Linear Regression / LDA

Linear Methods for Classification

2019.1.30 임소현

Contents

Contents

Classification ?

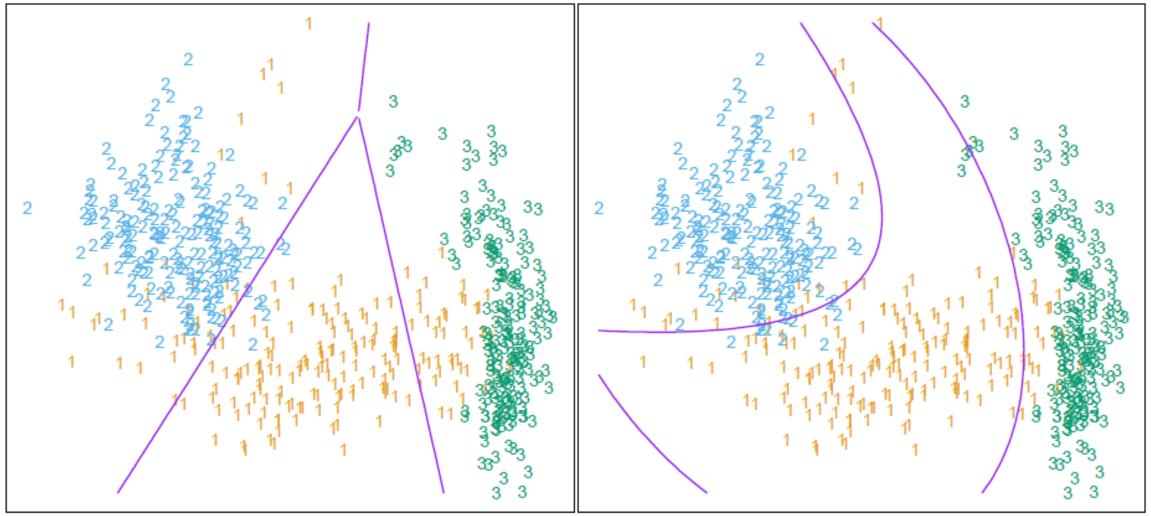
: 입력 벡터 x로부터 이에 대응되는 타겟 클래스 K에 대해 어떤 하나의 클래스에 속하도록 선정하는 작업

- 타겟 t에 대해 K의 크기를 가지는 이진 벡터로 정의

ex > K=5, 클래스가 2에 속하는 경우

$$\mathbf{t} = (0, 1, 0, 0, 0)^T$$

4.1 Introduction



- 클래스는 상호 배타적인 관계
- 이렇게 나누어진 지역을 ***decision region*** 라고 부른다
- 이를 나누는 경계면을 ***decision boundaries*** 라고 부른다

Model

- discriminant functions $\delta_k(x)$ for each class
- Posterior probabilities $\Pr(G = k|X = x)$

$$\Pr(G = 1|X = x) = \frac{\exp(\beta_0 + \beta^T x)}{1 + \exp(\beta_0 + \beta^T x)},$$
$$\Pr(G = 2|X = x) = \frac{1}{1 + \exp(\beta_0 + \beta^T x)}.$$

logit transformation

$$\log \frac{\Pr(G = 1|X = x)}{\Pr(G = 2|X = x)} = \beta_0 + \beta^T x$$

Contents

4.2 Linear Regression of an Indicator Matrix

- there will be K such indicators $Y_k,\ k = 1, \ldots, K,$ with $Y_k = 1$ if $G = k$ else 0
- $Y = (Y_1, \ldots, Y_K)$, ***indicator response matrix*** $(\ N \times K\ )$
- Y is a matrix of 0's and 1's, with each row having a single 1
- Fit a linear regression model

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

- Coefficient matrix

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

4.2 Linear Regression of an Indicator Matrix

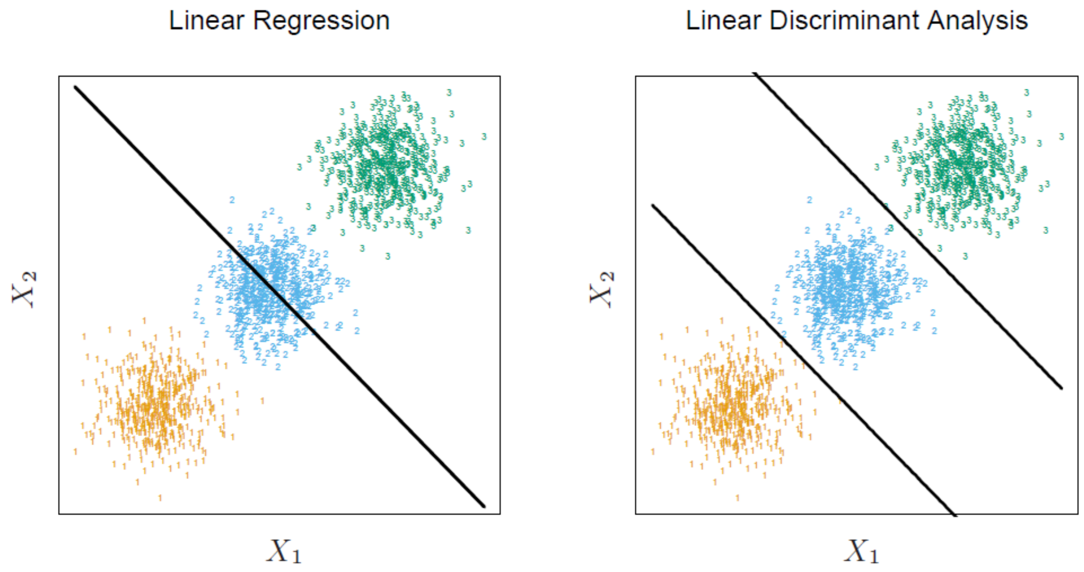
- A new observation with input x is classified as follows:
  - compute the fitted output $\hat{f}(x)^T = (1, x^T)\hat{\mathbf{B}}$, a K vector;
  - identify the largest component and classify accordingly: $\hat{G}(x) = \mathrm{argmax}_{k \in \mathcal{G}} \hat{f}_k(x)$
- The response vector y_i (ith row of $\mathbf{Y}$) for observation i has the value $y_i = t_k$ if $g_i = k$
- Fit the linear model by least squares $\min_{\mathbf{B}} \sum_{i=1}^{N} ||y_i - [(1, x_i^T)\mathbf{B}]^T||^2$
- The criterion is a sum-of-squared Euclidean distances of the fitted vectors from their targets

$$\hat{G}(x) = \underset{k}{\mathrm{argmin}} \, ||\hat{f}(x) - t_k||^2$$

4.2 Linear Regression of an Indicator Matrix

- There is a serious problem with the regression approach when the number of classes K ≥ 3
- Figure illustrates an extreme situation when K = 3.
- The three classes are perfectly separated by linear decision boundaries,

  yet linear regression misses the middle class completely.



Contents

4.3 Linear Discriminant Analysis

- Suppose $f_k(x)$ is the class conditional density of X in class G=k

multivariate Gaussian →

$$f_k(x) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}_k|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T\mathbf{\Sigma}_k^{-1}(x-\mu_k)}$$

- Let π_k be the prior probability of class k, with $\sum_{k=1}^{K} \pi_k = 1$

- By Bayes theorem,

$$\Pr(G=k|X=x) = \frac{f_k(x)\pi_k}{\sum_{\ell=1}^{K} f_\ell(x)\pi_\ell}$$

LDA

- assume that the classes have a common covariance matrix
- Comparing two classes *k* and *l*, log-ratio

$$
\begin{aligned}
\log \frac{\Pr(G=k|X=x)}{\Pr(G=\ell|X=x)} &= \log \frac{f_k(x)}{f_\ell(x)} + \log \frac{\pi_k}{\pi_\ell} \\
&= \log \frac{\pi_k}{\pi_\ell} - \frac{1}{2}(\mu_k + \mu_\ell)^T \boldsymbol{\Sigma}^{-1}(\mu_k - \mu_\ell) \\
&\quad + x^T \boldsymbol{\Sigma}^{-1}(\mu_k - \mu_\ell),
\end{aligned}
$$

LDA

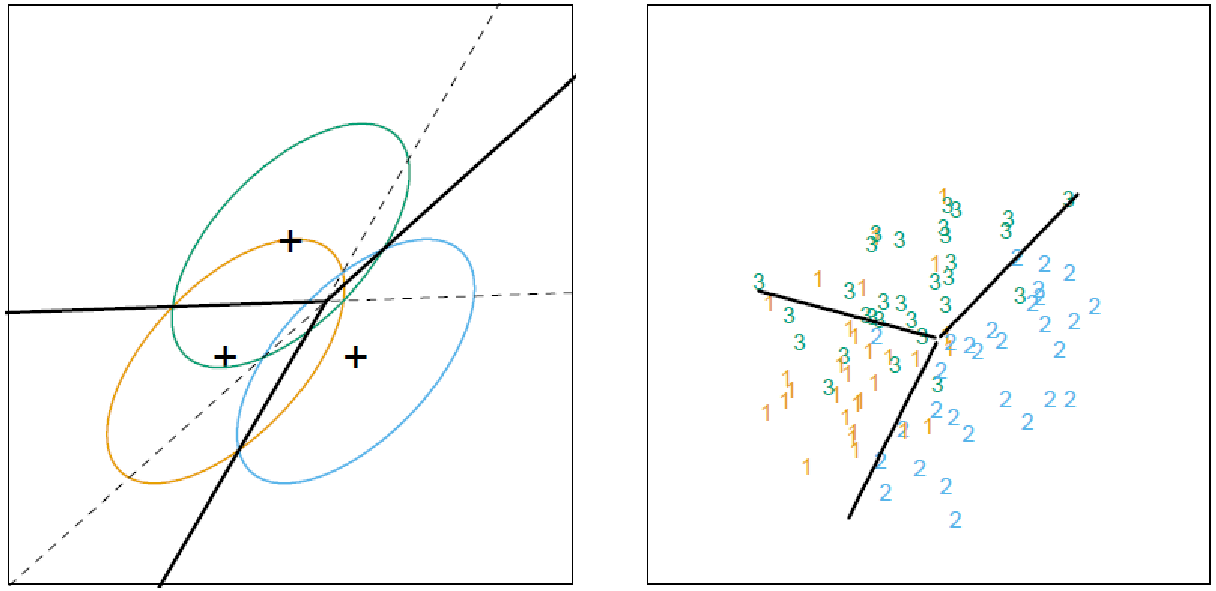
- *linear discriminant functions*

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

- We don't know the parameters of the Gaussian distributions, so we need to estimate them

  - $\hat{\pi}_k = N_k / N$, where N_k is the number of class-k observations;
  - $\hat{\mu}_k = \sum_{g_i = k} x_i / N_k$;
  - $\hat{\boldsymbol{\Sigma}} = \sum_{k=1}^{K} \sum_{g_i = k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T / (N - K)$.

4.3 Linear Discriminant Analysis



- shows three Gaussian distributions, with the same covariance and different means
- included are the contours of constant density enclosing 95% of the probability in each case
- the Bayes decision boundaries separating all three classes are the thicker solid lines

QDA (*quadratic discriminant functions*)

- if the Σ_k are not assumed to be equal
- *linear discriminant functions*

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

- The decision boundary between each pair of classes k and ℓ is described

  by a quadratic equation $\{x : \delta_k(x) = \delta_\ell(x)\}$

RDA (*Regularized Discriminant Analysis*)

- to shrink the separate covariances of QDA toward a common covariance as in LDA
- very similar in flavor to ridge regression

- $\alpha \in [0,1]$ allows a continuum of models between LDA and QDA

$$\hat{\Sigma}_k(\alpha) = \alpha\hat{\Sigma}_k + (1-\alpha)\hat{\Sigma} \qquad (4.13)$$

- Similar modifications to be shrunk toward the scalar covariance

$$\hat{\Sigma}(\gamma) = \gamma\hat{\Sigma} + (1-\gamma)\hat{\sigma}^2\mathbf{I}$$

- Replacing $\hat{\Sigma}$ in (4.13) by $\hat{\Sigma}(\gamma)$ leads to a more general family

Reference

- http://norman3.github.io/prml/docs/chapter04/1
- https://m.blog.naver.com/PostView.nhn?blogId=sw4r&logNo=221033110991&proxyReferer=https%3A%2F%2Fwww.google.com%2F

Thank you :-)