) = \frac{P(x|C_1)P(C_1)}{P(x|C_1)P(C_1) + P(x|C_2)P(C_2)}$$

$$= \frac{1}{1 + \frac{P(x|C_2)P(C_2)}{P(x|C_1)P(C_1)}} = \frac{1}{1 + exp(-z)} = \sigma(z)$$

Sigmoid function

$$z = ln \frac{P(x|C_1)P(C_1)}{P(x|C_2)P(C_2)}$$



From NTU Prof. H-Y. Lee's slide

Posterior Probability

$$P(C_1|x) = \sigma(z)$$ sigmoid $$z = ln\frac{P(x|C_1)P(C_1)}{P(x|C_2)P(C_2)}$$

$$z = ln\frac{P(x|C_1)}{P(x|C_2)} + ln\boxed{\frac{P(C_1)}{P(C_2)}} \Rightarrow \frac{\frac{N_1}{N_1+N_2}}{\frac{N_2}{N_1+N_2}} = \frac{N_1}{N_2}$$

$$P(x|C_1) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\Sigma^1|^{1/2}}exp\left\{-\frac{1}{2}(x-\mu^1)^T(\Sigma^1)^{-1}(x-\mu^1)\right\}$$

$$P(x|C_2) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\Sigma^2|^{1/2}}exp\left\{-\frac{1}{2}(x-\mu^2)^T(\Sigma^2)^{-1}(x-\mu^2)\right\}$$

From NTU Prof. H-Y. Lee's slide

$$z = \boxed{ln\frac{P(x|C_1)}{P(x|C_2)}} + ln\boxed{\frac{P(C_1)}{P(C_2)}} = \frac{N_1}{N_2}$$

$$P(x|C_1) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\Sigma^1|^{1/2}}exp\left\{-\frac{1}{2}(x-\mu^1)^T(\Sigma^1)^{-1}(x-\mu^1)\right\}$$

$$P(x|C_2) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\Sigma^2|^{1/2}}exp\left\{-\frac{1}{2}(x-\mu^2)^T(\Sigma^2)^{-1}(x-\mu^2)\right\}$$

$$ln\frac{\cancel{\frac{1}{(2\pi)^{D/2}}}\boxed{\frac{1}{|\Sigma^1|^{1/2}}}\boxed{exp\left\{-\frac{1}{2}(x-\mu^1)^T(\Sigma^1)^{-1}(x-\mu^1)\right\}}}{\cancel{\frac{1}{(2\pi)^{D/2}}}\boxed{\frac{1}{|\Sigma^2|^{1/2}}}\boxed{exp\left\{-\frac{1}{2}(x-\mu^2)^T(\Sigma^2)^{-1}(x-\mu^2)\right\}}}$$

$$= \boxed{ln\frac{|\Sigma^2|^{1/2}}{|\Sigma^1|^{1/2}}}\boxed{exp\left\{-\frac{1}{2}\left[(x-\mu^1)^T(\Sigma^1)^{-1}(x-\mu^1) - (x-\mu^2)^T(\Sigma^2)^{-1}(x-\mu^2)\right]\right\}}$$

$$= ln\frac{|\Sigma^2|^{1/2}}{|\Sigma^1|^{1/2}} - \frac{1}{2}\left[(x-\mu^1)^T(\Sigma^1)^{-1}(x-\mu^1) - (x-\mu^2)^T(\Sigma^2)^{-1}(x-\mu^2)\right]$$

From NTU Prof. H-Y. Lee's slide

$$z = \boxed{ln\frac{P(x|C_1)}{P(x|C_2)}} + ln\boxed{\frac{P(C_1)}{P(C_2)}} = \frac{N_1}{N_2}$$

$$= ln\frac{|\Sigma^2|^{1/2}}{|\Sigma^1|^{1/2}} - \frac{1}{2}\left[\underline{(x-\mu^1)^T(\Sigma^1)^{-1}(x-\mu^1)} - \underline{(x-\mu^2)^T(\Sigma^2)^{-1}(x-\mu^2)}\right]$$

$$(x-\mu^1)^T(\Sigma^1)^{-1}(x-\mu^1)$$

$$= x^T(\Sigma^1)^{-1}x \underline{- x^T(\Sigma^1)^{-1}\mu^1 - (\mu^1)^T(\Sigma^1)^{-1}x} + (\mu^1)^T(\Sigma^1)^{-1}\mu^1$$

$$= x^T(\Sigma^1)^{-1}x \underline{- 2(\mu^1)^T(\Sigma^1)^{-1}x} + (\mu^1)^T(\Sigma^1)^{-1}\mu^1$$

$$(x-\mu^2)^T(\Sigma^2)^{-1}(x-\mu^2)$$

$$= x^T(\Sigma^2)^{-1}x - 2(\mu^2)^T(\Sigma^2)^{-1}x + (\mu^2)^T(\Sigma^2)^{-1}\mu^2$$

$$\boxed{\begin{aligned} z = ln\frac{|\Sigma^2|^{1/2}}{|\Sigma^1|^{1/2}} &- \frac{1}{2}x^T(\Sigma^1)^{-1}x + (\mu^1)^T(\Sigma^1)^{-1}x - \frac{1}{2}(\mu^1)^T(\Sigma^1)^{-1}\mu^1 \\ &+ \frac{1}{2}x^T(\Sigma^2)^{-1}x - (\mu^2)^T(\Sigma^2)^{-1}x + \frac{1}{2}(\mu^2)^T(\Sigma^2)^{-1}\mu^2 + ln\frac{N_1}{N_2} \end{aligned}}$$

From NTU Prof. H-Y. Lee's slide

$$P(C_1|x) = \sigma(z)$$

$$\boxed{\begin{aligned} z = {} & \cancel{ln\frac{|\Sigma^2|^{1/2}}{|\Sigma^1|^{1/2}}} \cancel{-\frac{1}{2}x^T(\Sigma^1)^{-1}x} + (\mu^1)^T(\Sigma^1)^{-1}x - \frac{1}{2}(\mu^1)^T(\Sigma^1)^{-1}\mu^1 \\ & \cancel{+\frac{1}{2}x^T(\Sigma^2)^{-1}x} - (\mu^2)^T(\Sigma^2)^{-1}x + \frac{1}{2}(\mu^2)^T(\Sigma^2)^{-1}\mu^2 + ln\frac{N_1}{N_2} \end{aligned}}$$

$$\Sigma_1 = \Sigma_2 = \Sigma$$

$$z = \underbrace{(\mu^1 - \mu^2)^T\Sigma^{-1}}_{\boldsymbol{w^T}} x \underbrace{- \frac{1}{2}(\mu^1)^T\Sigma^{-1}\mu^1 + \frac{1}{2}(\mu^2)^T\Sigma^{-1}\mu^2 + ln\frac{N_1}{N_2}}_{\mathrm{b}}$$

$$P(C_1|x) = \sigma(w \cdot x + b)$$

How about directly find ***w*** and b?

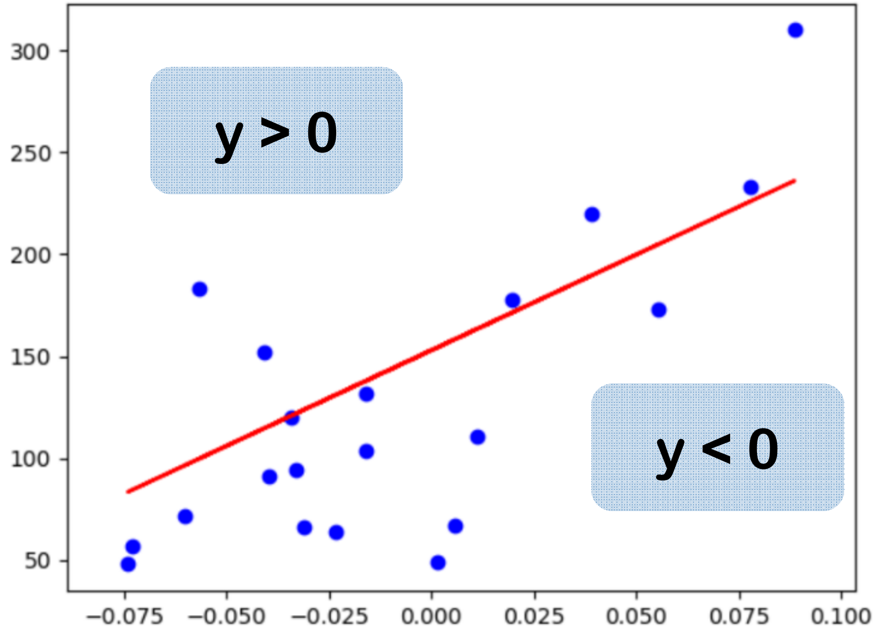
In generative model, we estimate $N_1, N_2, \mu^1, \mu^2, \Sigma$

Then we have ***w*** and b

From NTU Prof. H-Y. Lee's slide

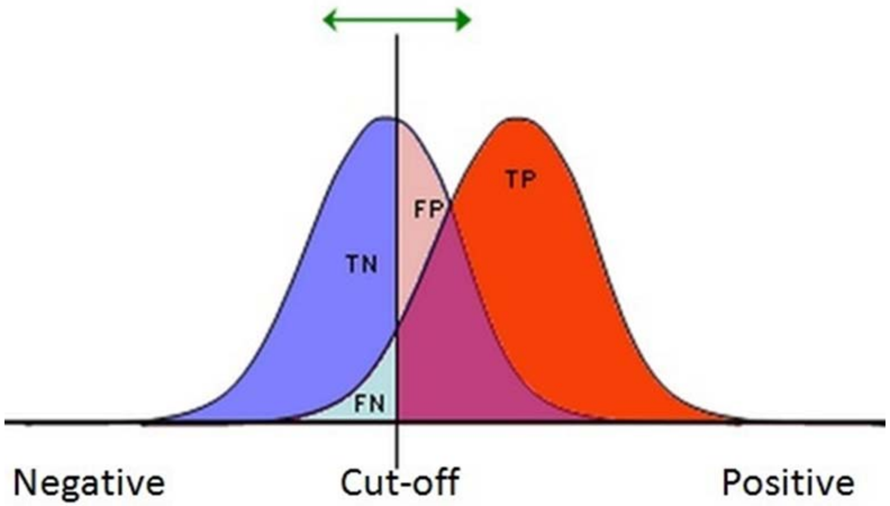
Binary classification

- Use linear regression for example, if x is assigned to class C_1 if $y(x) \geq 0$, and to class C_2 otherwise.



Cut-off point for binary classification

- The selection of cut-off will affect decision/prediction outcome
  - Actual positive: TP+FN   Actual negative: TN+FP.



	Actual Yes	**Actual No**
Predict Yes	TP	FP
Predict No	FN	TN

Confusion Matrix

	Actual Yes	Actual No
Predict Yes	TP (True Positive)	FP (False Positive)
Predict No	FN (False Negative)	TN (True Negative)

Accuracy	$\frac{TP + TN}{Total}$
Sensitivity (Recall)	$\frac{TP}{TP + FN}$
Precision	$\frac{TP}{TP + FP}$
Specificity	$\frac{TN}{TN + FP}$

Confusion Matrix

類的混淆矩陣

Actual

Predict

	Apple	Banana	Orange
Apple	10	2	1
Banana	1	15	4
Orange	4	2	6

Confusion Matrix of Apple

	Apple	No Apple
Apple	10(TP)	3(FP)
No Apple	5(FN)	27(TN)

F1-score

◈ Combine Recall and Precision.

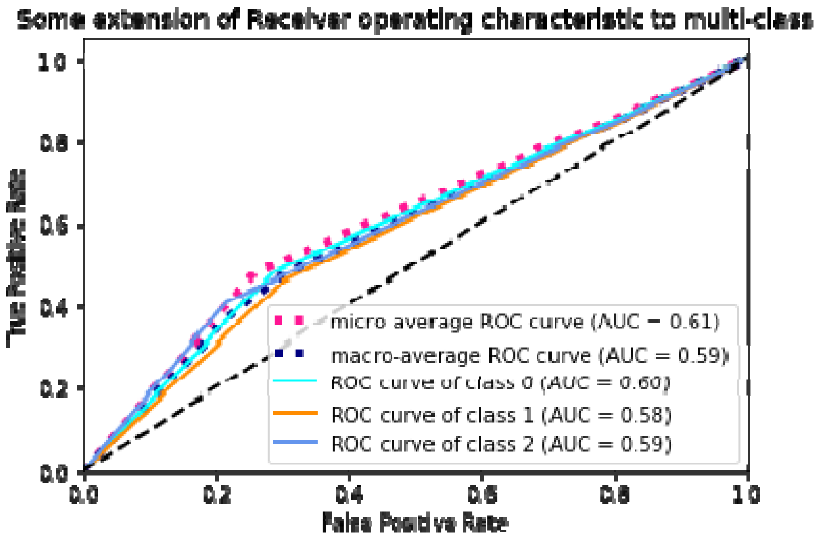
Sensitivity (Recall)	$\frac{TP}{TP+FN}$
Precision	$\frac{TP}{TP+FP}$

F1-Score

$$\frac{2}{\frac{1}{Recall}+\frac{1}{Precision}}$$

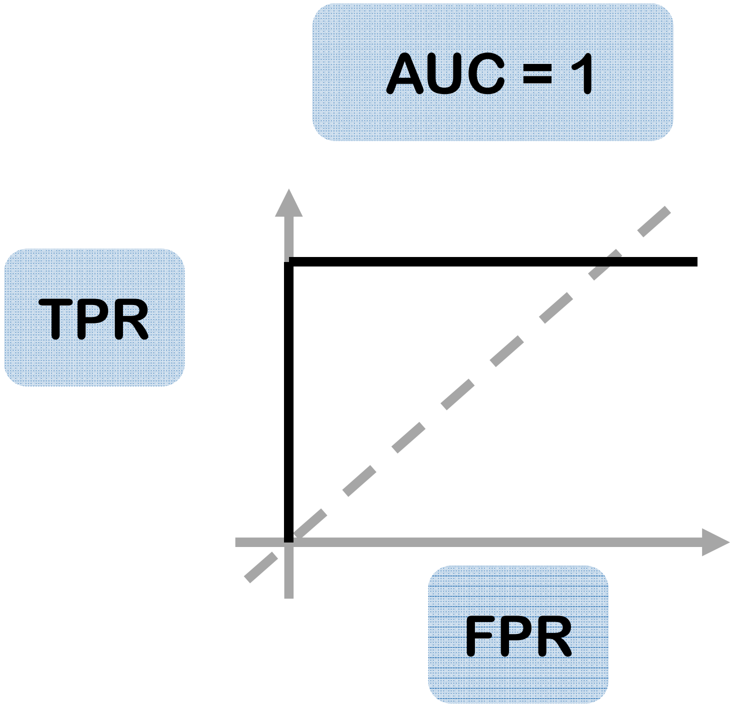
ROC

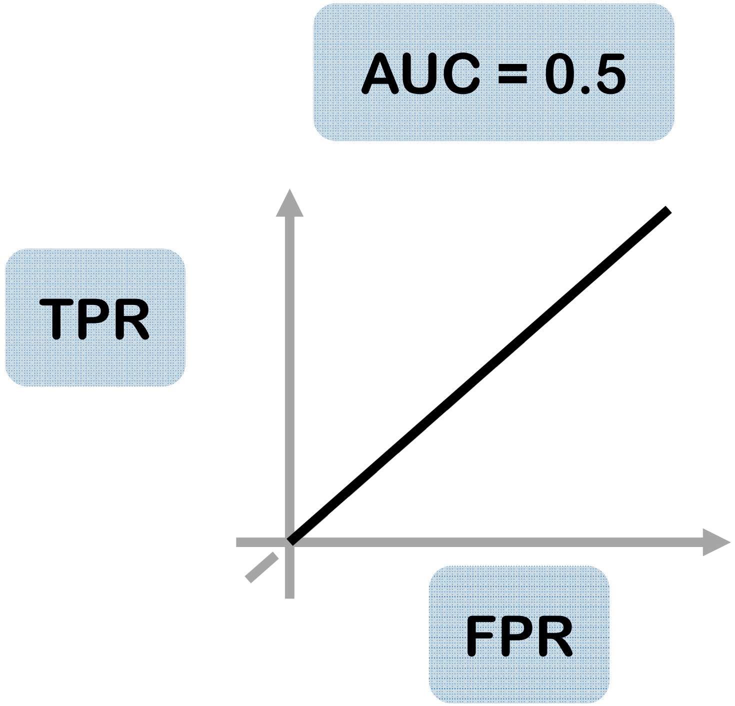
- Sensitivity (true positive rate/recall) vs 1-Specifity (true negative rate)
- The larger sensitivity is better.
- The smaller FPR is better.
- Therefore, the larger ***sensitivity-FPR*** is better.
  - The cut-off value which leads to the maximum is usually used as the final decision point.
- Sensitivity-FPR =0 can be regarded as the reference line
  - Different methods could lead to different curves.
  - Larger AUC (Area under the Curve of ROC) is better.



AUC

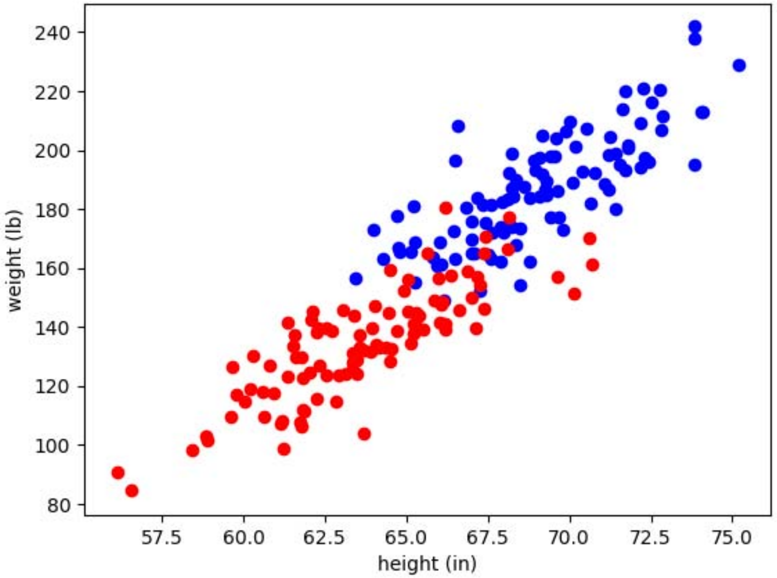
- ◇TPR, true positive rate
- ◇FPR, false positive rate





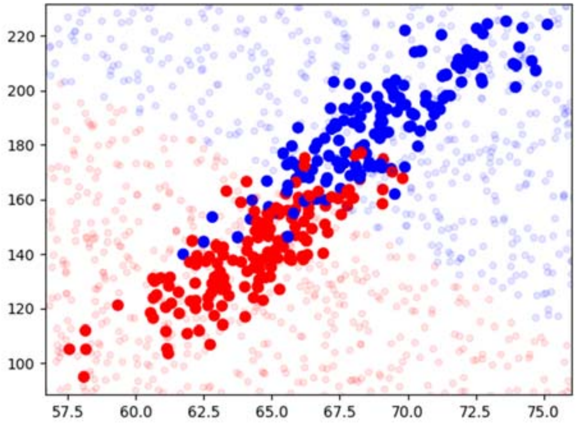
Binary classification example

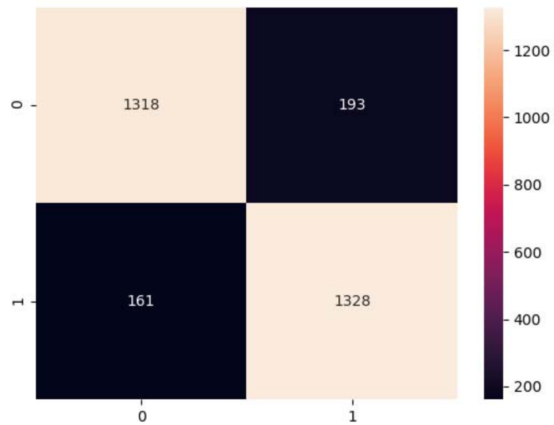
- Dataset : People’s height and weight
- Purpose : Predict Male or Female



Logistic Regression scikit learn

寺間的話





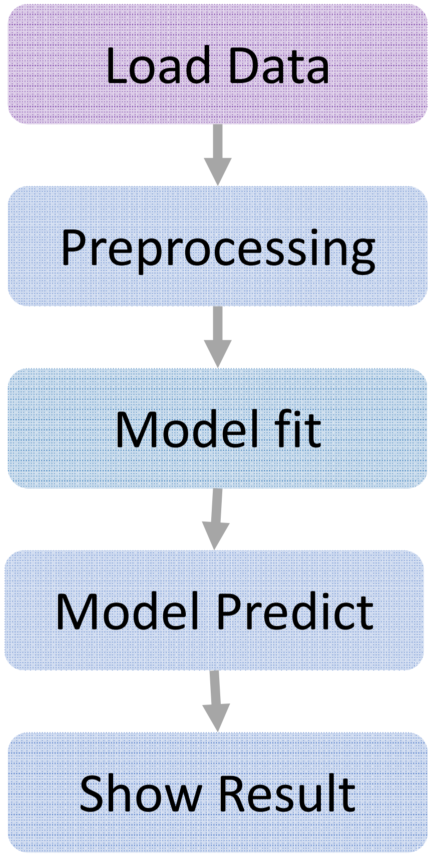

```
### Train
# read data
df_gender=pd.read_csv('./data/weight-height.csv')
df_gender=df_gender.replace('Male','0')
df_gender=df_gender.replace('Female','1')
df_gender.head()

y=df_gender['Gender']
df_gender.drop( ['Gender'],axis = 1,inplace = True)
X=df_gender

# split data
X_train, X_test, y_train, y_test=train_test_split(X,y,test_size=0.3, random_state=0)
# train
model = GaussianNB()
model.fit(X_train, y_train)

# predict
y_pred = model.predict(X_test)
# confusion matrix
print(confusion_matrix(y_test, y_pred))
ax = sns.heatmap(confusion_matrix(y_test, y_pred), annot=True, fmt="d")
plt.show()
```

How to use?



- use csv file
- Import pandas as pd

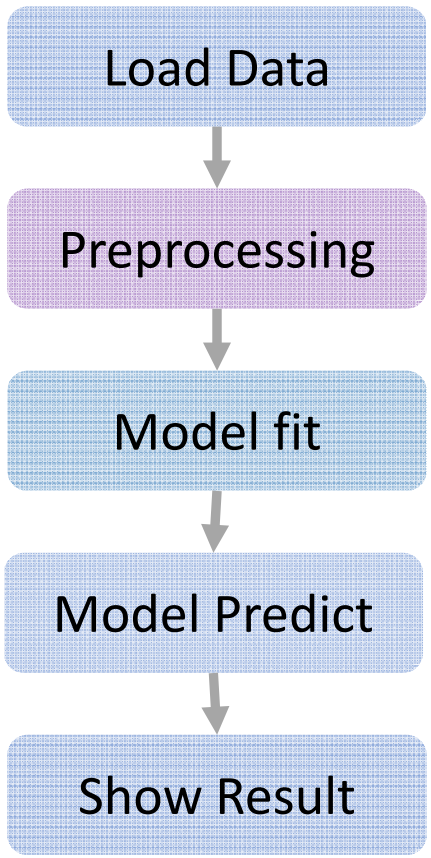
df_gender=pd.read_csv('./data/weight-height.csv')

df_gender=df_gender.replace('Male','0')

df_gender=df_gender.replace('Female','1')

df_gender.head()

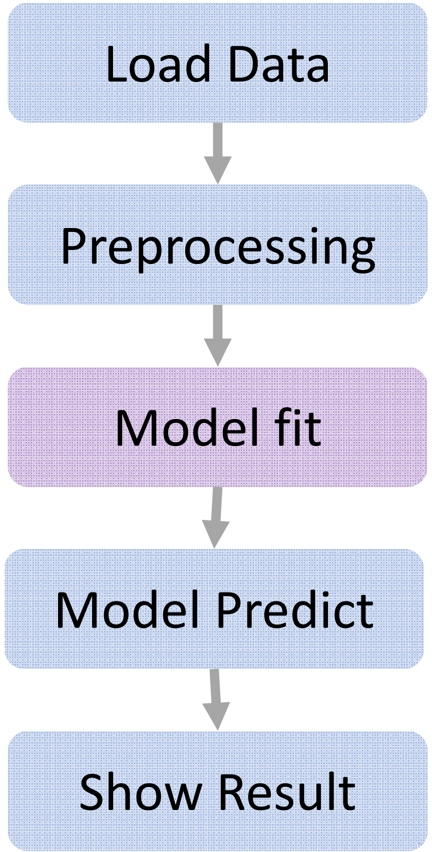
How to use?



◈Split dataset

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test
= train_test_split(X,y,test_size=0.3, random_state=0)
```

How to use?



◈Use Gaussian Naive Bayes model from sklearn

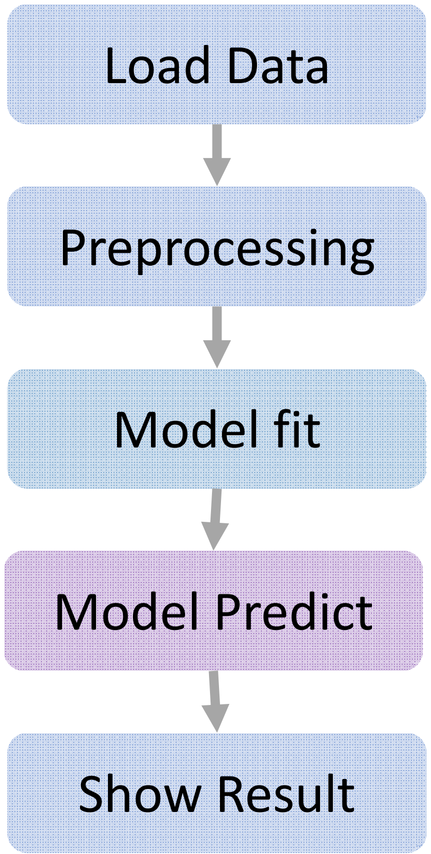
from sklearn.naive_bayes import GaussianNB

model = GaussianNB()

◈Use this model to train

model.fit(X_train, y_train)

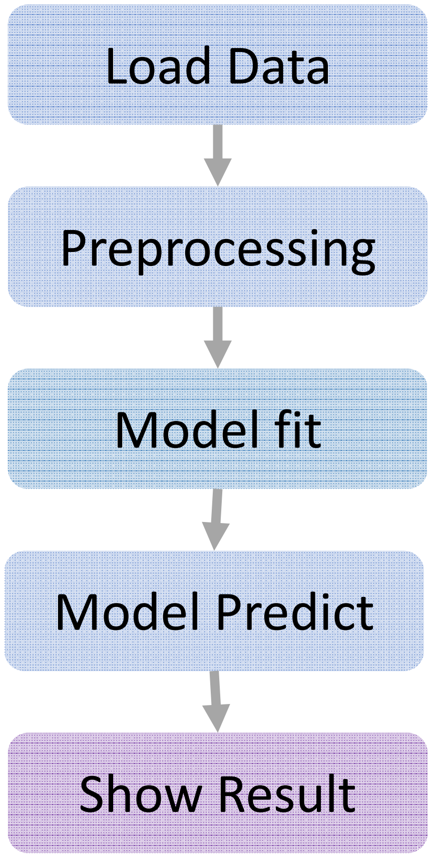
How to use?



◈Get predict

y_pred = model.predict(X_test)

How to use?



◈Confusion Matrix

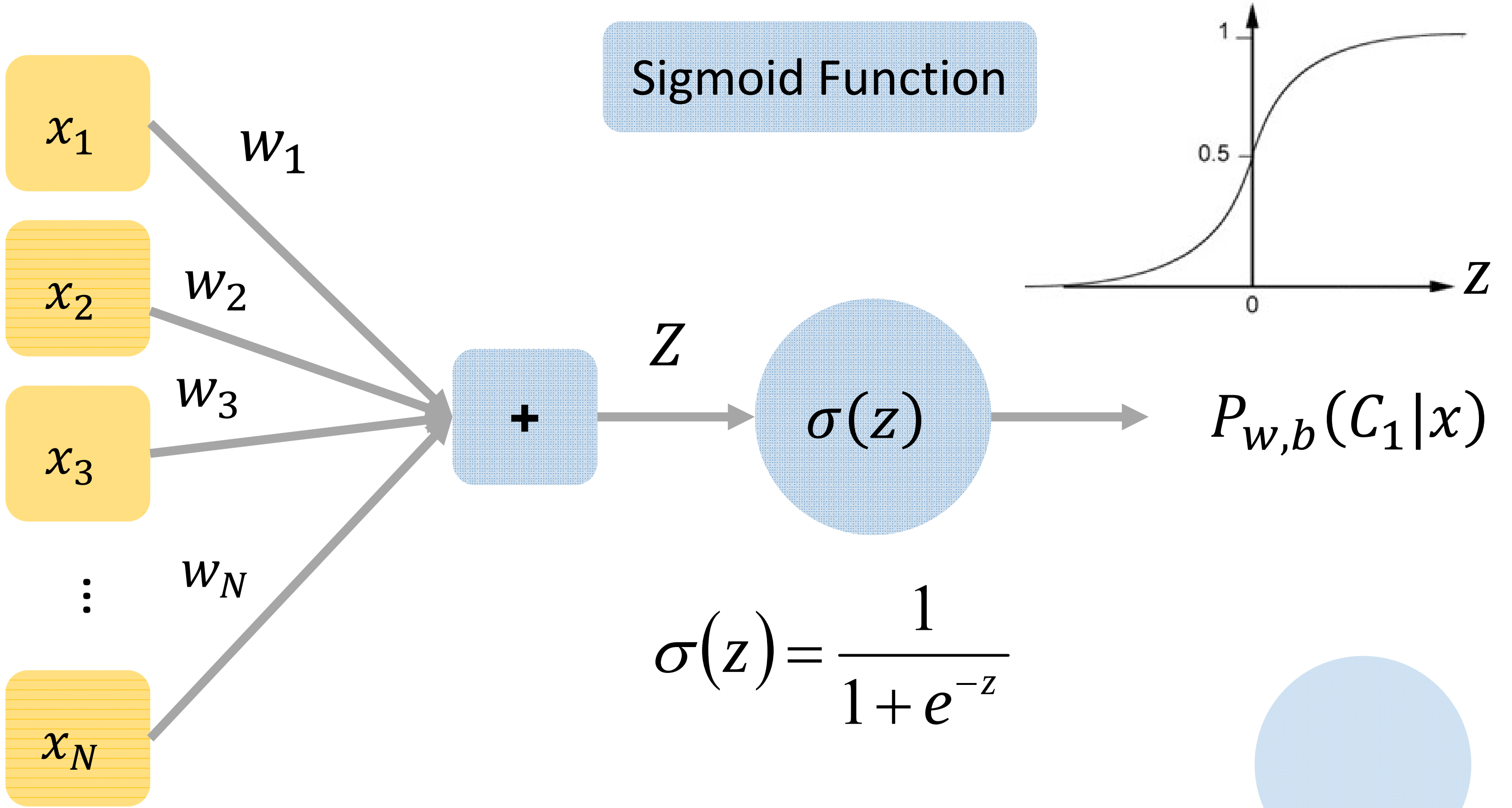
```
from sklearn.metrics import confusion_matrix
CM = confusion_matrix(y_test, y_pred)
```

◈Use matplotlib and seaborn

```
import matplotlib.pyplot as plt
import seaborn as sns
ax = sns.heatmap(CM, annot=True, fmt="d")
plt.show()
```

2 Logistic Regression

Logistic Regression



$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Setting of the object function

Training Data	x^1	x^2	x^3	…	x^N
	C_1	C_1	C_2		C_1

- Assume the data is generated based on $f_{w,b}(x) = P_{w,b}(C_1|x)$
- Given a set of w and b, what is its probability of generating the data?
- $L(w,b) = \underbrace{\boxed{f_{w,b}(x^1)}}_{C_1} f_{w,b}(x^2) \underbrace{\boxed{\left(1 - f_{w,b}(x^3)\right)}}_{C_2} \cdots f_{w,b}(x^N)$
- The most likely w^* and b^* is the one with the largest $L(w,b)$.

$$x^1 \quad x^2 \quad x^3 \quad \cdots\cdots$$
$$C_1 \quad C_1 \quad C_2$$

$$x^1 \quad x^2 \quad x^3 \quad \cdots\cdots$$
$$\hat{y}^1 = 1 \quad \hat{y}^2 = 1 \quad \hat{y}^3 = 0$$

$\hat{y}^n$: 1 for class 1, 0 for class 2

$$L(w,b) = f_{w,b}(x^1) f_{w,b}(x^2)\left(1 - f_{w,b}(x^3)\right)\cdots$$

$$\boxed{w^*, b^* = arg \max_{w,b} L(w,b)} = \boxed{w^*, b^* = arg \min_{w,b} -lnL(w,b)}$$

$$-lnL(w,b)$$
$$= -lnf_{w,b}(x^1) \Rightarrow -\left[1\ lnf(x^1) \cancel{+\ 0\ ln\left(1 - f(x^1)\right)}\right]$$
$$-lnf_{w,b}(x^2) \Rightarrow -\left[1\ lnf(x^2) \cancel{+\ 0\ ln\left(1 - f(x^2)\right)}\right]$$
$$-ln\left(1 - f_{w,b}(x^3)\right) \Rightarrow -\left[\cancel{0\ lnf(x^3) +}\ 1\ ln\left(1 - f(x^3)\right)\right]$$
$$\vdots$$

From NTU Prof. H-Y. Lee's slide

Setting of the object function

$$L(w,b) = f_{w,b}(x^1) f_{w,b}(x^2)\left(1 - f_{w,b}(x^3)\right) \cdots f_{w,b}(x^N)$$

$$-lnL(w,b) = lnf_{w,b}(x^1) + lnf_{w,b}(x^2) + ln\left(1 - f_{w,b}(x^3)\right) \cdots$$

$\hat{y}^n$: 1 for class 1, 0 for class 2

$$= \sum_n -\underbrace{\left[\hat{y}^n lnf_{w,b}(x^n) + (1-\hat{y}^n) ln\left(1 - f_{w,b}(x^n)\right)\right]}_{\text{Cross entropy between two Bernoulli distribution}}$$

Distribution p:

$p(x = 1) = \hat{y}^n$

$p(x = 0) = 1 - \hat{y}^n$

$\xleftrightarrow{\text{cross entropy}}$

Distribution q:

$q(x = 1) = f(x^n)$

$q(x = 0) = 1 - f(x^n)$

$$H(p,q) = -\sum_x p(x) ln(q(x))$$

From NTU Prof. H-Y. Lee's slide

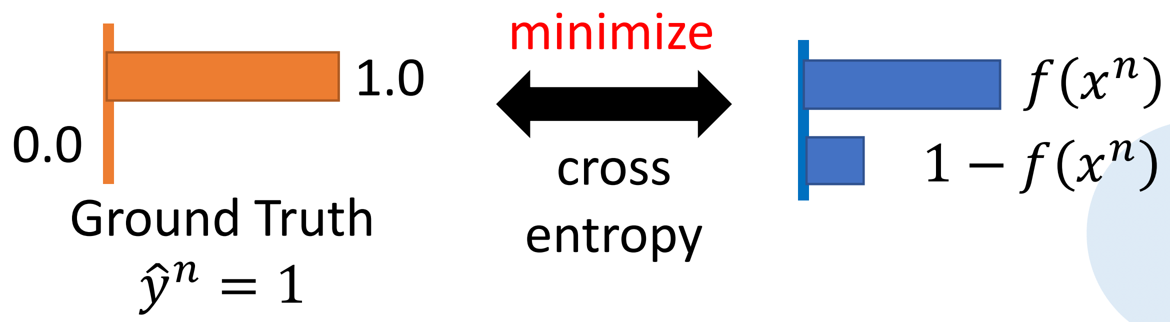
Setting of the object function

$$L(w,b) = f_{w,b}(x^1) f_{w,b}(x^2) \left(1 - f_{w,b}(x^3)\right) \cdots f_{w,b}(x^N)$$

$$-lnL(w,b) = lnf_{w,b}(x^1) + lnf_{w,b}(x^2) + ln\left(1 - f_{w,b}(x^3)\right) \cdots$$

$\hat{y}^n$: 1 for class 1, 0 for class 2

$$= \sum_n -\underbrace{\left[\hat{y}^n lnf_{w,b}(x^n) + (1-\hat{y}^n) ln\left(1 - f_{w,b}(x^n)\right)\right]}_{\text{Cross entropy between two Bernoulli distribution}}$$



From NTU Prof. H-Y. Lee's slide