

Statistical Machine Learning

2주차

담당: 11기 명재성

Review

- Risk Function

$$R(\theta, T(X)) = E[L(\tau(\theta), T(X))] \approx \frac{1}{n} L(\tau(\theta), T(X))$$

- Loss Function

$$\begin{aligned} L[\tau(\theta), T(X)] &= \sum (Y_i - \hat{Y}_i)^2 && \Rightarrow SSE \text{ (MSE)} \\ &= \sum |Y_i - \hat{Y}_i| && \Rightarrow SAE \text{ (MAE)} \end{aligned}$$

Review

- Regression

$$Y_i \stackrel{ind}{\sim} (\mu_i(\mathbf{X}_i), \sigma) \quad \text{where} \quad E[Y_i] = \mu_i(\mathbf{X}_i)$$

$$\mu_i(\mathbf{X}_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} = \boldsymbol{\beta}^T \mathbf{X}_i$$


$$\boldsymbol{\mu}(\mathbf{X}) = \mathbf{X} \boldsymbol{\beta}$$

Review

- Likelihood

$$\mathbf{Y} \sim N_n(\boldsymbol{\mu}(\mathbf{X}), \sigma^2 \mathbf{I}) \quad \text{where} \quad E[\mathbf{Y}] = \boldsymbol{\mu}(\mathbf{X}) = \mathbf{X} \boldsymbol{\beta}$$

$$L(\boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}} |\det \Sigma|^{\frac{1}{2}}} \exp \left(-\frac{1}{2} (\mathbf{Y} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{Y} - \boldsymbol{\mu}) \right)$$


$$L(\boldsymbol{\beta}, \sigma) = \frac{1}{(2\pi)^{\frac{n}{2}} |\det \Sigma|^{\frac{1}{2}}} \exp \left(-\frac{1}{2\sigma^2} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \right)$$

Review

- Likelihood

$$l(\boldsymbol{\beta}, \sigma) = \log L(\boldsymbol{\beta}, \sigma) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log |\det \Sigma| - \frac{1}{2\sigma} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

$$\frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\beta}, \sigma) = \mathbf{X}^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \stackrel{set}{=} 0$$

$$\text{Normal equation : } (\mathbf{X}^T \mathbf{X}) \boldsymbol{\beta} = \mathbf{X}^T \mathbf{Y}$$

Review

- Estimation

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum (Y_i - \hat{Y}_i)^2 \Leftrightarrow \underset{\boldsymbol{\beta}}{\operatorname{argmax}} L(\boldsymbol{\beta}, \sigma)$$

Normal equation : $(\mathbf{X}^T \mathbf{X})\boldsymbol{\beta} = \mathbf{X}^T \mathbf{Y}$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

Logistic Regression

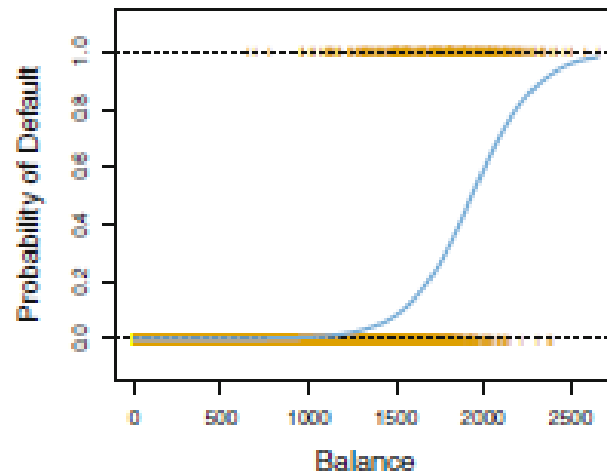
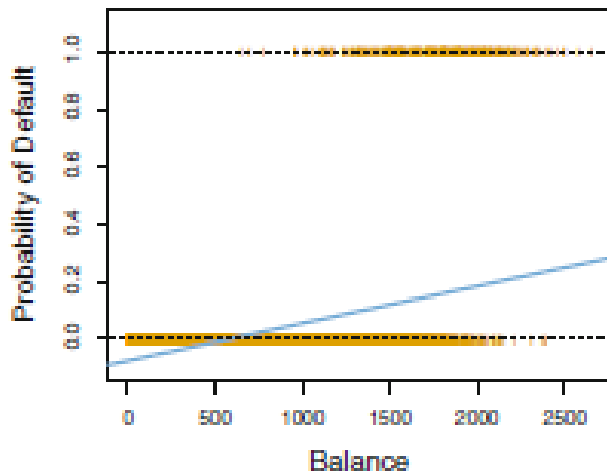
$$Y_i \stackrel{ind}{\sim} \text{Bernoulli}(\pi_i(\mathbf{X}_i)) \quad \text{where} \quad E[Y_i] = \pi_i(\mathbf{X}_i)$$

$$\log \left(\frac{\pi_i(\mathbf{X}_i)}{1 - \pi_i(\mathbf{X}_i)} \right) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi}$$

Logistic Regression

$$\begin{aligned} P(Y_i = 1 | \mathbf{X}_i) &= \pi_i(\mathbf{X}_i) = \frac{e^{\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_p X_{pi}}}{1 + e^{\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_p X_{pi}}} \\ &= \frac{e^{\beta^T \mathbf{X}_i}}{1 + e^{\beta^T \mathbf{X}_i}} = \frac{1}{1 + e^{-\beta^T \mathbf{X}_i}} \quad (\text{sigmoid function}) \end{aligned}$$

Logistic Regression



Logistic Regression

- How to Estimate? $\underset{\beta}{\operatorname{argmax}} L(\beta)$

$$L(\boldsymbol{\pi}; \mathbf{X}) = \prod_{i=1}^n \pi_i^{y_i} (1 - \pi_i)^{1-y_i}$$

$$l(\boldsymbol{\pi}; \mathbf{X}) = \sum_{i=1}^n [y_i \log \pi_i + (1 - y_i) \log(1 - \pi_i)]$$

Logistic Regression

- How to Estimate?

```
> fit.indep = glm(count ~ G + I + H, family=poisson(link=log), data=data2)
```

```
> summary(fit.indep) # loglinear model (G, I, H)
```

Call:

```
glm(formula = count ~ G + I + H, family = poisson(link = log),  
    data = data2)
```

Deviance Residuals:

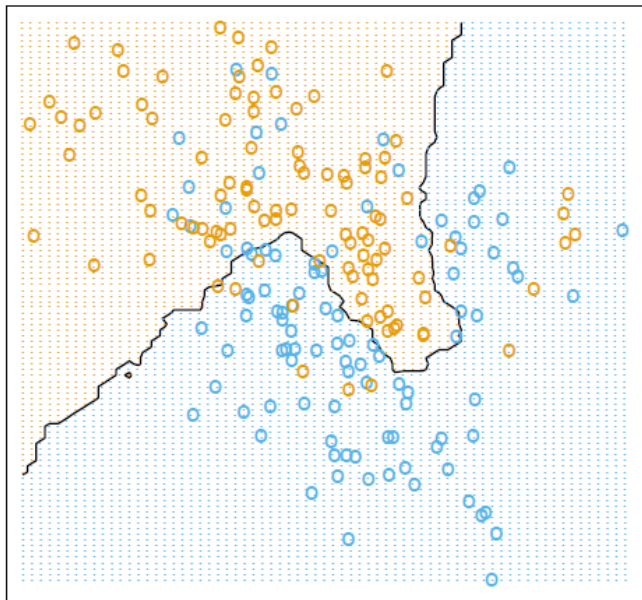
| | | | | | | |
|----------|---------|----------|----------|---------|----------|----------|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| -0.01163 | 0.62672 | -2.14775 | -0.15776 | 1.27750 | -1.49031 | -1.57956 |
| 8 | | | | | | |
| 2.22245 | | | | | | |

Coefficients:

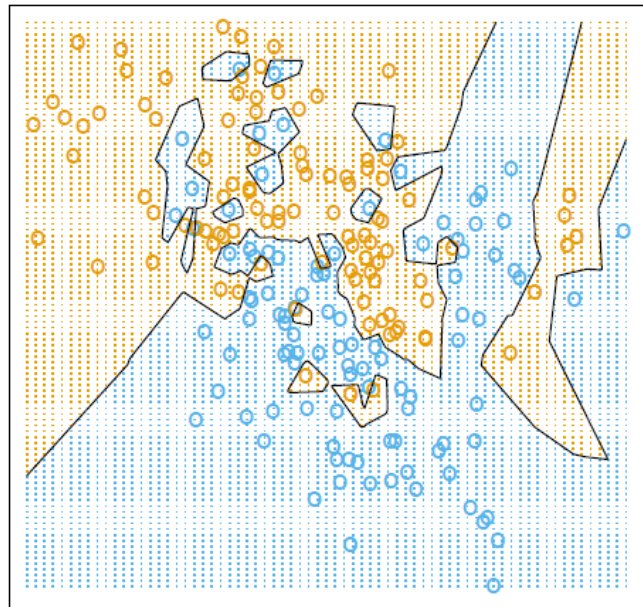
| | Estimate | Std. Error | z value | Pr(> z) |
|-------------|----------|------------|---------|--------------|
| (Intercept) | 3.53231 | 0.11459 | 30.826 | < 2e-16 *** |
| Gmale | -0.28205 | 0.08106 | -3.480 | 0.000502 *** |
| Isupport | 1.77495 | 0.11399 | 15.571 | < 2e-16 *** |
| Hsupport | -0.69315 | 0.08513 | -8.143 | 3.87e-16 *** |

KNN Classifier

15-Nearest Neighbor Classifier



1-Nearest Neighbor Classifier

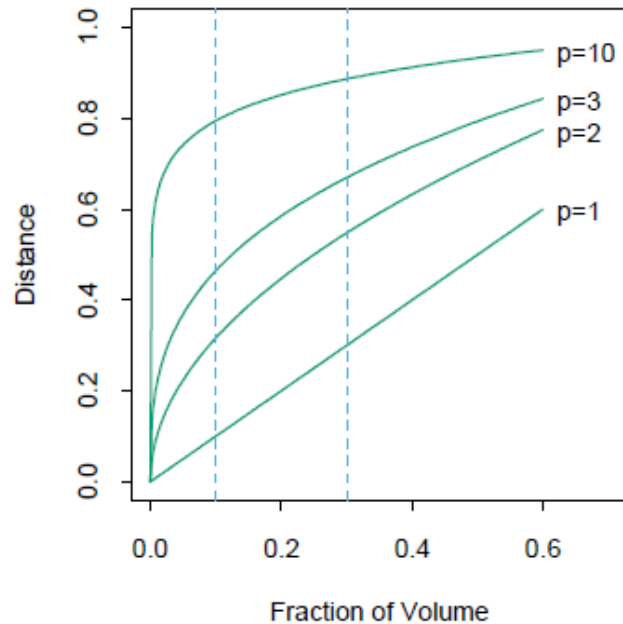
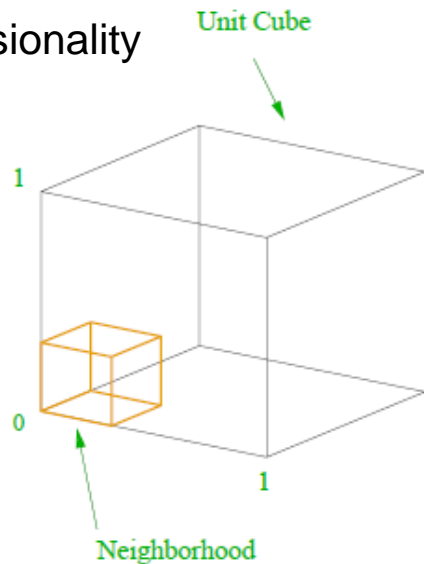


KNN Classifier

- Scenario 1
 - The training data in each class were generated from bivariate Gaussian distributions with uncorrelated components and different means.
- Scenario 2
 - The training data in each class came from a mixture of 10 low-variance Gaussian distributions, with individual means themselves distributed as Gaussian.

KNN Classifier

- Curse of dimensionality



KNN Classifier

- Distance measure

$$d(\mathbf{u}, \mathbf{v}) = (\sum |u_i - v_i|^2)^{\frac{1}{2}} = ||\mathbf{u} - \mathbf{v}||_2 \quad \textit{Euclidean (L2 norm)}$$

$$d(\mathbf{u}, \mathbf{v}) = \sum |u_i - v_i| = ||\mathbf{u} - \mathbf{v}||_1 \quad \textit{Manhattan (L1 norm)}$$

$$d(\mathbf{u}, \mathbf{v}) = (\sum |u_i - v_i|^p)^{\frac{1}{p}} = ||\mathbf{u} - \mathbf{v}||_p \quad \textit{Minkowski (Lp norm)}$$

$$d(\mathbf{u}, \mathbf{v}) = \sqrt{(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})} \quad \textit{Mahalanobis Distance}$$

Kernel Density Estimation

- Kernel function

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma(\mathbf{x}_i - \mathbf{x}_j)^T(\mathbf{x}_i - \mathbf{x}_j))$$

*Gaussian Kernel
(Radial Basis function)*

$$K(\mathbf{x}_i, \mathbf{x}_j) = (1 + \mathbf{x}_i^T \mathbf{x}_j)^p$$

polynomial Kernel

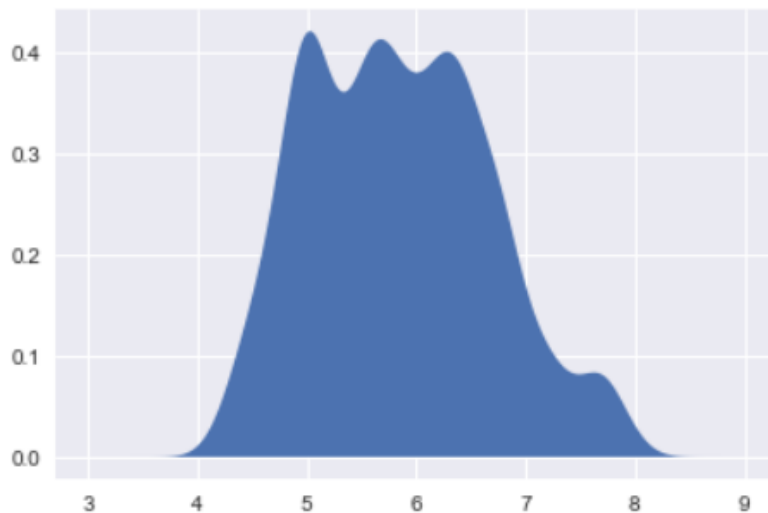
$$K(\mathbf{x}_i, \mathbf{x}_j) = \tanh(k_1 \mathbf{x}_i^T \mathbf{x}_j + k_2)$$

Sigmoid Kernel

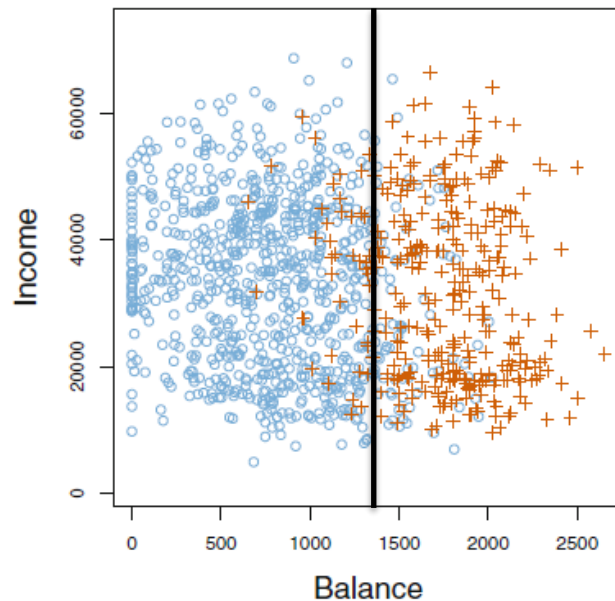
Kernel Density Estimation

- Density estimation at $x = x_0$

$$\hat{f}_X(x_0) = \frac{1}{n} \sum K(x_0, x_i)$$

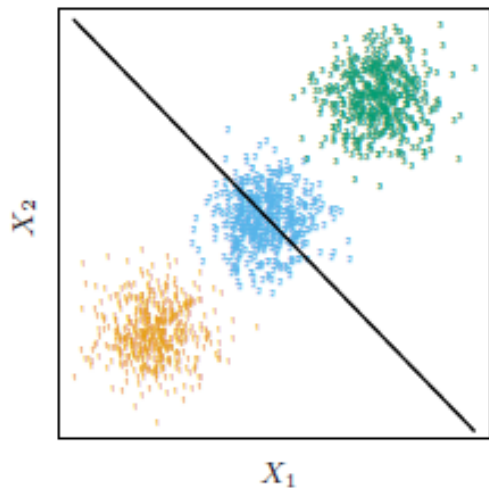


Classification with regression

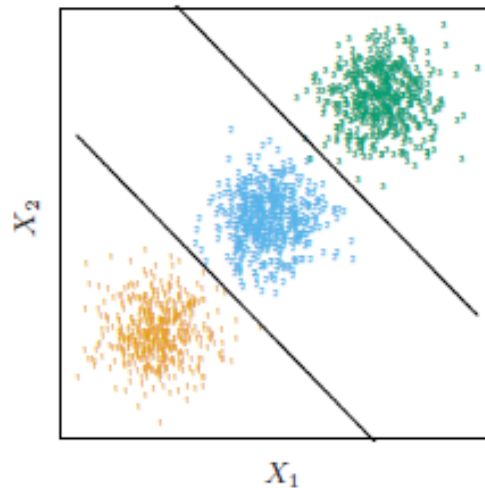


Discriminant Analysis

Linear Regression



Linear Discriminant Analysis



Naïve Bayes Classifier

$$P(Y_i = k | \mathbf{X}_i) = \frac{P(\mathbf{X}_i | k)P(k)}{\sum_k P(\mathbf{X}_i | k)P(k)}$$

Bayes' Theorem

$$\text{where } P(\mathbf{X}_i | k) = \prod_j^p P(X_{ij} | k)$$

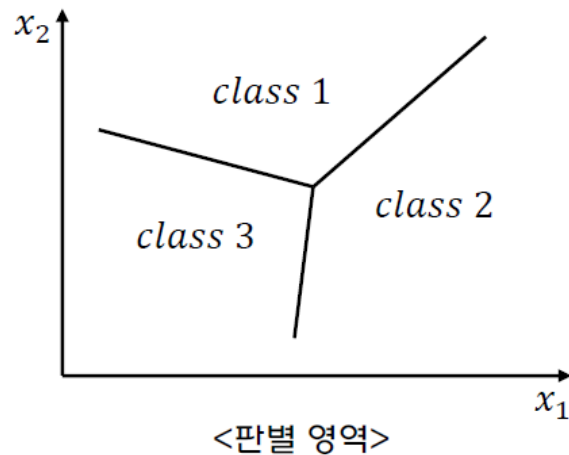
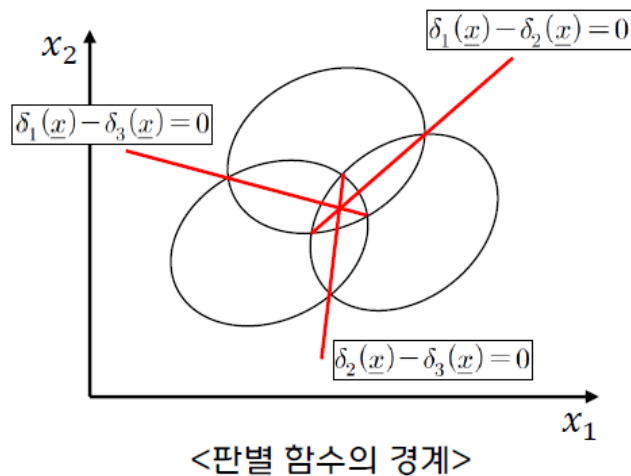
Linear Discriminant Analysis

$$P(Y_i = k | \mathbf{X}_i) = \frac{P(\mathbf{X}_i | k)P(k)}{\sum_k P(\mathbf{X}_i | k)P(k)}$$

Bayes' Theorem

where $P(\mathbf{X}_i | k) \sim N_p(\boldsymbol{\mu}_k, \Sigma)$

Linear Discriminant Analysis



Linear Discriminant Analysis

IF $P(Y_i = k|\mathbf{X}_i) > P(Y_i = l|\mathbf{X}_i) \rightarrow$ *estimate class of Y_i to k*

$$\log \frac{P(Y_i = k|\mathbf{X}_i)}{P(Y_i = l|\mathbf{X}_i)} = \delta_k(\mathbf{X}_i) - \delta_l(\mathbf{X}_i)$$

$$\text{where } \delta_k(\mathbf{X}_i) = \mathbf{X}_i^T \Sigma^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \boldsymbol{\mu}_k^T \Sigma^{-1} \boldsymbol{\mu}_k + \log P(k)$$

Quadratic Discriminant Analysis

$$P(Y_i = k | \mathbf{X}_i) = \frac{P(\mathbf{X}_i | k)P(k)}{\sum_k P(\mathbf{X}_i | k)P(k)}$$

Bayes' Theorem

where $P(\mathbf{X}_i | k) \sim N_p(\boldsymbol{\mu}_k, \Sigma_k)$

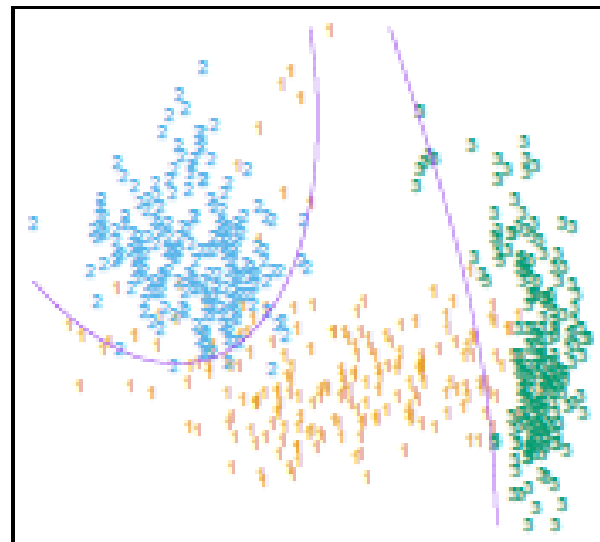
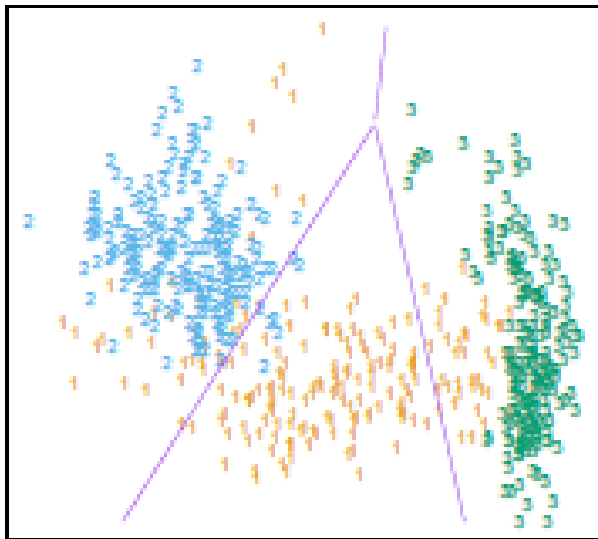
Quadratic Discriminant Analysis

IF $P(Y_i = k|\mathbf{X}_i) > P(Y_i = l|\mathbf{X}_i) \rightarrow$ *estimate class of Y_i to k*

$$\log \frac{P(Y_i = k|\mathbf{X}_i)}{P(Y_i = l|\mathbf{X}_i)} = \delta_k(\mathbf{X}_i) - \delta_l(\mathbf{X}_i)$$

$$\text{where } \delta_k(\mathbf{X}_i) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(\mathbf{X}_i - \boldsymbol{\mu}_k)^T \Sigma_k^{-1}(\mathbf{X}_i - \boldsymbol{\mu}_k) + \log P(k)$$

LDA and QDA



Loss Function for Classification

- 0-1 Loss

$$L[\tau(\theta), T(X)] = \sum I(Y_i \neq \hat{Y}_i)$$

- The Bayes decision rule for minimizing the loss ($P(Y_i \neq \hat{Y}_i)$) is

$$\underset{k}{\operatorname{argmax}} P(Y = k|\mathbf{X})$$

Loss Function for Classification

- Categorical Cross Entropy

$$CE_i = - \sum_{k=1}^C y_{ik} \log \pi_i(k)$$

- Binary Cross Entropy

$$\begin{aligned} CE_i &= -[y_{i1} \log \pi_i(1) + y_{i0} \log \pi_i(0)] \\ &= -[y_i \log \pi_i + (1 - y_i) \log(1 - \pi_i)] \end{aligned}$$

Loss Function for Classification

- Binary Cross Entropy

$$\sum_{i=1}^n CE_i = - \sum_{i=1}^n [y_i \log \pi_i + (1 - y_i) \log(1 - \pi_i)]$$

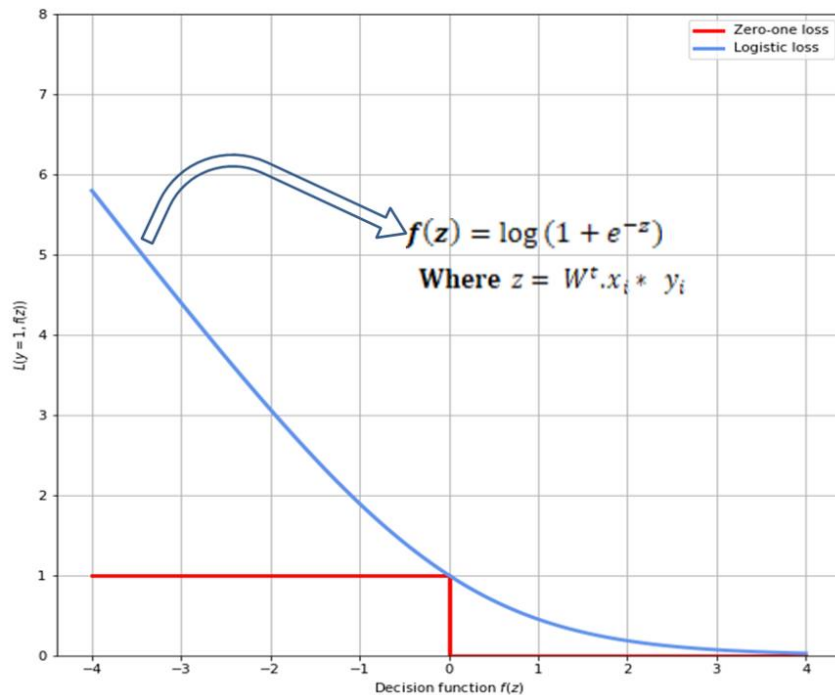
$$l(\boldsymbol{\pi}; \mathbf{X}) = \sum_{i=1}^n [y_i \log \pi_i + (1 - y_i) \log(1 - \pi_i)]$$

Loss Function for Classification

- For Logistic Regression

$$\underset{\beta}{\operatorname{argmin}} \text{ "Cross Entropy"} \Leftrightarrow \underset{\beta}{\operatorname{argmax}} \text{ "Likelihood"}$$

Loss Function for Classification



reference

자료

19-2 STAT424 통계적 머신러닝 - 박유성 교수님

교재

파이썬을 이용한 통계적 머신러닝 (2020) - 박유성

ISLR (2013) - G. James, D. Witten, T. Hastie, R. Tibshirani

The elements of Statistical Learning (2001) - J. Friedman, T. Hastie, R. Tibshirani

Hands on Machine Learning (2017) - Aurelien Geron