Linear Methods for Classification

4.4 Logistic Regression
4.5 Separating Hyperplanes

19.01.08 임소현

CONTENTS

01. Regularized Logistic Regression

$$\max_{\beta_0,\beta} \left\{ \sum_{i=1}^{N} \left[y_i(\beta_0 + \beta^T x_i) - \log(1 + e^{\beta_0 + \beta^T x_i}) \right] - \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

- L1 penalty used in the lasso can be used for variable selection and shrinkage with any linear regression model
- do not penalize the intercept term, and standardize the predictors for the penalty to be meaningful

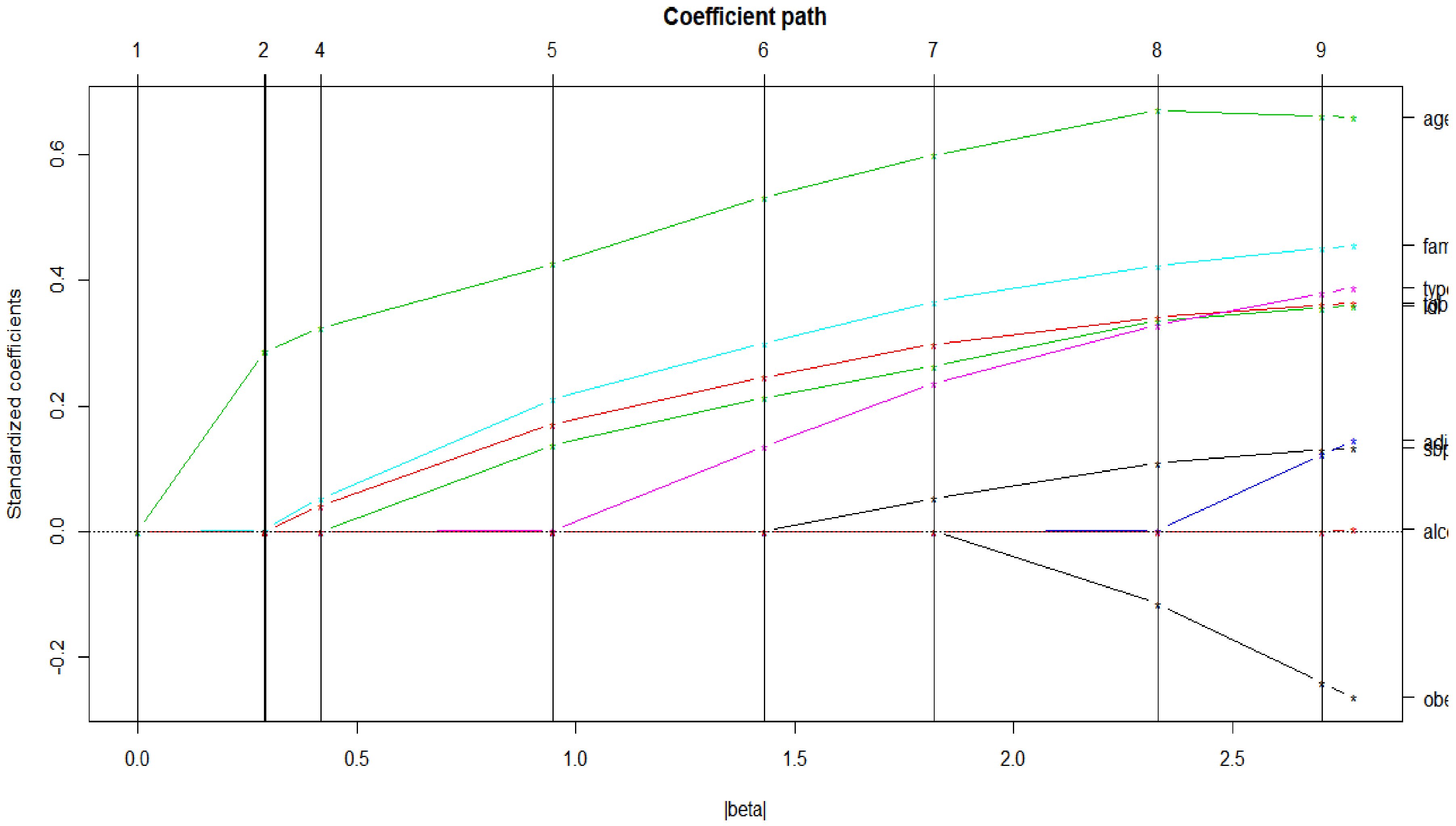
2. the South African heart disease data

02. the South African heart disease data

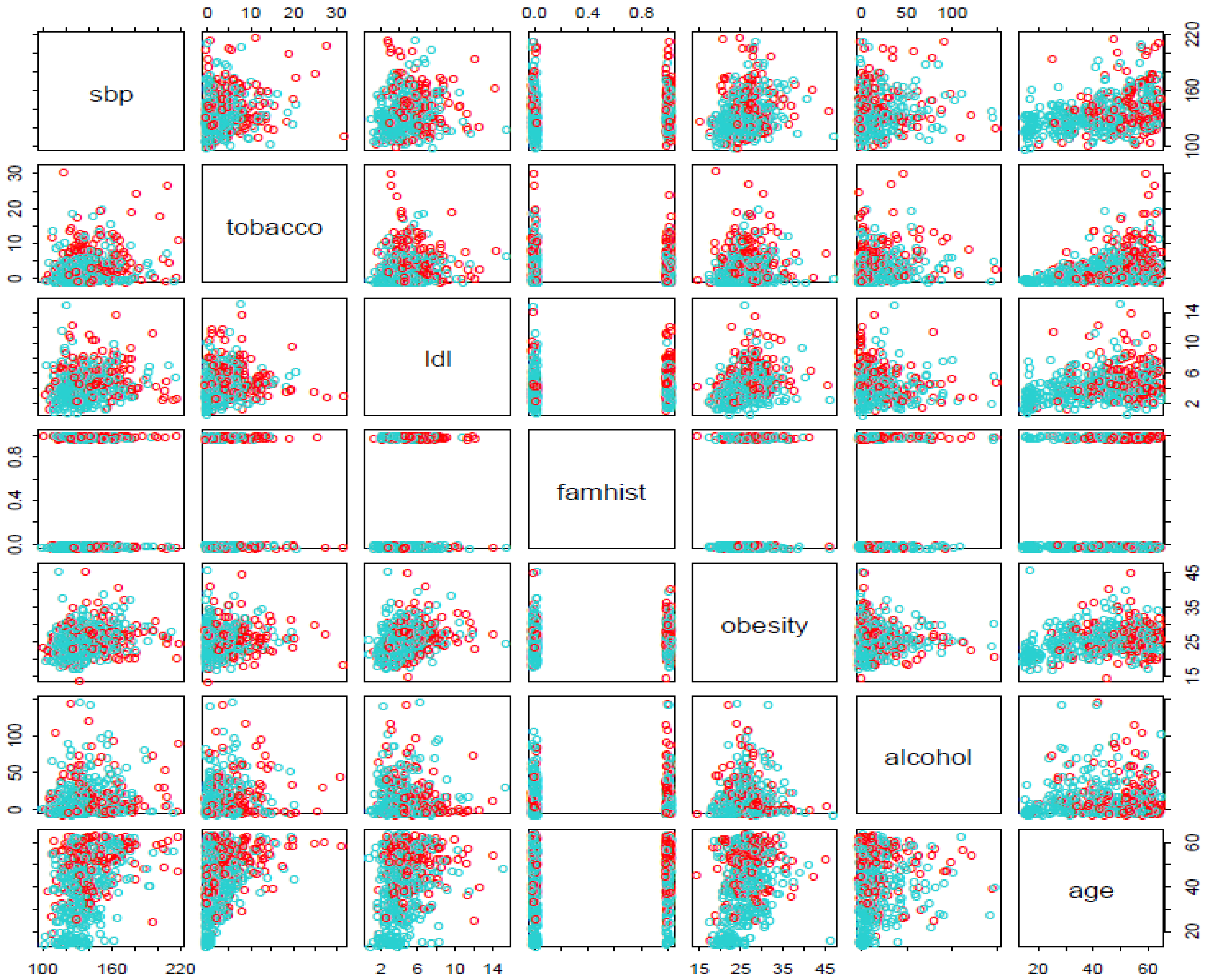
- A total of 462 samples are included in this data set
- Adiposity is a measure of % bodyfat, whereas obesity measures weight-to-height rations (body-mass-index, bmi). Type-A behaviour pattern is characterised by an excessive competitive drive, impatience and anger/hostility.

- systolic blood pressure (**sbp**)
- cumulative tobacco (**tobacco**)
- low density lipoprotein cholesterol (**ldl**)
- **Adiposity**
- family history of heart disease (**famhist**)
- type-A behavior (**typea**)
- **Obesity**
- **alcohol**
- **Age**

02. the South African heart disease data



02. the South African heart disease data



02. the South African heart disease data

1) Logistic Regression

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept) -6.1507209  1.3082600  -4.701 2.58e-06 ***
sbp          0.0065040  0.0057304   1.135 0.256374    
tobacco      0.0793764  0.0266028   2.984 0.002847 ** 
ldl          0.1739239  0.0596617   2.915 0.003555 ** 
adiposity    0.0185866  0.0292894   0.635 0.525700    
famhist      0.9253704  0.2278940   4.061 4.90e-05 ***
typea        0.0395950  0.0123202   3.214 0.001310 ** 
obesity     -0.0629099  0.0442477  -1.422 0.155095    
alcohol      0.0001217  0.0044832   0.027 0.978350    
age          0.0452253  0.0121298   3.728 0.000193 ***
```

73.37%

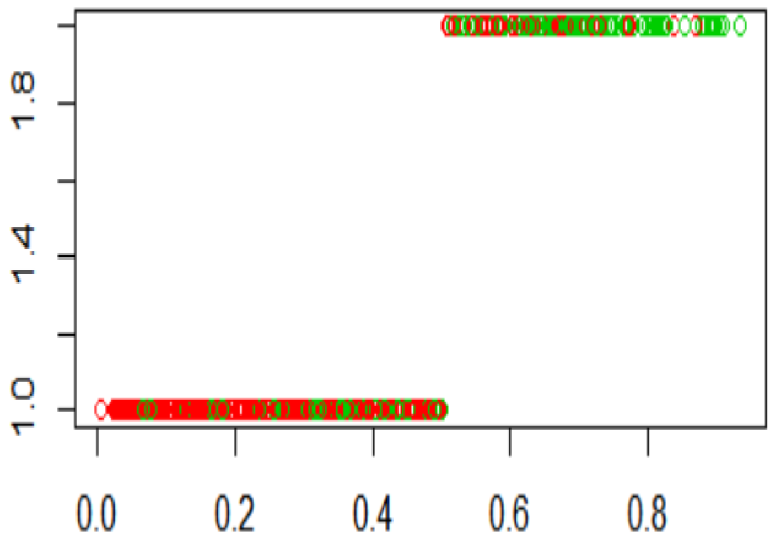
2) Logistic Regression -> stepwise

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -6.44644    0.92087  -7.000 2.55e-12 ***
tobacco      0.08038    0.02588   3.106  0.00190 ** 
ldl          0.16199    0.05497   2.947  0.00321 ** 
famhist      0.90818    0.22576   4.023 5.75e-05 ***
typea        0.03712    0.01217   3.051  0.00228 ** 
age          0.05046    0.01021   4.944 7.65e-07 ***
```

74.24%

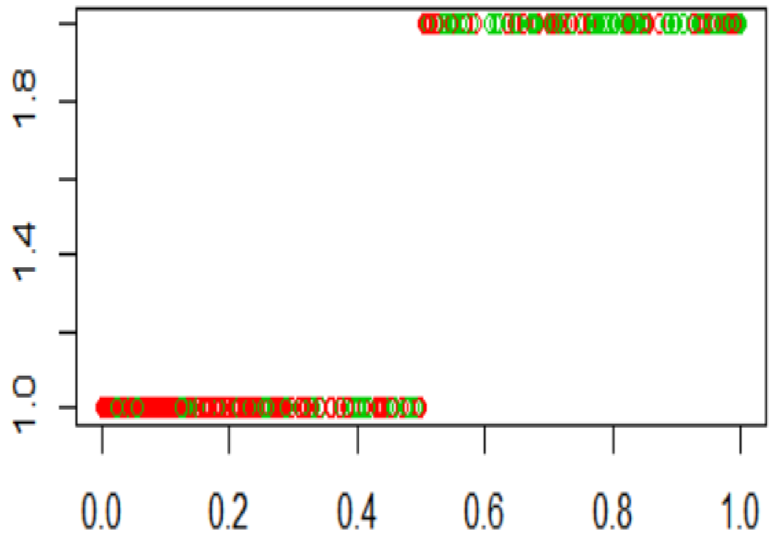
02. the South African heart disease data

3) LDA



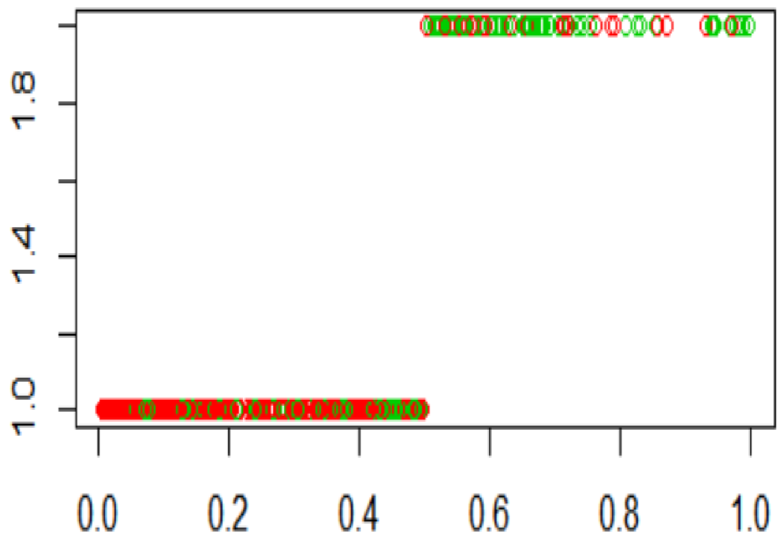
72.94%

4) QDA



75.75%

5) RDA



75.54%

3. Logistic Regression or LDA

03. Logistic Regression or LDA?

- the log-posterior odds between class k and K are linear functions of x

$$\begin{aligned}\log \frac{\Pr(G=k|X=x)}{\Pr(G=K|X=x)} &= \log \frac{\pi_k}{\pi_K} - \frac{1}{2}(\mu_k+\mu_K)^T\mathbf{\Sigma}^{-1}(\mu_k-\mu_K) \\ &\quad + x^T\mathbf{\Sigma}^{-1}(\mu_k-\mu_K) \\ &= \alpha_{k0} + \alpha_k^T x.\end{aligned}$$

- the linear logistic model by construction has linear logits

$$\log \frac{\Pr(G=k|X=x)}{\Pr(G=K|X=x)} = \beta_{k0} + \beta_k^T x$$

- seems that the models are the same
  - > the difference lies in the way the linear coefficients are estimated.

- The logistic regression model is more general, in that it makes less assumptions.

03. Logistic Regression or LDA?

- We can write the joint density of X and G as, (Pr(X) : the marginal density of the inputs X)

$$\Pr(X, G = k) = \Pr(X)\Pr(G = k|X)$$

- The logistic regression model leaves the marginal density of X as an arbitrary density function Pr(X), and fits the parameters of Pr(G|X) by maximizing the conditional likelihood

- with LDA we fit the parameters by maximizing the full log-likelihood, based on the joint density, where ϕ is the Gaussian density function.

$$\Pr(X, G = k) = \phi(X; \mu_k, \Sigma)\pi_k$$

$$\Pr(X) = \sum_{k=1}^{K} \pi_k \phi(X; \mu_k, \Sigma).$$

unlike in the conditional case, the marginal density Pr(X) does play a role here

4. Separating Hyperplane
5. Perceptron Learning Algorithm
6. Optimal Separating Hyperplane

04. Separating Hyperplane

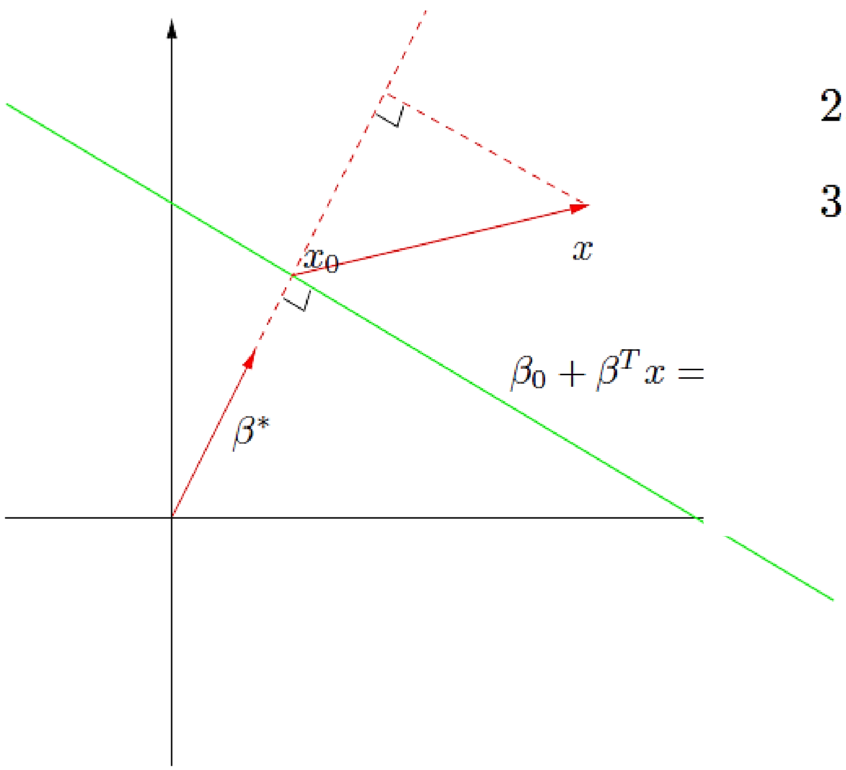


FIGURE 4.15. *The linear algebra of a hyperplane (affine set).*

1. For any two points x_1 and x_2 lying in L, $\beta^T(x_1 - x_2) = 0$, and hence $\beta^* = \beta/||\beta||$ is the vector normal to the surface of L.
2. For any point x_0 in L, $\beta^T x_0 = -\beta_0$.
3. The signed distance of any point x to L is given by

$$\begin{aligned} \beta^{*T}(x - x_0) &= \frac{1}{||\beta||}(\beta^T x + \beta_0) \\ &= \frac{1}{||f'(x)||} f(x). \end{aligned} \tag{4.40}$$

f(x) is proportional to the signed distance

from x to the hyperplane defined by f(x) = 0.

05. Rosenblatt’s Perceptron Learning Algorithm

- compute a linear combination of the input features and return the sign
- obtained by regressing the −1/1 response Y on X
- to find a separating hyperplane by minimizing the distance of misclassified points

  to the decision boundary. M indexes the set of misclassified points

$$D(\beta, \beta_0) = -\sum_{i \in \mathcal{M}} y_i (x_i^T \beta + \beta_0)$$

- If x in K=1, $x_i^T \beta + \beta_0 > 0$ / If x in K=2, $x_i^T \beta + \beta_0 < 0$

05. Rosenblatt's Perceptron Learning Algorithm

- The gradient is given by

$$\partial\frac{D(\beta,\beta_0)}{\partial\beta} = -\sum_{i\in\mathcal{M}} y_i x_i,$$

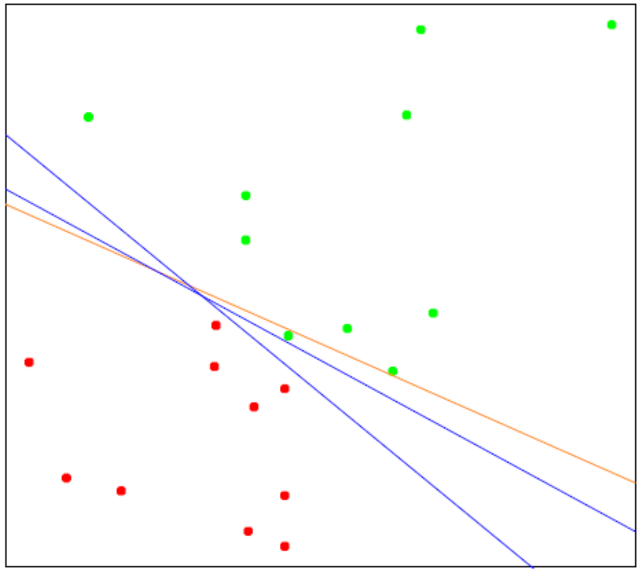
$$\partial\frac{D(\beta,\beta_0)}{\partial\beta_0} = -\sum_{i\in\mathcal{M}} y_i.$$

- The algorithm in fact uses *stochastic gradient descent* to minimize this piecewise linear criterion.
- the misclassified observations are visited in some sequence, and the β are updated via

$$\begin{pmatrix}\beta\\ \beta_0\end{pmatrix} \leftarrow \begin{pmatrix}\beta\\ \beta_0\end{pmatrix} + \rho\begin{pmatrix}y_i x_i\\ y_i\end{pmatrix}$$

, ρ is the learning rate, which in this case can be taken to be 1 without loss in generality

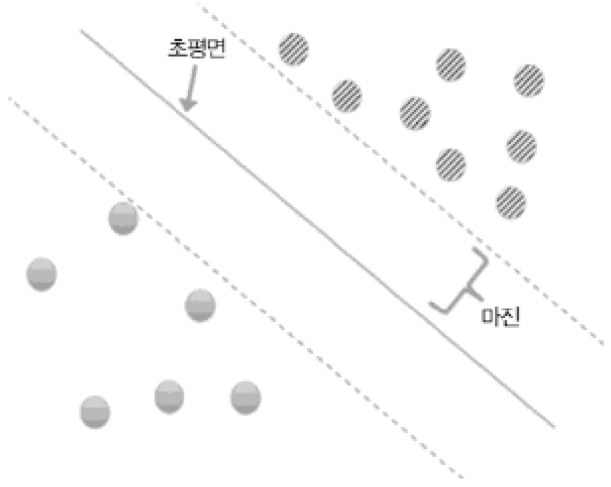
05. Rosenblatt's Perceptron Learning Algorithm



Problems with this algorithm

- When the data are separable, there are many solutions, and which one is found depends on the starting values.
- The "finite" number of steps can be very large. The smaller the gap, the longer the time to find it.
- When the data are not separable, the algorithm will not converge, and cycles develop. The cycles can be long and therefore hard to detect.

06. Optimal Separating Hyperplane



- the optimization problem

$$\max_{\beta, \beta_0, ||\beta||=1} M$$

$$\text{subject to } y_i(x_i^T\beta + \beta_0) \geq M, \; i = 1, \ldots, N.$$

- get rid of the ||β|| =1 constraint by replacing the conditions with

$$\frac{1}{||\beta||} y_i(x_i^T\beta + \beta_0) \geq M, \quad \Longrightarrow \quad y_i(x_i^T\beta + \beta_0) \geq M||\beta||.$$

- we can arbitrarily set ||β|| = 1/M

$$\min_{\beta, \beta_0} \frac{1}{2}||\beta||^2$$

$$\text{subject to } y_i(x_i^T\beta + \beta_0) \geq 1, \; i = 1, \ldots, N.$$

06. Optimal Separating Hyperplane

- the Lagrange (primal) function, to be minimized w.r.t. β and β0

$$L_P = \frac{1}{2}||\beta||^2 - \sum_{i=1}^{N} \alpha_i[y_i(x_i^T\beta + \beta_0) - 1] \qquad (4.49)$$

- setting the derivatives to zero,

$$\beta = \sum_{i=1}^{N} \alpha_i y_i x_i, \qquad 0 = \sum_{i=1}^{N} \alpha_i y_i,$$

- substituting these in (4.49) obtain the so-called Wolfe dual, the solution is obtained by maximizing L_D

$$L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{N} \alpha_i \alpha_k y_i y_k x_i^T x_k$$

$$\text{subject to} \quad \alpha_i \geq 0.$$

Thank you

ponybuhagom.tistory.com