Chapter 5: Linear Models for Classification

Richard Liu

School of Mathematics, XMU

June 24, 2020

Content

Source

- Trevor Hastie, et al. *The Elements of Statistical Learning*
- Jie Hu, *Applied Linear Models Course in XMU*
- Zhihu: Gaussian Mixture Methods
- Zhihua Zhang, *Deep Learning Basics Course in PKU*

Section 1

Introduction

Introduction

- In chapter 3-4 we have discussed linear models for regression. However, there exist some other linear models used for solving classification problems.
- Examples: LDA, QDA, Logistic Regression (main topics in this chapter)
- In fact, these are parametric methods, and also there exist some other non-parametric methods.
    - Examples: Decision Trees (Introduced in Chapter 2), SVM, Ensemble Methods, Xgboost (Introduced in the further chapters).

Section 2

Multivariate Methods

Introduction

- We could use the methods introduced before.
- Suppose we have K classes with K indicators Y_k, where $Y_k = 1$ if $G = k$ and otherwise 0, then there will be an indicator response matrix $\mathbf{Y}$. By multivariate regression we have

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

- For a new independent variable x (which is a vector), we could have $\hat{f}(x)^T = (1, x^T)\hat{\mathbf{B}}$, where $\hat{\mathbf{B}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$

Introduction

Criterion

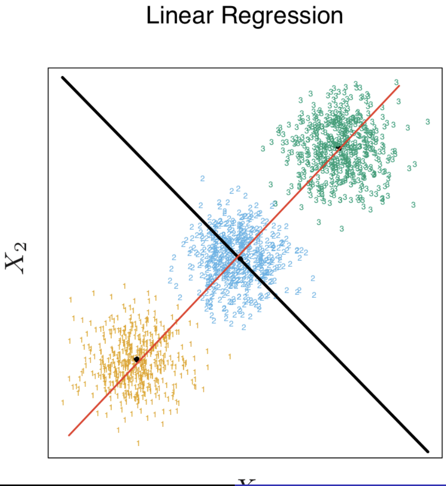
$\hat{G}(x) = \arg\max_k \hat{f}_k(x)$, where k denotes the indicators of classes.

Explanation

Note that $E(Y_k|X = x) = P(G = k|X = x)$

Problem: Masking!

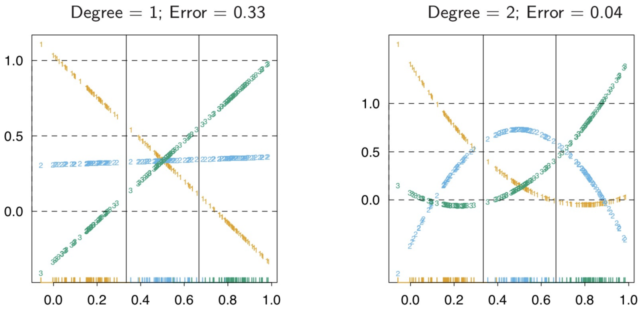
Masking



What happened?

- We projected the data onto the plane joining the three centroids. Then we could run three regression lines and draw them on the same graph.
- For comparison we draw the graphs for quadratic regression, too.

Graphs



We should attribute this to the natural rigidity of multivariate regression.

Section 3

Linear Discriminant Analysis

Introduction

- Could we solve classification by some alternative linear models?
    - Linear Discriminant Analysis (LDA)
- Discriminant Analysis: Given the number of groups, identify where the observations locate by specified characteristic values.
- Relies on Bayesian Statistics.

Introduction

- Prior: Suppose we have π_k be the prior probability of class k, which means $P(G = k) = \pi_k$.
- Posterior:

$$P(G = k|X = x) = \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\pi_l}$$

Where $f_k(x)$ is the pdf of data in the k-th class.

Theorem 1

Prove the posterior probability.

Introduction

Suppose we have the probability

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

And suppose that $\Sigma_k = \Sigma, \forall k$, then

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

Question: Why log-ratio?

Why call LDA?

- Note that the log-ratio is linear w.r.t x.
- In fact, we have the following linear discriminant functions

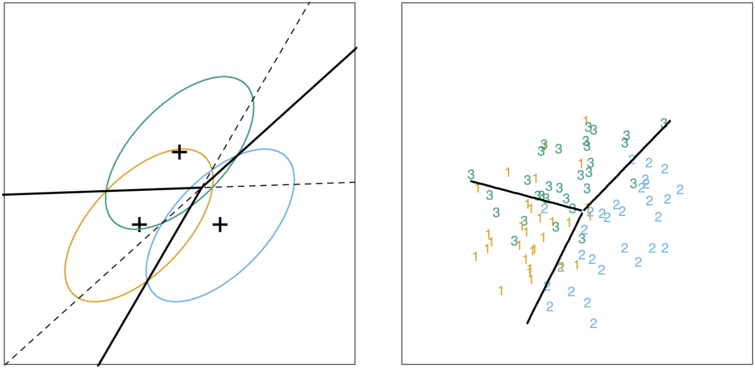
$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k$$

(Why defined as this?)

Criterion

$G(x) = \arg\max_k \delta_k(x)$

Decision Boundary——Graphs



Decision Boundary——Explanation

- The independent variables on decision boundaries have the same discriminant function values.
- Consider a simple 1-d case with 2 classes and the same prior probability, then we have

$$\frac{x\mu_1}{\sigma} - \frac{1}{2}\frac{\mu_1^2}{\sigma} = \frac{x\mu_2}{\sigma} - \frac{1}{2}\frac{\mu_2^2}{\sigma}$$

- Solve this equation, we obtain $x = \frac{\mu_1+\mu_2}{2}$.
- You could draw two graphs to find some interesting phenomenons.

How to apply this model into the real data?

- Problem: In real sample data, all parameters are unknown.
    - So we need some estimators.

Estimators

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

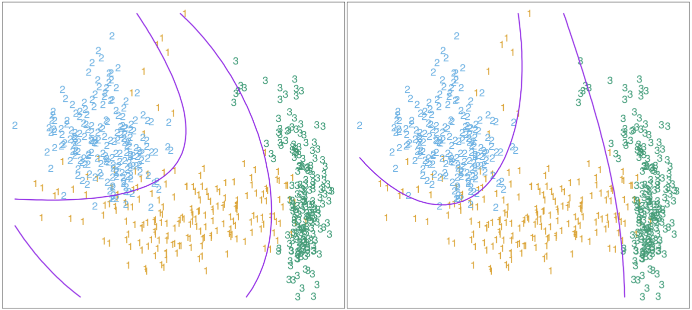
Quadratic Discriminant Analysis

- If we drop the assumptions $\Sigma_k = \Sigma$, then the discriminant functions become

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

- We call it Quadratic Discriminant Analysis (QDA) because it is a quadratic function w.r.t. x.

Decision Boundaries (Difference?)



Bias-Variance Trade-off

- Background: Many models based on LDA apply well on real data (e.g. Naive Bayes).
- Why?
  - Bias-Variance Trade-off
- Generally, higher bias, lower variance.
- Linear or quadratic models are simple and have few parameters, this may lead to higher bias, which means they have a good generality (lower variance).

Section 4

Logistic Regression

Introduction

- Problem: For predicting binary values, could we use traditional multivariate regression models?
    - No!
    - Rigidity, unboundedness, etc.
- Do an exponential transformation and normalize.
- A simple case (with only two classes): $P(G=1|X=x)=\frac{\exp(\beta_0+\beta^T x)}{1+\exp(\beta_0+\beta^T x)}$, $P(G=2|X=x)=\frac{1}{1+\exp(\beta_0+\beta^T x)}$. It is widely used in Biostatistics.

General Form

Problem Formulation (Part 1)

$$
\begin{aligned}
\log \frac{\Pr(G=1|X=x)}{\Pr(G=K|X=x)} &= \beta_{10} + \beta_1^T x \\
\log \frac{\Pr(G=2|X=x)}{\Pr(G=K|X=x)} &= \beta_{20} + \beta_2^T x \\
&\vdots \\
\log \frac{\Pr(G=K-1|X=x)}{\Pr(G=K|X=x)} &= \beta_{(K-1)0} + \beta_{K-1}^T x
\end{aligned} \tag{1}
$$

For the same reason, you know that we could use log-ratio here.

Problem Formulation (Part 2)

$$
\begin{aligned}
\Pr(G=k|X=x) &= \frac{\exp\left(\beta_{k0}+\beta_k^T x\right)}{1+\sum_{\ell=1}^{K-1}\exp\left(\beta_{\ell 0}+\beta_\ell^T x\right)}, k=1,\ldots,K-1 \\
\Pr(G=K|X=x) &= \frac{1}{1+\sum_{\ell=1}^{K-1}\exp\left(\beta_{\ell 0}+\beta_\ell^T x\right)}
\end{aligned} \tag{2}
$$

- How to find the estimators of β?
  - Maximum Likelihood Estimator (MLE)!
- We will introduce the simple case with 2 classes and defer the general discussion later.

Maximum Likelihood Estimator

- Note that the log-likelihood for N observations is

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

  where $p_k(x_i; \theta) = P(G = k | X = x_i; \theta)$
- Consider the two-class case, where $y_i = 1$ when $g_i = 1$ and $y_i = 0$ when $g_i = 2$. Then we could write the likelihood as

$$l(\beta) = \sum_{i=1}^{N} [y_i \beta^T x_i - \log(1 + e^{\beta^T x_i})]$$

- Why?

Bad News

- The analytical solution of the objective does not exist, which means we need to rely on Numerical Optimization to solve the problem.

Section 5

LDA or Logistic Regression?

Comparison

- The formula for LDA is

$$\log \frac{\Pr(G=k|X=x)}{\Pr(G=K|X=x)} = \log \frac{\pi_k}{\pi_K} - \frac{1}{2}\left(\mu_k+\mu_K\right)^T \Sigma^{-1}\left(\mu_k-\mu_K\right)$$
$$+ x^T\Sigma^{-1}\left(\mu_k-\mu_K\right)$$
$$= \alpha_{k0} + \alpha_k^T x$$

- The formula for Logistic Regression is

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

- What is the difference?

Comparison

Note that $P(X, G = k) = P(X)P(G = k|X)$. So the key difference is the assumptions put on the prior probability $P(X)$. For these two models have the same conditional probability form.

$$\Pr(G = k|X = x) = \frac{e^{\beta_{k0} + \beta_k^T x}}{1 + \sum_{\ell=1}^{K-1} e^{\beta_{\ell 0} + \beta_\ell^T x}}$$

Comparison

That is to say, in LDA, we have

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

while we do not add on much information on the prior probability in Logistic Regression.

- For LDA: more accurate (lower bias) but less robust (higher variance).
- Why?

Subsection 1

Another View: Generative Model v.s. Discriminative Model

Generative Model: Assumptions

- Assume that $x = (x_1, \dots, x_p)^T \in \mathbb{R}^p, y = \{0, 1\}$, then we assume the joint-distribution of X, Y is known and parameterized by θ, which means

$$p(x, y|\theta) = p(y|\theta_1)p(x|y, \theta_2)$$

- Example: Bernoulli Prior and Gaussian Posterior. This implies $p(y|\pi) = \pi^y(1-\pi)^{1-y}, \pi \in (0, 1)$ and $p(x_j|Y = k, \theta_j) \sim N(\mu_{kj}, \sigma_j^2), k = 0, 1$

Generative Model: Example

- To provide a posterior distribution of $p(y|x,\theta)$ with given data (x here means all the data, not just an observation). we need to find the joint distribution first, and then use Bayesian Formula.
- Here, we introduce Naive Bayes, having assumption

$$p(x,y|\theta) = p(y|\pi)p(x|y,\hat{\theta}) = p(y|\pi)\prod_{j=1}^{p} p(x_j|y,\theta_j)$$

Generative Model: Example

- With previous assumptions we have

$$p(x|y=k,\theta)=\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\exp\{-\frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k)\}$$

with $\mu_0=(\mu_{01},\ldots,\mu_{0p})^T, \mu_1=(\mu_{11},\ldots,\mu_{1p})^T, \Sigma=\text{diag}(\sigma_1^2,\ldots,\sigma_p^2)$.

- This formula is the same as that in the previous sections.

Generative Model: Example

- With Bayesian Formula, we could find

$$P(Y=1|x,\theta)=\frac{P(x|Y=1,\theta)P(Y=1|\pi)}{P(x|Y=1,\theta)P(Y=1|\pi)+P(x|Y=0,\theta)P(Y=}$$

$$=\frac{1}{1+\frac{1-\pi}{\pi}\exp\{-(\mu_1-\mu_0)^T\Sigma^{-1}x+\frac{1}{2}(\mu_1-\mu_0)^T\Sigma^{-1}(\mu_1+\mu_0)\}}$$

$$=\frac{1}{1+\exp\{-\beta^T x-r\}}$$

where $\beta=\Sigma^{-1}(\mu_1-\mu_0)$,
$r=-\frac{1}{2}(\mu_1-\mu_0)^T\Sigma^{-1}(\mu_1+\mu_0)+\ln\frac{\pi}{1-\pi}$

Generative Model: Parameters Estimation

- We need to use MLE with data $D = \{(x_n, y_n), n = 1, \ldots, N\}$, which means

$$l(\theta|D) = \log\{\prod_{n=1}^{N} p(y_n|\pi) \prod_{j=1}^{p} p(x_{nj}|y_n, \theta_j)\}$$

$$= \sum_{n=1}^{N} \log p(y_n|\pi) + \sum_{n=1}^{N} \sum_{j=1}^{p} \log p(x_{n,j}|y_n, \theta_j)$$

Discriminative Model: Assumptions

- Assume that $p(y=1|x,\beta)=\frac{1}{1+\exp\{-\beta^T x\}}$. This means we do not care much about the distribution of X.
- Example: $p(y|x,\beta)=(\mu(x))^y(1-\mu(x))^{1-y}$, where $\mu(x)\in(0,1)$.
- MLE: $l(\beta)=\sum_{i=1}^n [y_i \log\mu_i+(1-y_i)\log(1-\mu_i)]$ (similar to Logistic Regression).
- No analytic results, we need numerical optimization (e.g. Gradient Descent).

Thank you!