Modern Statistical Methods

Rajen Shah

Chapters

1. Kernel machines
2. The Lasso and extensions
3. Graphical modelling and causal inference
4. Multiple testing and high-dimensional inference

Books

"Elements of statistical learning"
"Statistics for high-dimensional data"
"Statistical learning with sparsity"

Webpage

- Notes
- Feedback form

Introduction

Ordinary Least Squares (OLS)

Data $(Y_i, x_i) \in \mathbb{R} \times \mathbb{R}^p$, $i = 1, \dots, n$

(Y_i: response, x_i: predictors)

Linear model assumes $Y = X\beta^0 + \varepsilon$ where $X = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix}$ $\quad E\varepsilon = 0$ $\quad \mathrm{Var}(\varepsilon) = \sigma^2 I$

Can estimate β^0 by OLS provided X has full col rank (so X^TX invertible).

$$\hat{\beta}^{OLS} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \|Y - X\beta\|_2^2 = (X^T X)^{-1} X^T Y$$

- $E\hat{\beta}^{OLS} = E(X^TX)^{-1}X^T(X\beta^0 + \varepsilon) = \beta^0$ unbiased
- $Var(\hat{\beta}^{OLS}) = \sigma^2 (X^TX)^{-1} X^T X (X^TX)^{-1} = \sigma^2 (X^TX)^{-1}$

Gauss-Markov theorem: $\hat{\beta}^{OLS}$ is the best linear unbiased estimator of β^0

Any other unbiased linear estimator $\tilde{\beta} = AY$

$Var(\tilde{\beta}) - \sigma^2 (X^TX)^{-1}$ positive semi-definite

Maximum Likelihood

Specify a density for Y, $f(y;\theta)$, θ is an unknown vector of parameters in $\mathbb{R}^d$

log-likelihood $\ell(\theta) = \log(f(Y;\theta))$

Maximum likelihood estimation maximize $\ell(\theta)$

- Call maximizer $\hat{\theta}$.

Fisher information matrix $i_{jk}(\theta) = -E_\theta\left(\frac{\partial^2}{\partial\theta_j \partial\theta_k}\ell(\theta)\right)$

Cramér-Rao lower bound if $\tilde{\theta}$ is an unbiased estimator, then

$Var(\tilde{\theta}) - i^{-1}(\theta)$ is positive semi-definite

MLEs - asymptotically ($n \to \infty$) unbiased, achieves the C-R bound, normally distributed

Chapter 1 Kernel methods

Consider linear model $Y = X\beta^0 + \varepsilon$ $E\varepsilon = 0$ $Var(\varepsilon) = \sigma^2 I$

For a biased estimator $\tilde{\beta}$ we should study the mean-squared error matrix

$$E(\tilde{\beta} - \beta^0)(\tilde{\beta} - \beta^0)^T = E\{(\tilde{\beta} - E\tilde{\beta} + E\tilde{\beta} - \beta^0)(\tilde{\beta} - E\tilde{\beta} + E\tilde{\beta} + \beta^0)^T\}$$

$$= Var(\tilde{\beta}) + \underbrace{E(\tilde{\beta} - \beta^0)\{E(\tilde{\beta} - \beta^0)\}^T}_{\text{squared bias}}$$

1.1. Ridge regression

Idea: Reduce variance of OLS by shrinking estimated coefficients towards 0

Ridge regression solves

$$(\hat{\mu}_\lambda^R, \hat{\beta}_\lambda^R) = \underset{(\mu,\beta)\in\mathbb{R}\times\mathbb{R}^p}{\arg\min} \left\{ \|Y - \mu \mathbf{1} - X\beta\|_2^2 + \lambda \|\beta\|_2^2 \right\}$$

where $\mathbf{1}$ is an n-vector of 1's

$\lambda \geq 0$ is a tuning parameter

Intercept term is not penalized, so that if response is $Y' = Y + c\mathbf{1}$ then

$$\hat{\mu}_\lambda^R(Y') = \hat{\mu}_\lambda^R(Y) + c$$

It is also standard practice to center each col of X and scale them to have equal l_2-norm. Then have $\hat{\mu}_\lambda^R = \bar{Y} = \frac{1}{n}\sum Y_i$ so by replacing Y with $Y - \bar{Y}\mathbf{1}$ can omit intercept from the objective.

Have $\hat{\beta}_\lambda^R = (X^T X + \lambda I)^{-1} X^T Y$

Theorem 1 Suppose $\text{rank}(X) = p$. For λ sufficiently small (depending on β^0 and σ^2)

$$E(\hat{\beta}^{OLS} - \beta^0)(\hat{\beta}^{OLS} - \beta^0)^T - E(\hat{\beta}_\lambda^R - \beta^0)(\hat{\beta}_\lambda^R - \beta^0)^T \text{ is positive definite.}$$

- EoSL p 61 – 68
  241 – 249 cross-validation