Statistical Learning

https://github.com/ggorr/Machine-Learning/tree/master/ISLR

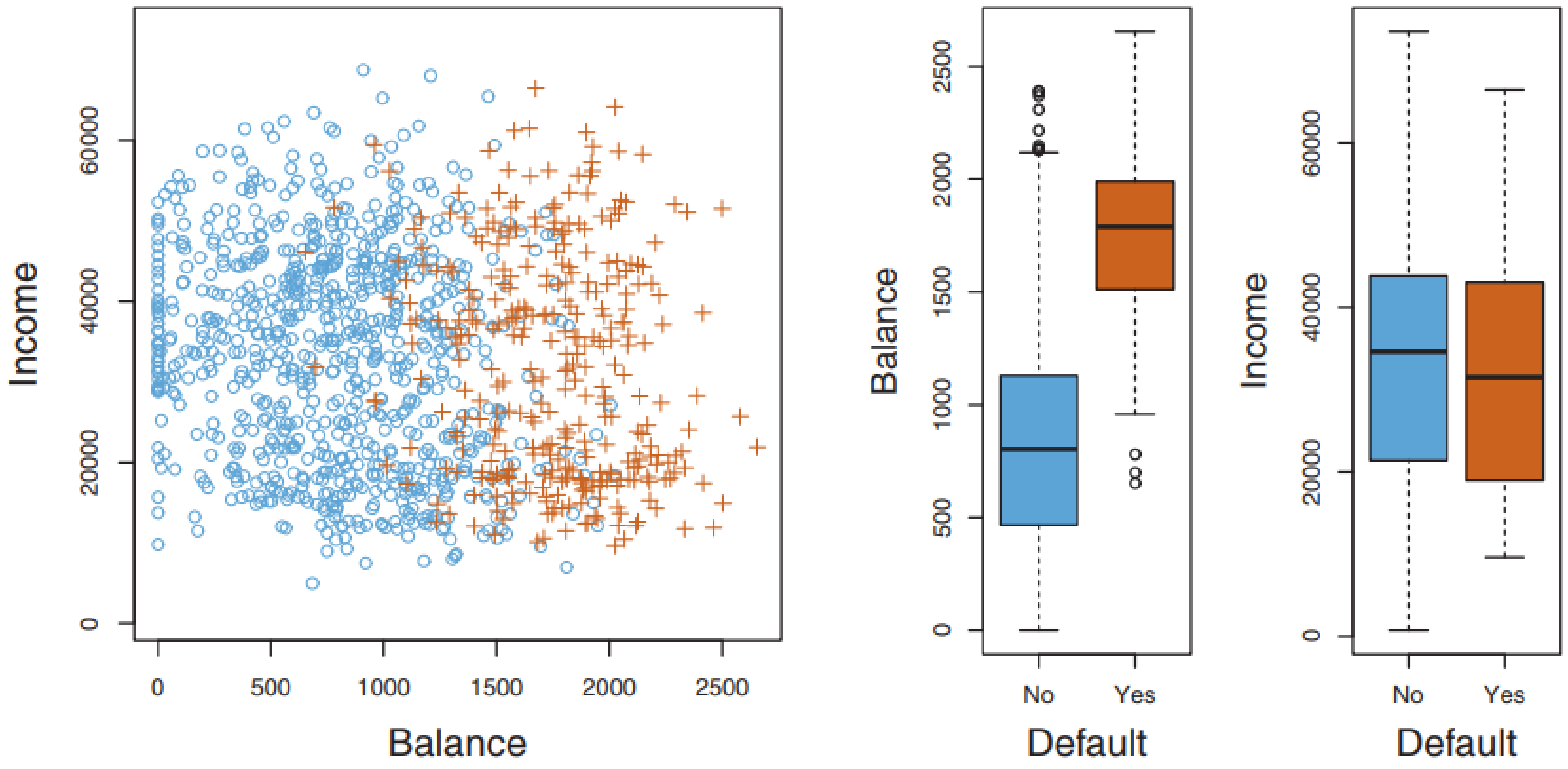
4 Classification

- 4.1 An Overview of Classification
- 4.2 Why Not Linear Regression?
- 4.3 Logistic Regression
- 4.4 Linear Discriminant
- 4.5 A Comparison of Classification Methods
- 4.6 Lab: Logistic Regression, LDA, QDA, and KNN
- 4.7 Exercises

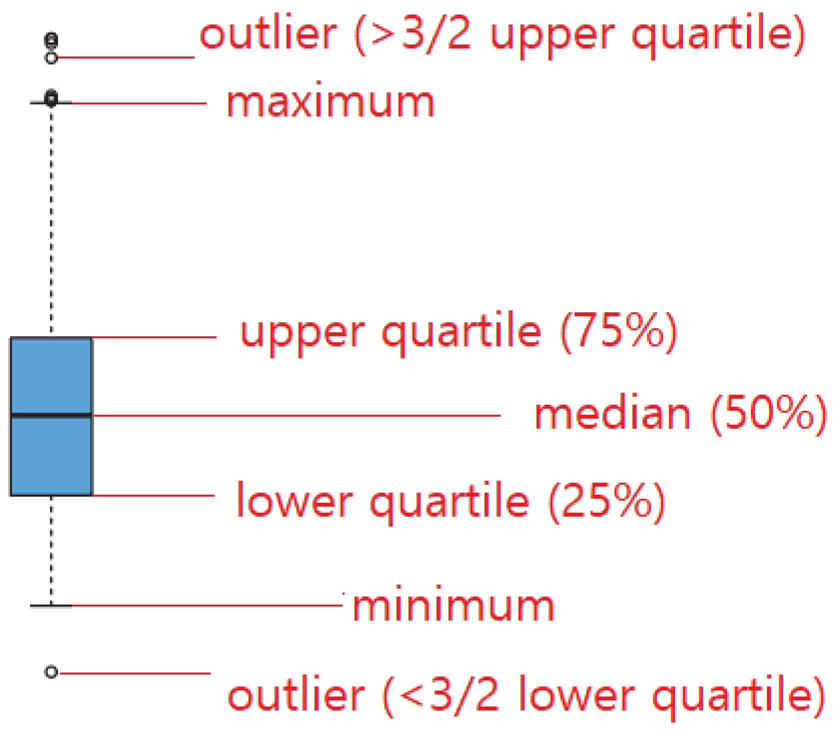
4.1 An Overview of Classification

- classification
  - response Y is qualitive
    - true/false, yes/no, A/B/C
- example - default
  - predictors : income, valance(credit card debt)
  - response : default(不付)

Income, balance, default



Box plot


outlier (>3/2 upper quartile)
maximum
upper quartile (75%)
median (50%)
lower quartile (25%)
minimum
outlier (<3/2 lower quartile)

4.2 Why Not Linear Regression?

- example – emergency room(急诊室)
  - predictors - symptoms
  - response – stroke, drug overdose, epileptic seizure

  - Is there any ordering in responses?

4.3 Logistic Regression

- logistic regression models the probability that Y belongs to a particular category

- example - default
  - predictor : valance(credit card debt)
  - response : default
  - logistic regression models

$$\Pr(\text{default} = \text{yes} \mid \text{balance})$$

4.3 Logistic Regression

- 4.3.1 The Logistic Model
- 4.3.2 Estimating the Regression Coefficients
- 4.3.3 Making Predictions
- 4.3.4 Multiple Logistic Regression
- 4.3.5 Logistic Regression for >2 Response Classes

4.3.1 The Logistic Model

- problem
  - predictor X
  - response Y with $Y = 0$ or 1
  - let $p(X) = \mathrm{Pr}(Y = 1 \mid X)$ for simplicity
  - find a relationship between $p(X)$ and X

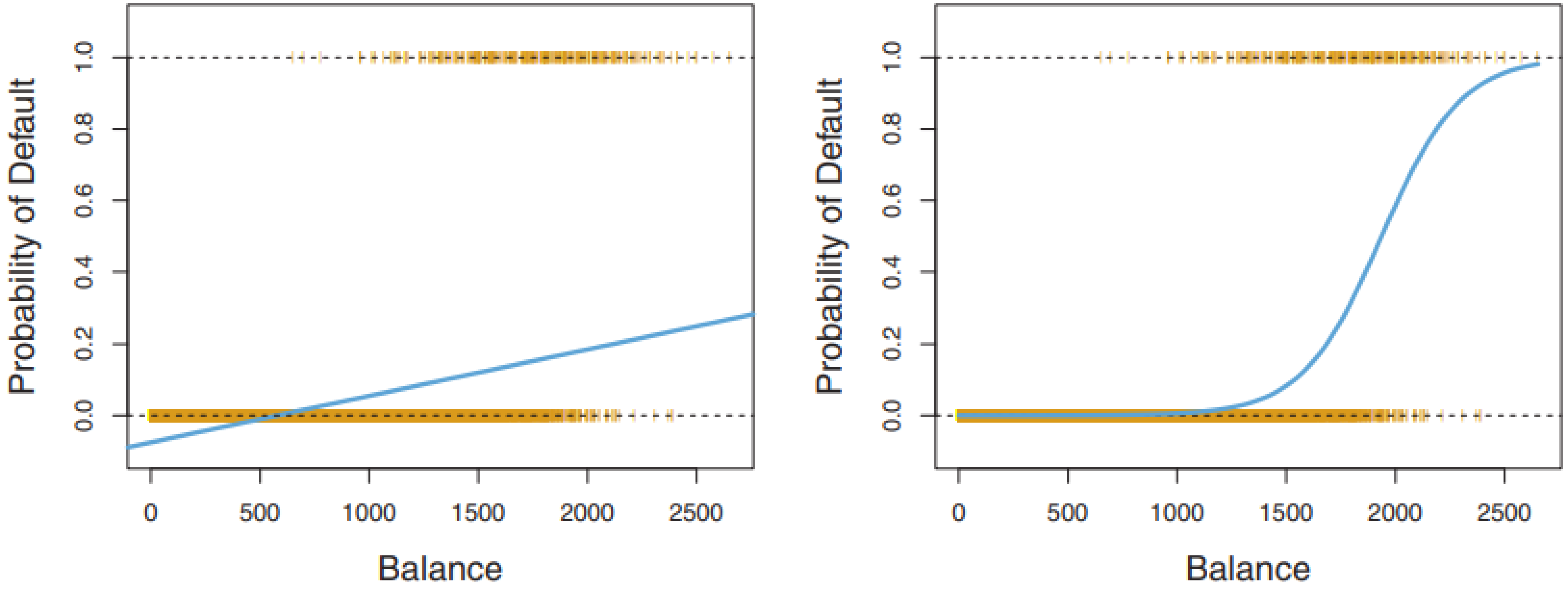
- linear regression model

$$p(X) = \beta_0 + \beta_1 X$$

- logistic regression model

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

linear model vs logistic model



- linear regression model is not suitable for probability, because

$$0 \leq p(X) \leq 1$$

Some calculations

- $p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} = \frac{1}{1+\frac{1}{e^{\beta_0+\beta_1 X}}}$
- $e^{\beta_0+\beta_1 X} = \frac{p(X)}{1-p(X)}$
  - $\frac{p(X)}{1-p(X)}$ is called odds
- $\log \frac{p(X)}{1-p(X)} = \beta_0 + \beta_1 X$
  - $\log \frac{p(X)}{1-p(X)}$ is called log-odds or logit

4.3.2 Estimating the Regression Coefficients

- estimating β_0, β_1
  - linear regression for the model

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

  - maximum likelihood(ML)
    - maximize the likelihood function

$$l(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i:y_i=0} \left(1 - p(x_i)\right)$$

Example: Linear Regression

- data

X	1	1	1	2	2	3	3	3	3
Y	0	0	1	0	1	0	1	1	1

- probabilities

$$p(X=1)=1/3,\ p(X=2)=1/2,\ p(X=3)=3/4$$

$$p(X) = \Pr(Y=1\,|\,X)$$

- log-odds

$$\log\frac{1/3}{1-1/3} = -\log 2,\ \log\frac{1/2}{1-1/2} = 0,\ \log\frac{3/4}{1-3/4} = \log 3$$

$$\log\frac{p(X)}{1-p(X)}$$

- regression

$$\hat{\beta}_0 = -1.66,\ \hat{\beta}_1 = 0.90$$

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

```
X = np.array([[1, 1], [1,2], [1,3]], float)
y = np.array([[-np.log(2)], [0], [np.log(3)]])
beta = matmul(inv(matmul(X.T, X)), matmul(X.T, y))
print(beta)
```

```
[[-1.65660443]
 [ 0.89587973]]
```

Maximum Likelihood

- likelihood function

$$l(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i:y_i=0} \big(1 - p(x_i)\big) = \prod p(x_i)^{y_i} \big(1 - p(x_i)\big)^{1-y_i}$$

- log

$$\log l(\beta_0, \beta_1) = \sum (y_i \log p(x_i) + (1 - y_i) \log(1 - p(x_i)))$$

- maximize $l(\beta_0, \beta_1)$, or equivalently $\log l(\beta_0, \beta_1)$

Example: Maximum Likelihood

- solve
$$\frac{\partial \log l(\beta_0, \beta_1)}{\partial \beta_0} = 0$$
$$\frac{\partial \log l(\beta_0, \beta_1)}{\partial \beta_1} = 0$$
- calculation

let $\sigma(z) = \frac{1}{1+e^{-z}}$ and $z = \beta_0 + \beta_1 x$ then

$$p(x) = \sigma(z) = \sigma(\beta_0 + \beta_1 x)$$
$$\sigma'(z) = \sigma(z)\big(1 - \sigma(z)\big)$$
$$\frac{\partial p(x)}{\partial \beta_0} = \sigma(z)\big(1 - \sigma(z)\big) = p(x)\big(1 - p(x)\big)$$
$$\frac{\partial p(x)}{\partial \beta_1} = \sigma(z)\big(1 - \sigma(z)\big)x = xp(x)\big(1 - p(x)\big)$$

Example: Maximum Likelihood

$$\frac{\partial \log p(x)}{\partial \beta_0} = \frac{p(x)\big(1-p(x)\big)}{p(x)} = 1 - p(x)$$
$$\frac{\partial \log p(x)}{\partial \beta_1} = \frac{xp(x)(1-p(x))}{p(x)} = x\big(1 - p(x)\big)$$
$$\frac{\partial \log(1-p(x))}{\partial \beta_0} = -\frac{p(x)\big(1-p(x)\big)}{1-p(x)} = -p(x)$$
$$\frac{\partial \log(1-p(x))}{\partial \beta_1} = -\frac{xp(x)\big(1-p(x)\big)}{1-p(x)} = -xp(x)$$

- $\frac{\partial \log l(\beta_0,\beta_1)}{\partial \beta_0} = \sum \Big(y_i\big(1 - p(x_i)\big) - (1 - y_i)p(x_i)\Big) = \sum y_i - \sum p(x_i)$
- $\frac{\partial \log l(\beta_0,\beta_1)}{\partial \beta_1} = \sum \Big(x_i y_i\big(1 - p(x_i)\big) - x_i(1 - y_i)p(x_i)\Big) = \sum x_i y_i - \sum x_i p(x_i)$

Example: Maximum Likelihood

- solve

$$\sum y_i = \sum p(x_i)$$
$$\sum x_i y_i = \sum x_i p(x_i)$$

X	1	1	1	2	2	3	3	3	3
Y	0	0	1	0	1	0	1	1	1

- solve

$$5 = 3p(1) + 2p(2) + 4p(3)$$
$$12 = 3p(1) + 4p(2) + 12p(3)$$

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

- solution

```
[-1.64933202  0.89956862]
```

```python
import numpy as np
from scipy.special import expit  # sigmoid

X = np.array([1, 1, 1, 2, 2, 3, 3, 3, 3])
Y = np.array([0, 0, 1, 0, 1, 0, 1, 1, 1])

def log_likelihood(beta):
    p = expit(beta[0] + beta[1] * X)
    return np.sum(Y * np.log(p) + (1 - Y) * np.log(1 - p))

def gradient(beta):
    p = expit(beta[0] + beta[1] * X)
    return np.array([np.sum(Y - p), np.sum(X * (Y - p))])

learning_rate = 0.05
beta = 10 * np.random.rand(2) - 5 # starting point
value = log_likelihood(beta)
while True:
    beta = beta + learning_rate * gradient(beta) # update
    new_value = log_likelihood(beta)
    if np.abs(value - new_value) < 1.0e-13:
        break
    value = new_value

print(beta)
print(new_value)
```

```python
# from mpl_toolkits.mplot3d import Axes3D
import matplotlib.pyplot as plt
from matplotlib import cm

import numpy as np
from scipy.special import expit  # sigmoid

X = np.array([1, 1, 1, 2, 2, 3, 3, 3, 3])
Y = np.array([0, 0, 1, 0, 1, 0, 1, 1, 1])

def log_likelihood(beta):
    p = expit(beta[0] + beta[1] * X)
    return np.sum(Y * np.log(p) + (1 - Y) * np.log(1 - p))

fig = plt.figure()
ax = fig.gca(projection='3d')

Beta0 = np.arange(-5, 5, 0.1)
Beta1 = np.arange(-5, 5, 0.1)
Beta0, Beta1 = np.meshgrid(Beta0, Beta1)
Z = np.empty_like(Beta0)
for i in range(Z.shape[0]):
    for j in range(Z.shape[1]):
        Z[i, j] = log_likelihood([Beta0[i, j], Beta1[i, j]])

ax.plot_surface(Beta0, Beta1, Z, cmap=cm.coolwarm)
plt.show()
```

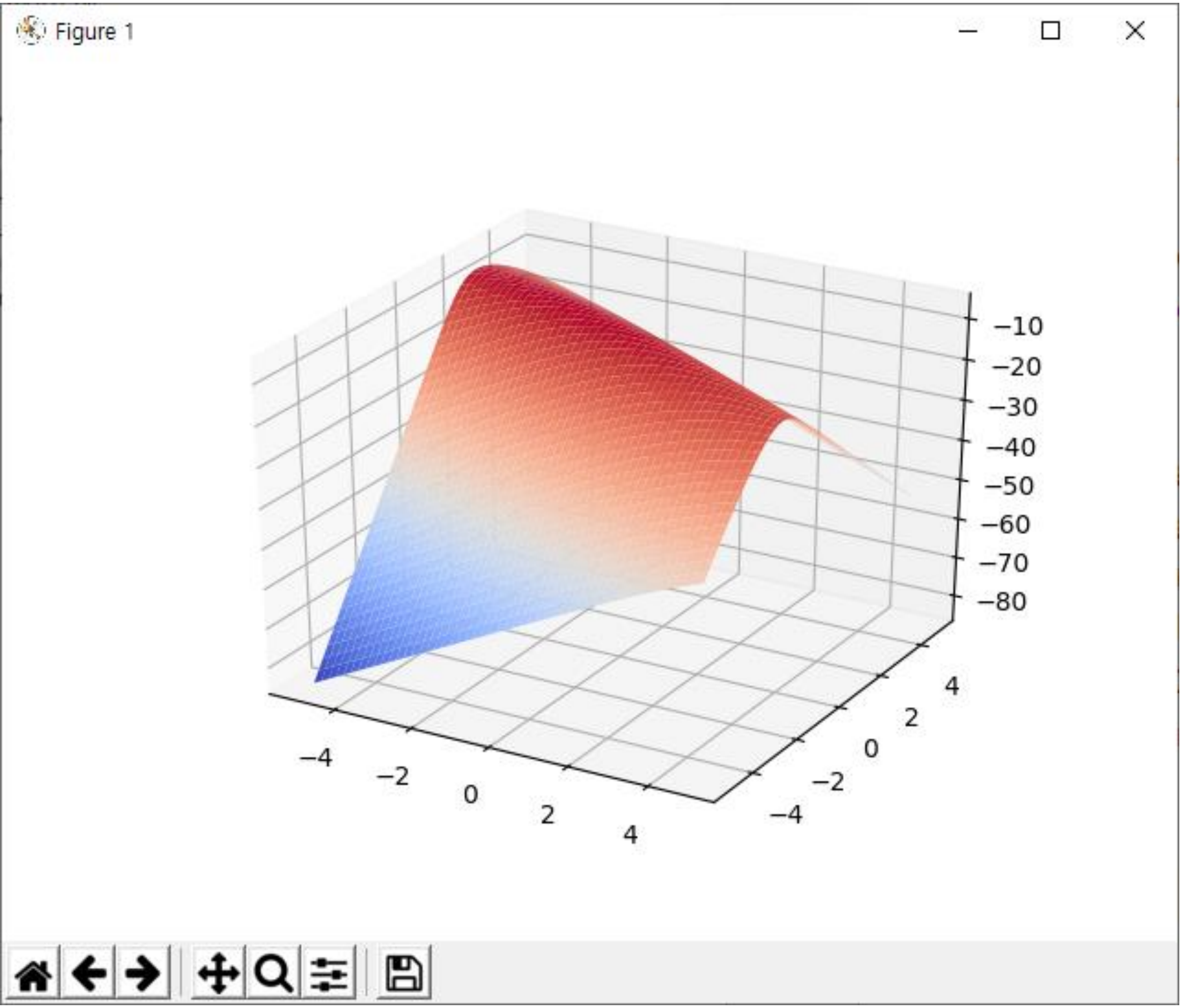


Figure 1
−10
−20
−30
−40
−50
−60
−70
−80
−4
−2
0
2
4
−4
−2
0
2
4

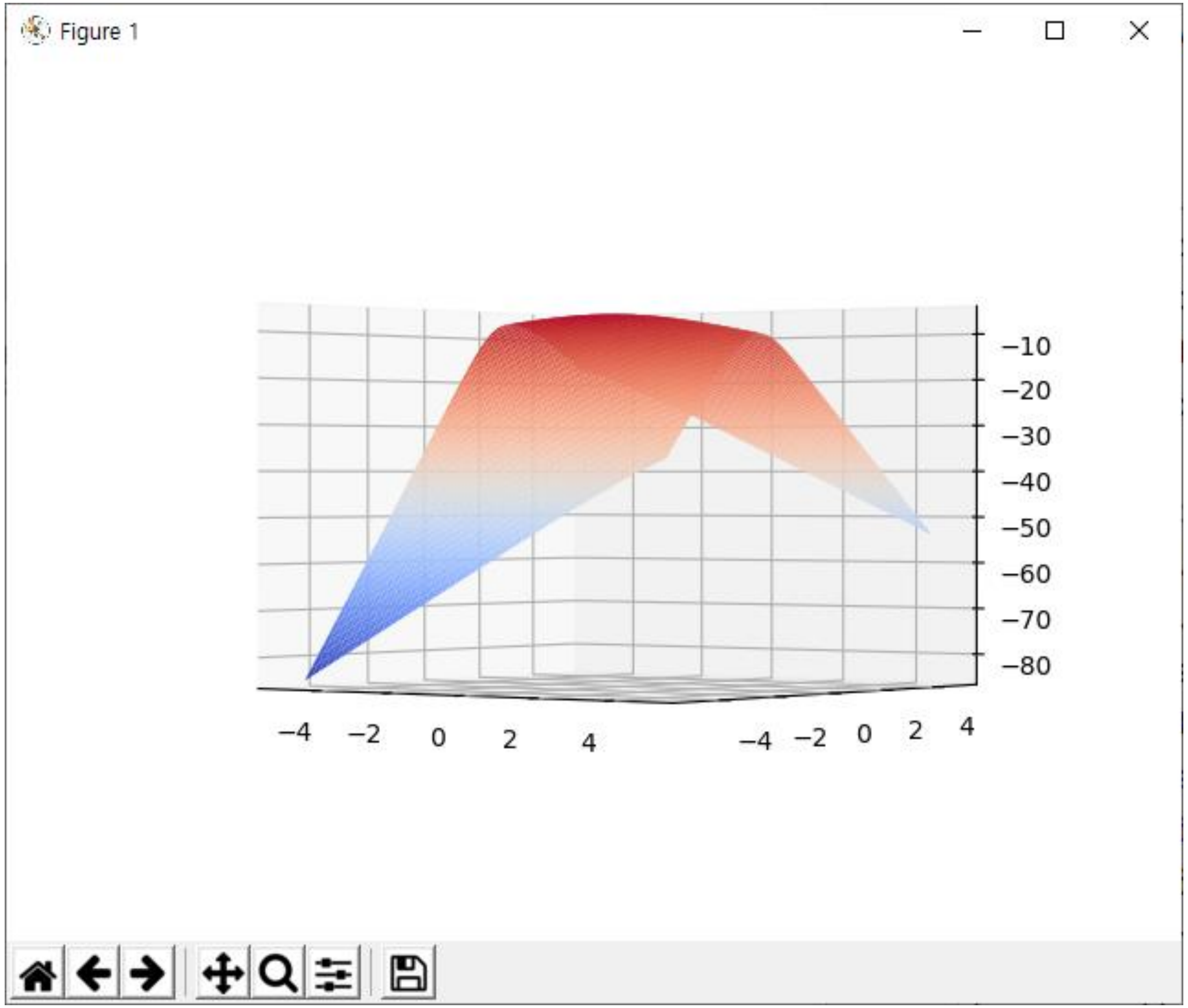


Figure 1
−10
−20
−30
−40
−50
−60
−70
−80
−4
−2
0
2
4
−4
−2
0
2
4

4.3.3 Making Predictions

- prediction

$$\Pr(Y = 1 \mid X = x) = \frac{1}{1+e^{-(-1.65+0.9x)}}$$

  - $X = 1.5$

$$\Pr(Y = 1 \mid X = 1.5) = \frac{1}{1 + e^{0.3}} = 0.43$$

4.3.4 Multiple Logistic Regression

- multiple logistic regression
  - predictors: $X_1, \ldots, X_p$
  - response: Y, binary
- logistic regression model

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

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

4.3.5 Logistic Regression for >2 Response Classes

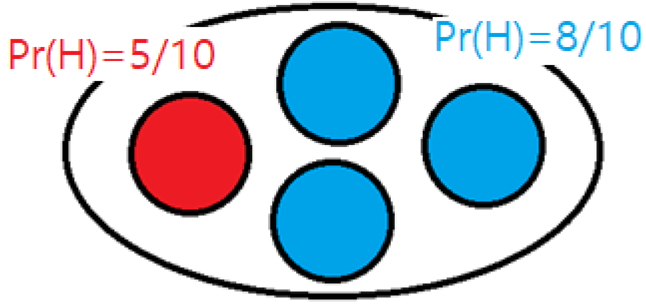
- example – emergency room
  - response - stroke, drug overdose, epileptic seizure.

$$p(X) = \Pr(Y = \text{stroke} \mid X)$$
$$q(X) = \Pr(Y = \text{drug overdose} \mid X)$$
$$1 - p(X) - q(X)$$

4.4 Linear Discriminant Analysis

- 4.4.1 Using Bayes' Theorem for Classification
- 4.4.2 Linear Discriminant Analysis for $p = 1$
- 4.4.3 Linear Discriminant Analysis for $p > 1$
- 4.4.4 Quadratic Discriminant Analysis

Example - coins



- choose and toss
- $X \in \{\text{red}, \text{blue}\}$
- $Y \in \{H, T\}$

	$X = \text{red}$	$X = \text{blue}$	
$Y = H$	$1/4 \times 5/10$	$3/4 \times 8/10$	$29/40$
$Y = T$	$1/4 \times 5/10$	$3/4 \times 2/10$	$11/40$
	$1/4$	$3/4$	

- $\Pr(X = \text{red} \mid Y = T) = ?$

Bayes' Theorem

- joint probability

$$\Pr(X = x_i, Y = y_j)$$

- conditional probability(Bayes' Theorem)

$$\Pr(X = x_i \mid Y = y_j) = \frac{\Pr(X = x_i, Y = y_j)}{\Pr(Y = y_j)}$$

- marginal probability

$$\Pr(X = x_i) = \sum_j \Pr(X = x_i, Y = y_j)$$

	$X = x_1$	$\cdots$	$X = x_m$	
$Y = y_1$	$\Pr(X = x_1, Y = y_1)$	$\cdots$	$\Pr(X = x_m, Y = y_1)$	$\Pr(Y = y_1)$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$Y = y_n$	$\Pr(X = x_1, Y = y_n)$	$\cdots$	$\Pr(X = x_m, Y = y_n)$	$\Pr(Y = y_n)$
	$\Pr(X = x_1)$	$\cdots$	$\Pr(X = x_m)$	

Bayes' Theorem

- $\Pr(X = x_i \mid Y = y_j) = \frac{\Pr(X=x_i, Y=y_j)}{\Pr(Y=y_j)}$
$$= \frac{\Pr(X=x_i, Y=y_j)}{\sum_k \Pr(X=x_k, Y=y_j)}$$
$$= \frac{\Pr(Y=y_j \mid X=x_i)\Pr(X=x_i)}{\sum_k \Pr(Y=y_j \mid X=x_k)\Pr(X=x_k)}$$
- $\Pr(X = x_i \mid Y = y_j)$: posterior
- $\Pr(X = x_i)$: prior
- $\Pr(Y = y_j \mid X = x_i)$: likelihood
- $\Pr(Y = y_j)$: evidence

Example 1.2 (Hamburgers)**.** Consider the following fictitious scientific information: Doctors find that people with Kreuzfeld-Jacob disease (KJ) almost invariably ate hamburgers, thus $p(\textit{Hamburger Eater}|KJ) = 0.9$. The probability of an individual having KJ is currently rather low, about one in 100,000.

1. Assuming eating lots of hamburgers is rather widespread, say $p(\textit{Hamburger Eater}) = 0.5$, what is the probability that a hamburger eater will have Kreuzfeld-Jacob disease?

   This may be computed as

   $$p(KJ\,|\textit{Hamburger Eater}) = \frac{p(\textit{Hamburger Eater}, KJ)}{p(\textit{Hamburger Eater})} = \frac{p(\textit{Hamburger Eater}|KJ)p(KJ)}{p(\textit{Hamburger Eater})} \tag{1.2.1}$$

   $$= \frac{\frac{9}{10} \times \frac{1}{100000}}{\frac{1}{2}} = 1.8 \times 10^{-5} \tag{1.2.2}$$

2. If the fraction of people eating hamburgers was rather small, $p(\textit{Hamburger Eater}) = 0.001$, what is the probability that a regular hamburger eater will have Kreuzfeld-Jacob disease? Repeating the above calculation, this is given by

   $$\frac{\frac{9}{10} \times \frac{1}{100000}}{\frac{1}{1000}} \approx 1/100 \tag{1.2.3}$$

   This is much higher than in scenario (1) since here we can be more sure that eating hamburgers is related to the illness.

Example 1.3 (Inspector Clouseau)**.** Inspector Clouseau arrives at the scene of a crime. The victim lies dead in the room alongside the possible murder weapon, a knife. The Butler (B) and Maid (M) are the inspector's main suspects and the inspector has a prior belief of 0.6 that the Butler is the murderer, and a prior belief of 0.2 that the Maid is the murderer. These beliefs are independent in the sense that $p(B, M) = p(B)p(M)$. (It is possible that both the Butler and the Maid murdered the victim or neither). The inspector's *prior* criminal knowledge can be formulated mathematically as follows:

$$\mathrm{dom}(B) = \mathrm{dom}(M) = \{\mathsf{murderer}, \mathsf{not\ murderer}\}, \mathrm{dom}(K) = \{\mathsf{knife\ used}, \mathsf{knife\ not\ used}\} \tag{1.2.4}$$

$$p(B = \mathsf{murderer}) = 0.6, \qquad p(M = \mathsf{murderer}) = 0.2 \tag{1.2.5}$$

$$\begin{array}{llr} p(\mathsf{knife\ used}|B = \mathsf{not\ murderer}, & M = \mathsf{not\ murderer}) & = 0.3 \\ p(\mathsf{knife\ used}|B = \mathsf{not\ murderer}, & M = \mathsf{murderer}) & = 0.2 \\ p(\mathsf{knife\ used}|B = \mathsf{murderer}, & M = \mathsf{not\ murderer}) & = 0.6 \\ p(\mathsf{knife\ used}|B = \mathsf{murderer}, & M = \mathsf{murderer}) & = 0.1 \end{array} \tag{1.2.6}$$

In addition $p(K, B, M) = p(K|B, M)p(B)p(M)$. Assuming that the knife is the murder weapon, what is the probability that the Butler is the murderer? (Remember that it might be that neither is the murderer). Using b for the two states of B and m for the two states of M,

$$p(B|K) = \sum_m p(B, m|K) = \sum_m \frac{p(B, m, K)}{p(K)} = \frac{\sum_m p(K|B, m)p(B, m)}{\sum_{m,b} p(K|b, m)p(b, m)} = \frac{p(B)\sum_m p(K|B, m)p(m)}{\sum_b p(b)\sum_m p(K|b, m)p(m)} \tag{1.2.7}$$

where we used the fact that in our model $p(B, M) = p(B)p(M)$. Plugging in the values we have (see also `demoClouseau.m`)

$$p(B = \mathsf{murderer}|\mathsf{knife\ used}) = \frac{\frac{6}{10}\left(\frac{2}{10} \times \frac{1}{10} + \frac{8}{10} \times \frac{6}{10}\right)}{\frac{6}{10}\left(\frac{2}{10} \times \frac{1}{10} + \frac{8}{10} \times \frac{6}{10}\right) + \frac{4}{10}\left(\frac{2}{10} \times \frac{2}{10} + \frac{8}{10} \times \frac{3}{10}\right)} = \frac{300}{412} \approx 0.73 \tag{1.2.8}$$

Hence knowing that the knife was the murder weapon strengthens our belief that the butler did it.

Example 1.7 (Soft XOR Gate).

A standard XOR logic gate is given by the table on the right. If we observe that the output of the XOR gate is 0, what can we say about A and B? In this case, either A and B were both 0, or A and B were both 1. This means we don't know which state A was in – it could equally likely have been 1 or 0.

A	B	A xor B
0	0	0
0	1	1
1	0	1
1	1	0

Consider a 'soft' version of the XOR gate given on the right, so that the gate stochastically outputs $C = 1$ depending on its inputs, with additionally $A \perp\!\!\!\perp B$ and $p(A = 1) = 0.65$, $p(B = 1) = 0.77$. What is $p(A = 1|C = 0)$?

A	B	$p(C = 1\|A, B)$
0	0	0.1
0	1	0.99
1	0	0.8
1	1	0.25

$$
\begin{aligned}
p(A=1, C=0) &= \sum_B p(A=1, B, C=0) = \sum_B p(C=0|A=1, B)p(A=1)p(B) \\
&= p(A=1)\left(p(C=0|A=1, B=0)p(B=0) + p(C=0|A=1, B=1)p(B=1)\right) \\
&= 0.65 \times (0.2 \times 0.23 + 0.75 \times 0.77) = 0.405275
\end{aligned} \tag{1.2.20}
$$

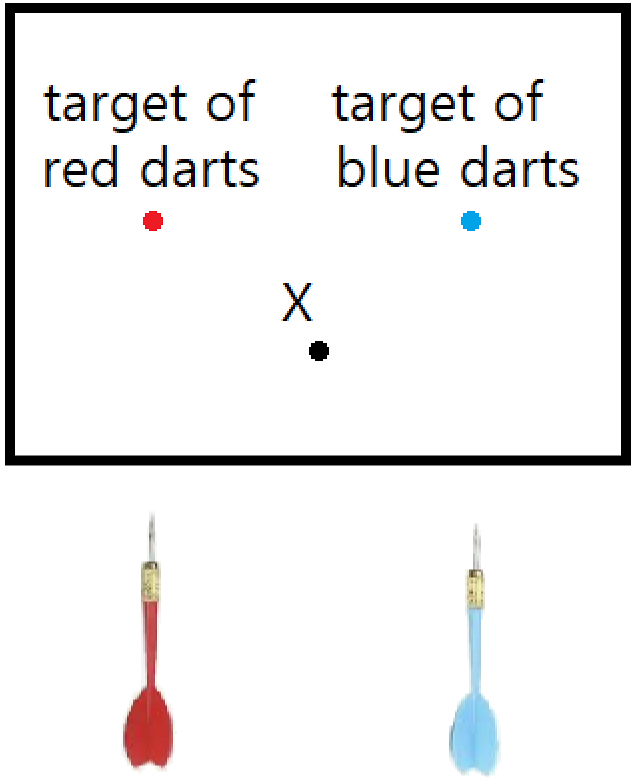
$$
\begin{aligned}
p(A=0, C=0) &= \sum_B p(A=0, B, C=0) = \sum_B p(C=0|A=0, B)p(A=0)p(B) \\
&= p(A=0)\left(p(C=0|A=0, B=0)p(B=0) + p(C=0|A=0, B=1)p(B=1)\right) \\
&= 0.35 \times (0.9 \times 0.23 + 0.01 \times 0.77) = 0.075145
\end{aligned}
$$

Then

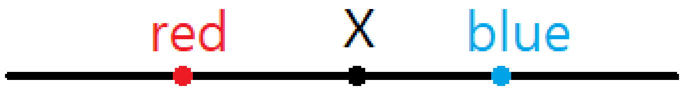
$$
p(A=1|C=0) = \frac{p(A=1, C=0)}{p(A=1, C=0) + p(A=0, C=0)} = \frac{0.405275}{0.405275 + 0.075145} = 0.8436 \tag{1.2.21}
$$

Example: dart

- Problem
  - Infer the color of the dart that hit X
- Approach
  - Compute the Bayes classifier
    $\Pr(Y = \text{red} \mid X)$ or $\Pr(Y = \text{blue} \mid X)$
  - Choose the color which gains the larger value



1-dimensional dart



- Apply Bayes' Theorem

$$\Pr(Y = \text{red} \mid X) = \frac{\Pr(X|Y = \text{red})\Pr(Y = \text{red})}{\Pr(X)}$$

- Assume
  - $\Pr(Y = \text{red}) = \Pr(Y = \text{blue}) = 1/2$
  - $\Pr(X|Y = \text{color})$ is a normal distribution with mean $= \mu_{\text{color}}$, variance $= \sigma^2$.

$$\Pr(X = x|Y = \text{red}) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}(x - \mu_{\text{red}})^2\right)$$

$$\Pr(X = x|Y = \text{blue}) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}(x - \mu_{\text{blue}})^2\right)$$

- Computation

$$\log \Pr(Y = \text{red}|X = x) = \log \Pr(X = x|Y = \text{red}) + \log \Pr(Y = \text{red}) - \log \Pr(X = x)$$

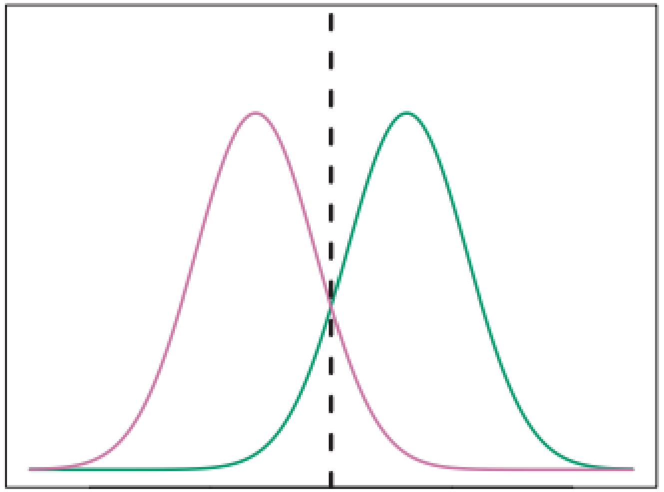
$$= x\frac{\mu_{\text{red}}}{\sigma^2} - \frac{\mu_{\text{red}}^2}{2\sigma^2} + \text{const.}$$

$$\log \Pr(Y = \text{blue}|X = x) = x\frac{\mu_{\text{blue}}}{\sigma^2} - \frac{\mu_{\text{blue}}^2}{2\sigma^2} + \text{const.}$$

- Decision boundary

$$\log \Pr(Y = \text{red}|X = x) = \log \Pr(Y = \text{blue}|X = x)$$

$$x = \frac{\mu_{\text{red}} + \mu_{\text{blue}}}{2}$$



$$\frac{\mu_{\text{red}} + \mu_{\text{blue}}}{2}$$

Linear discriminant analysis(LDA)

- Estimate the mean μ_{color} and the variance σ^2.
  - Observations
    red darts: x_1, …, x_m
    blue darts: x_{m+1}, …, x_{m+n}
  - mean
    $\hat{\mu}_{\text{red}} = \frac{x_1+\cdots+x_m}{m}$, $\hat{\mu}_{\text{blue}} = \frac{x_{m+1}+\cdots+x_{m+n}}{n}$
  - variance
    $\hat{\sigma}^2 = \frac{1}{m+n-2}\left(\sum_{i=1}^{m}(x_i - \hat{\mu}_{\text{red}})^2 + \sum_{i=1}^{n}(x_{m+i} - \hat{\mu}_{\text{blue}})^2\right)$

4.4.1 Using Bayes' Theorem for Classification

- predictor $X \in \mathbb{R}^p$
  - p predictor variables
- response $Y \in \{1, \ldots, K\}$
  - K classes
- Bayes classifier

$$\Pr(Y = k \mid X = x) = \frac{\Pr(X=x|Y=k)\,\Pr(Y=k)}{\Pr(X=x)}$$

- Notation

$p_k(x) = \Pr(Y = k \mid X = x)$ — posterior

$\pi_k = \Pr(Y = k)$ — prior

$f_k(x) = \Pr(X = x|Y = k)$ — likelihood or density function

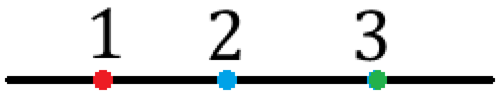
- Bayes classifier

$$p_k(x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

- Note

  posterior ~ prior · likelihood

4.4.2 Linear Discriminant Analysis for $p = 1$



- predictor $X \in \mathbb{R}$
- response $Y \in \{1, \dots, K\}$
- Bayes classifier

$$\Pr(Y = k|X)$$

- inference

$$\mathrm{y} = \underset{k}{\operatorname{argmax}} \Pr(Y = k|X)$$

- Bayes' Theorem

$$\Pr(Y = k|\, X) = \frac{\Pr(X|Y = k)\Pr(Y = k)}{\Pr(X)}$$

- Observations

$$(x_1, y_1), \dots, (x_n, y_n)$$

$n_k = \#$ of k, $\sum n_k = n$

- Assume

$$\Pr(X = x | Y = k) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x - \mu_k)^2\right)$$

$$\sigma = \sigma_1 = \cdots = \sigma_k$$

- Estimation - LDA

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

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

$$\pi_k \sim \hat{\pi}_k = n_k / n$$

- Computation

$$\log \Pr(Y = k|X = x) = \log \Pr(X = x|Y = k) + \log \Pr(Y = k) + \text{const}$$
$$\log p_k(x) = \log f_k(x) + \log \pi_k + \text{const}$$
$$\log p_k(x) \sim \hat{\delta}_k(x) = x\frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log \hat{\pi}_k$$
$$y = \underset{k}{\operatorname{argmax}}\, \hat{\delta}_k(x)$$

- Discriminant function

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

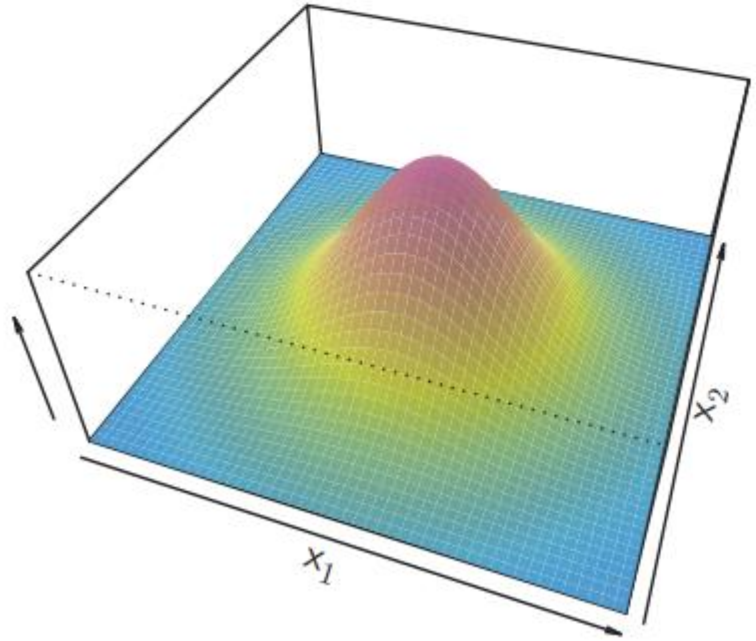
4.4.3 Linear Discriminant Analysis for $p > 1$

- $X \in \mathbb{R}^p$
- $Y \in \{1, \ldots, k\}$

Multivariate Gaussian distribution

- $X \sim N(\mu, \Sigma)$
  - $\mu = E(X)$ – mean
  - $\Sigma = \mathrm{cov}(X)$ – $p \times p$ covariance matrix of X
  - density function

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



Covariance Matrix

- $\operatorname{cov}(X) = \mathrm{E}\left[\left(X - \mathrm{E}(X)\right)\left(X - \mathrm{E}(X)\right)^T\right]$
- $X = \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix}$, $\mathrm{E}[X] = \begin{bmatrix} \mathrm{E}[X_1] \\ \vdots \\ \mathrm{E}[X_p] \end{bmatrix}$
- $\operatorname{cov}(X) = \mathrm{E}[(X - \mathrm{E}[X])(X - \mathrm{E}[X])^T]$

$$= \begin{bmatrix} \mathrm{E}[(X_1 - \mathrm{E}[X_1])(X_1 - \mathrm{E}[X_1])] & \cdots & \mathrm{E}\left[(X_1 - \mathrm{E}[X_1])\left(X_p - \mathrm{E}\left[X_p\right]\right)\right] \\ \vdots & \ddots & \vdots \\ \mathrm{E}\left[\left(X_p - \mathrm{E}\left[X_p\right]\right)(X_1 - \mathrm{E}[X_1])\right] & \cdots & \mathrm{E}\left[\left(X_p - \mathrm{E}\left[X_p\right]\right)\left(X_p - \mathrm{E}\left[X_p\right]\right)\right] \end{bmatrix}$$

LDA

- Assume

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

$$\Sigma = \Sigma_1 = \cdots = \Sigma_K$$

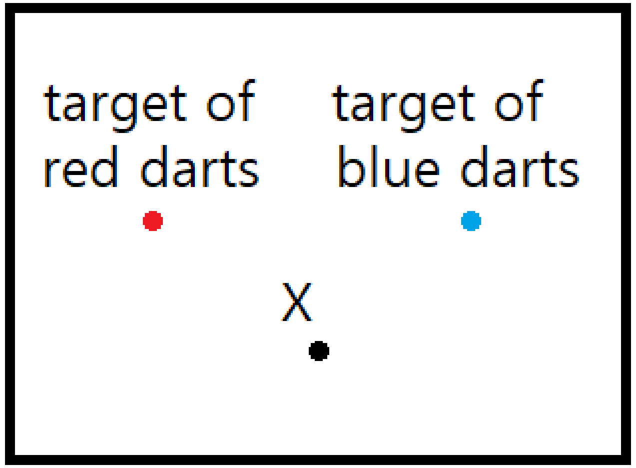
- log-posterior

$$\log f_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2}\mu_k^T \Sigma^{-1} \mu_k + \log \pi_k + \text{const}$$

- Discriminant

$$\hat{\delta}_k(x) = x^T \hat{\Sigma}^{-1} \hat{\mu}_k - \frac{1}{2}\hat{\mu}_k^T \hat{\Sigma}^{-1} \hat{\mu}_k + \log \hat{\pi}_k$$

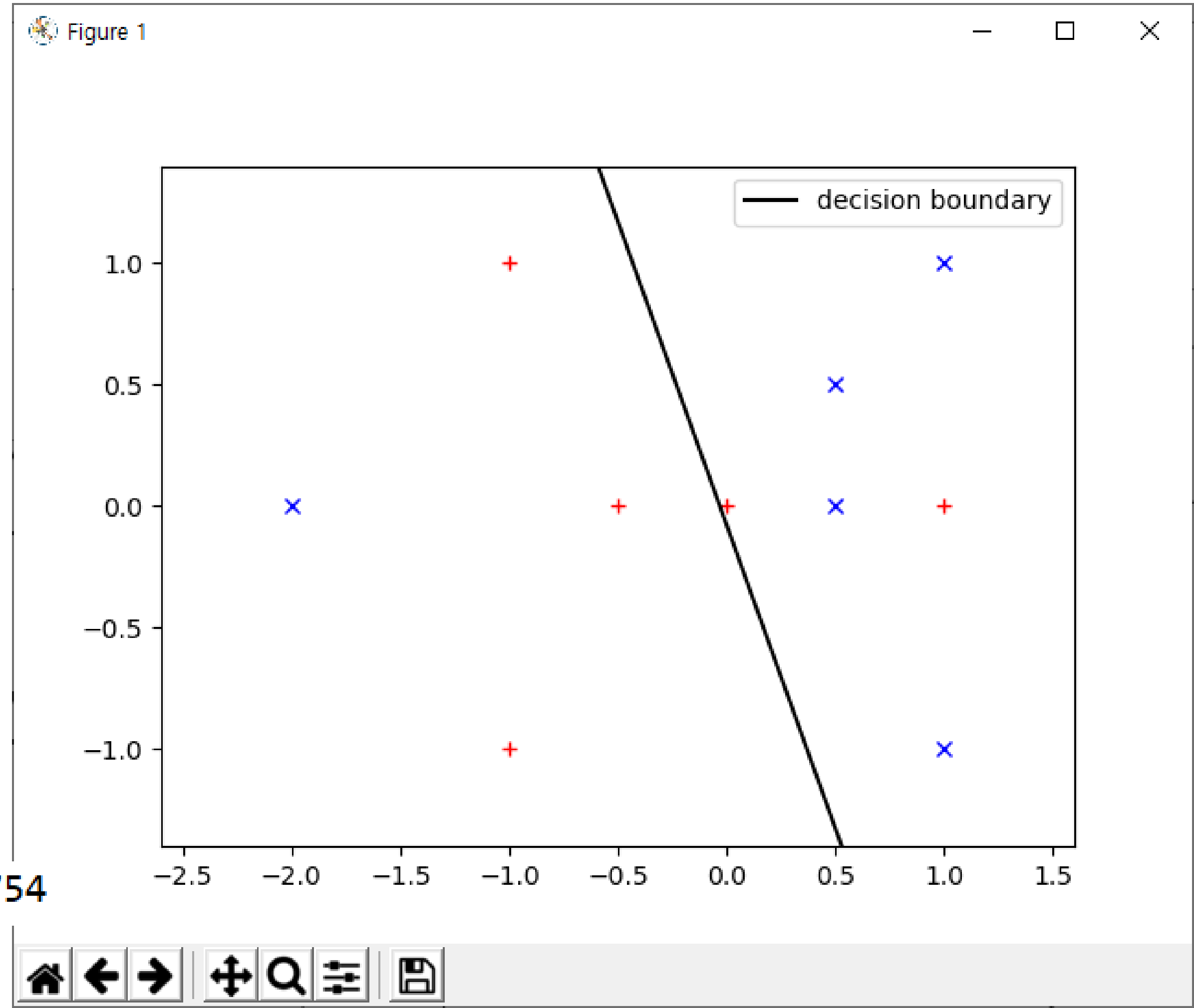
2-dimensional dart



- Observations
  - red: $(-0.5, 0.0), (-1.0, -1.0), (1.0, 0.0), (-1.0, 1.0), (0.0, 0.0)$
  - blue: $(0.5, 0.0), (1.0, -1.0), (-2.0, 0.0), (1.0, 1.0), (0.5, 0.5)$
- Computations
  - $\hat{\mu}_1 = (-0.3, 0.0),\ \hat{\mu}_2 = (0.2, 0.1)$
  - $\hat{\Sigma}_{11} = \frac{1}{10-2}(0.2^2 + 0.7^2 + 1.3^2 + 0.7^2 + 0.3^2 + 0.3^2 + 0.8^2 + 2.2^2 + 0.8^2 + 0.3^2)$
  - $\hat{\Sigma}_{12} = \frac{1}{10-2}(0.0 + (-0.7) \cdot (-1.0) + 0.0 + (-0.7) \cdot 1.0 + 0.0 + 0.3 \cdot (-0.1) + \cdots)$
  - $\hat{\Sigma}_{21} = \Sigma_{12}$
  - $\hat{\Sigma}_{22} = \frac{1}{10-2}(0.0 + 1.0^2 + \cdots)$
  - $\hat{\Sigma} = \begin{bmatrix} 1.1375 & 0.01875 \\ 0.01875 & 0.525 \end{bmatrix},\ \hat{\Sigma}^{-1} =$

```
[[ 0.87963872 -0.03141567]
 [-0.03141567  1.90588389]]
```

- Discriminant
  - $\hat{\delta}_1(x, y) = -0.26x + 0.01y - 0.04$
  - $\hat{\delta}_2(x, y) = 0.17x + 0.18y - 0.03$
  - $(0,0)$ is blue
- Bayes decision boundary
  - $0.43x + 0.17y + 0.01 = 0$

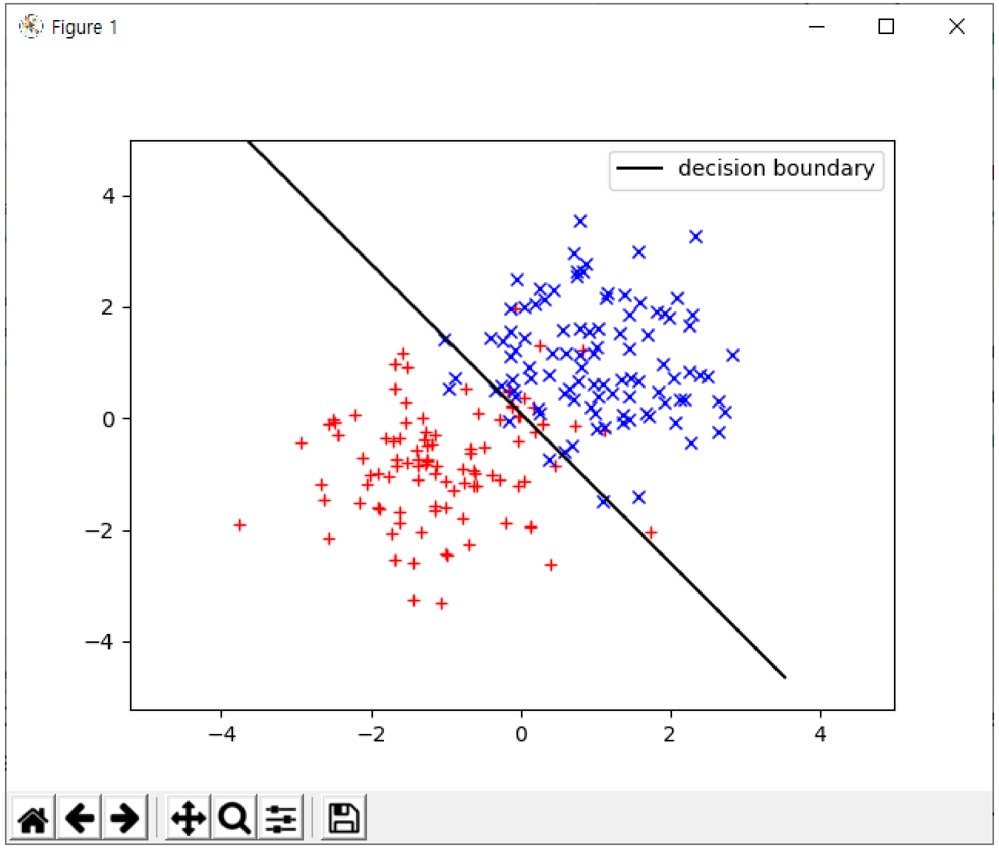



```
slope = -2.497005988023951 , intercept = -0.07485029940119754
```

Generated data

- red $\sim N\big((-1,-1),1\big)$
- blue $\sim N\big((1,1),1\big)$

```
means
    mu1 = [-1.04093541 -0.79564609], mu2 = [1.00551314 1.02143964]
covariance
    [0.85056616 0.03585688]
    [0.03585688 0.99567372]
discriminant
    coef1 = [-1.19193692 -0.75617838], con1 = 0.9211898566020682
    coef2 = [1.14065352 0.98479987], con2 = 1.0764278635214006
decision boundary
    y = -1.3398159485794257 x + 0.08916711413748392
error rate: 0.09
```



4.4.4 Quadratic Discriminant Analysis

- Bayes classifier

$$\Pr(Y = k | X) = \frac{\Pr(X|Y = k)\Pr(Y = k)}{\Pr(X)}$$

- log

$$\log p_k(x) = \log f_k(x) + \log \pi_k + \text{const}$$

- QDA
  - assume

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

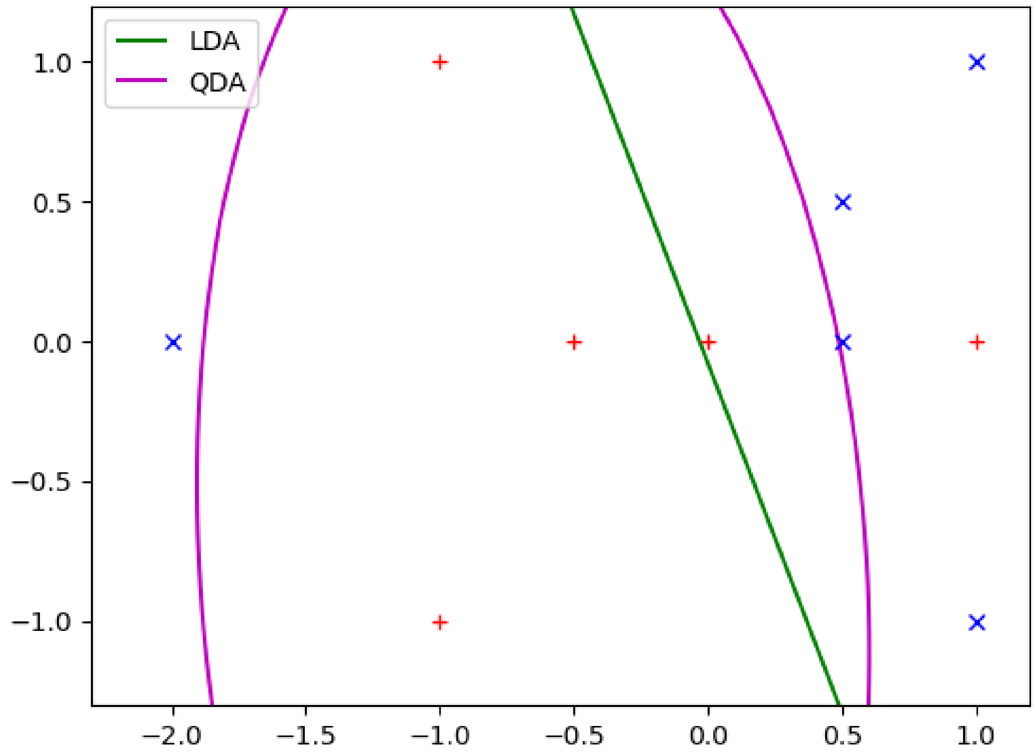
- discriminant

$$\log p_k(x) = -\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) + \log \pi_k - \frac{1}{2}\log|\Sigma_k| + \text{const}$$

$$\delta_k(x) = -\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) + \log \pi_k - \frac{1}{2}\log|\Sigma_k|$$

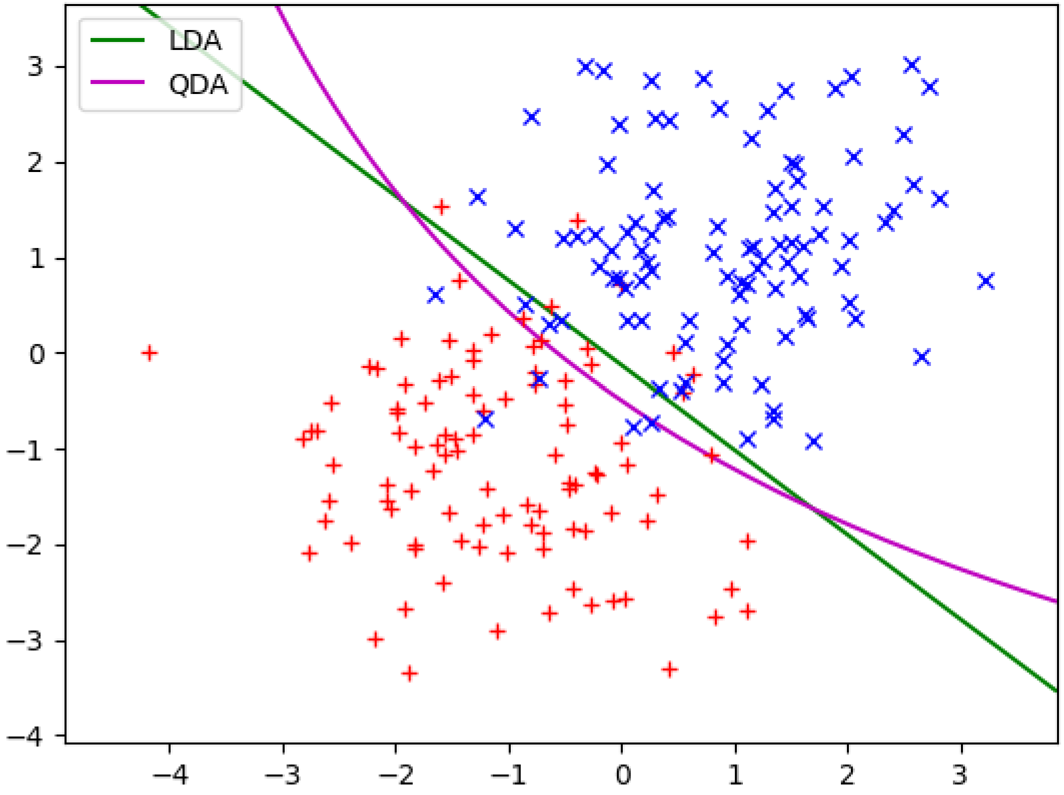
$$= -\frac{1}{2}x^T\Sigma_k^{-1}x + x^T\Sigma_k^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma_k^{-1}\mu_k + \log \pi_k - \frac{1}{2}\log|\Sigma_k|$$

- The discriminant $\delta_k(x)$ is a quadratic form



Generated data

- red $\sim N\big((-1,-1),1\big)$
- blue $\sim N\big((1,1),1\big)$




```
means
    [-1.0843 -1.1289], [0.8574 1.0747]
covariance
    [[ 1.0138 -0.14  ], [-0.14   1.039]]
    [[1.0137 0.1351], [0.1351 0.998 ]]
discriminant
    -3.473, [-1.2428 -1.254 ], [[-0.5026 -0.0677], [-0.0677 -0.4904]]
    -2.356, [0.7152 0.98  ], [[-0.5023  0.068 ], [ 0.068  -0.5102]]
decision boundary
    1.117 + 1.958 x + 2.234 y + 0.0002721 x^2 + 0.2714 xy + -0.01981 y^2
training error rate: 0.08
```

4.5 A Comparison of Classification Methods

- Logistic regression
- LDA
- QDA
- KNN