Chapter 4: Linear Methods for Classification

Junrui Di

Contents

1. Linear discriminant analysis

Question set up for classification:

- Goal: To know the class posteriors $Pr(G|X)$ for optimal classification
- Parameters: $f_k(x)$ is the class conditional density of X in class $G = k$, and π_k is the probability of class k, with $\sum_{k=1}^K \pi_k = 1$.
- Bayes theorem gives that $Pr(G = k|X = x) = \frac{f_k(x)\pi_k}{\sum_{l=1}^K f_l(x)\pi_l}$

Suppose each class density is a multivariate Gaussian

$$f_k(x) = \frac{1}{(2\pi)^p|\mathbf{\Sigma}|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T\mathbf{\Sigma}^{-1}(x-\mu_k)} \quad \text{assume equal variance across classes}$$

To compare the two classes k and l, we have

$$\log\frac{Pr(G = k|X = x)}{Pr(G = l|X = x)} = \log\frac{\pi_k}{\pi_l} - \frac{1}{2}(\mu_k + \mu_l)^T\mathbf{\Sigma}^{-1}(\mu_k - \mu_l) + x^T\mathbf{\Sigma}^-1(\mu_k - \mu_l)$$

which is linear in x.

- Linear discriminant function is to solve for $G(x) = \text{argmax}_k\delta_k(x)$

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

π_k, μ_k, $\mathbf{\Sigma}$ all needs to be estimated empirically.

- Quadratic discriminant function is where $\mathbf{\Sigma}_k$ are not all the same, then the function becomes

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

1.1 Regularized disciminant analysis

Allow one to shrink the separate covariance of QDA toward a common covariance as in LDA

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

2. Logistic regression

For multiple K class,

$$Pr(G = k|X = x) = \frac{\exp(\beta_{k0} + \beta_k^T x)}{1 + \sum_{l=1}^{K-1} \exp(\beta_{l0} + \beta_l^T x)} \quad k = 1, ..., K-1$$
$$Pr(G = l|X = x) = \frac{1}{1 + \sum_{l=1}^{K-1} \exp(\beta_{l0} + \beta_l^T x)}$$

2.1 Fitting a binary class logistic regression

- Log likelihood for multiclass

$$l(\theta) = \sum_{i=1}^{N} \log p_{g_i}(x_i; \theta)$$

where $p_k(x_i; \theta) = Pr(G = k|X = x_i; \theta)$.

- For a two class case, $p_1(x; \theta) = p(x; \theta)$ corresponding to $y_i = 1$, and $p_2(x; \theta) = 1 - p(x; \theta)$ corresponding to $y_i = 0$. Then the log likelihood becomes

$$\begin{aligned} l(\beta) &= \sum_{i=1}^{N} \{y_i \log p(x_i; \theta) + (1 - y_i \log(1 - p(x_i; \theta)))\} \\ &= \sum_i \{y_i \beta^T x_i - \log(1 + e^{\beta^T x_i})\} \end{aligned}$$

- First order derivative

$$\begin{aligned} \frac{\partial l(\beta)}{\partial \beta} &= \sum_i x_i (y_i - p(x_i; \beta)) = 0 \\ &= \mathbf{X}^T(\mathbf{y} - \mathbf{p}) \end{aligned}$$

- Second order derivative

$$\begin{aligned} \frac{\partial^2 l(\beta)}{\partial \beta \partial \beta^T} &= -\sum_i x_i x_i^T p(x_i; \beta)(1 - p(x_i; p)) \\ &= -\mathbf{X}^T \mathbf{W} \mathbf{X} \end{aligned}$$

- Newton Raphson

$$\begin{aligned}\beta^{\text{new}} &= \beta^{\text{old}} - (\frac{\partial^2 l(\beta)}{\partial\beta\partial\beta^T})^{-1}\frac{\partial l(\beta)}{\partial\beta} \\ &= \beta^{\text{old}} + (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{y}-\mathbf{p}) \\ &= (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}(\mathbf{X}\beta_{old} + \mathbf{W}^{-1}(\mathbf{y}-\mathbf{p})) \\ &= (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{z}\end{aligned}$$

The last step can be considered as iteratively reweighted least squares

$$\beta^{\text{new}} \to \text{argmin}_\beta(\mathbf{z}-\mathbf{X}\beta)^T\mathbf{W}(\mathbf{z}-\mathbf{X}\beta)$$

2.1 L1 regularized logistic regression

$$\text{max}_{\beta_0,\beta_1}\{\sum_{i=1}^N[y_i(\beta_0+\beta^Tx_i) - \log(1+e^{\beta_0+\beta^Tx_i})] - \lambda\sum_{j=1}^p|\beta_j|\}$$

3. Separating hyperplanes

Hyperplane (affine set) L defined by the equation $f(x) = \beta_0 + \beta^T x = 0$, in $\mathbb{R}^2$, is a line, with the properties

- For any two points in L, $\beta^T(x_1 - x_2) = 0$
- For any point x_0 in L, $\beta^T x_0 = -\beta_0$
- The signed distance of any point x to L is $\frac{1}{||\beta||}(\beta^T x + \beta_0) = \frac{1}{||f'(x)||}f(x)$

3.1 Perpcetron learning algorithm

For a two class problem $y_i \in \{-1, 1\}$

$$\begin{aligned}x_i^T\beta + \beta_0 < 0 \quad &\text{if } y_i = 1 \text{ is misclassied} \\ x_i^T\beta + \beta_0 > 0 \quad &\text{if } y_i = -1 \text{ is misclassied}\end{aligned}$$

Therefore, the goal is to minimize

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

where $\mathcal{M}$ is the set of misclassified points. This quantity is nonnegative and proportional to the distance of the misclassfied points to the decision boundary $\beta^T x + \beta_0 = 0$. The gradient is

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

3.2 Optimal separating hyperplanes

Definition: OSH separates the two classes and maximizes the distance to the closest point from either class.

$$\max_{\beta,\beta_0,||\beta||=1} M$$
$$\text{subject to } y_i(x_i^T\beta+\beta_0) \geq M \quad \forall i$$

Interpretation: all points are at least a signed distance M from the decision boundary defined by β and β_0, and seek the largest the M. $||\beta|| = 1$ can be removed by changing the condition to $y_i(x_i^T\beta + \beta_0) \geq M||\beta||$

If we arbitrarily set $||\beta|| = 1/M$, the question becomes

$$\min_{\beta,\beta_0} \frac{1}{2}||\beta||^2$$
$$\text{subject to } y_i(x_i^T\beta+\beta_0) \geq 1 \quad \forall i$$

The constraints define a margin around the linear decision boundary of thickness $1/||\beta||$.

The question is to mimimize the Lagrange function

$$L_p = \frac{1}{2}||\beta||^2 - \sum_i \alpha_i[y_i(x_i^T\beta + \beta_0) - 1]$$