Part III — Modern Statistical Methods

Theorems

Based on lectures by R. D. Shah

Notes taken by Dexter Chua

Michaelmas 2017

The remarkable development of computing power and other technology now allows scientists and businesses to routinely collect datasets of immense size and complexity. Most classical statistical methods were designed for situations with many observations and a few, carefully chosen variables. However, we now often gather data where we have huge numbers of variables, in an attempt to capture as much information as we can about anything which might conceivably have an influence on the phenomenon of interest. This dramatic increase in the number variables makes modern datasets strikingly different, as well-established traditional methods perform either very poorly, or often do not work at all.

Developing methods that are able to extract meaningful information from these large and challenging datasets has recently been an area of intense research in statistics, machine learning and computer science. In this course, we will study some of the methods that have been developed to analyse such datasets. We aim to cover some of the following topics.

- Kernel machines: the kernel trick, the representer theorem, support vector machines, the hashing trick.
- Penalised regression: Ridge regression, the Lasso and variants.
- Graphical modelling: neighbourhood selection and the graphical Lasso. Causal inference through structural equation modelling; the PC algorithm.
- High-dimensional inference: the closed testing procedure and the Benjamini–Hochberg procedure; the debiased Lasso

Pre-requisites

Basic knowledge of statistics, probability, linear algebra and real analysis. Some background in optimisation would be helpful but is not essential.

Contents

0 Introduction

1 Classical statistics

Theorem (Cramér–Rao bound)**.** If $\tilde{\theta}$ is an unbiased estimator for θ, then $\mathrm{var}(\tilde{\theta}) - I^{-1}(\theta)$ is positive semi-definite.

Moreover, asymptotically, as $n \to \infty$, the maximum likelihood estimator is unbiased and achieves the Carmér–Rao bound.

2 Kernel machines

2.1 Ridge regression

Theorem. Suppose $\operatorname{rank}(X) = p$. Then for $\lambda > 0$ sufficiently small (depending on β^0 and σ^2), the matrix

$$\mathbb{E}(\hat{\beta}^{OLS} - \beta^0)(\hat{\beta}^{OLS} - \beta^0)^T - \mathbb{E}(\hat{\beta}^R_\lambda - \beta^0)(\hat{\beta}^R_\lambda - \beta^0)^T \qquad (*)$$

is positive definite.

Theorem (Singular value decomposition)**.** Let $X \in \mathbb{R}^{n\times p}$ be any matrix. Then it has a *singular value decomposition* (SVD)

$$\underset{n\times p}{X} = \underset{n\times n}{U}\ \underset{n\times p}{D}\ \underset{p\times p}{V^T},$$

where U, V are orthogonal matrices, and $D_{11} \geq D_{22} \geq \cdots \geq D_{mm} \geq 0$, where $m = \min(n, p)$, and all other entries of D are zero.

2.2 v-fold cross-validation

2.3 The kernel trick

Proposition. Given $\phi : \mathcal{X} \times \mathcal{X} \to \mathcal{H}$, define $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ by

$$k(x, x') = \langle \phi(x), \phi(x') \rangle.$$

Then for any $x_1, \ldots, x_n \in \mathcal{X}$, the matrix $K \in \mathbb{R}^n \times \mathbb{R}^n$ with entries

$$K_{ij} = k(x_i, x_j)$$

is positive semi-definite.

Theorem (Moore–Aronszajn theorem)**.** For every kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, there exists an inner product space $\mathcal{H}$ and a feature map $\phi : \mathcal{X} \to \mathcal{H}$ such that

$$k(x, x') = \langle \phi(x), \phi(x') \rangle.$$

2.4 Making predictions

Theorem (Representer theorem)**.** Let $\mathcal{H}$ be an RKHS with reproducing kernel k. Let c be an arbitrary loss function and $J : [0, \infty) \to \mathbb{R}$ any strictly increasing function. Then the minimizer $\hat{f} \in \mathcal{H}$ of

$$Q_1(f) = c(Y, x_1, \ldots, x_n, f(x_1), \ldots, f(x_n)) + J(\|f\|^2_{\mathcal{H}})$$

lies in the linear span of $\{k(\,\cdot\,, x_i)\}_{i=1}^n$.

Theorem. We have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \mathbb{E}(f^0(x_i) - \hat{f}_\lambda(x_i))^2 &\leq \frac{\sigma^2}{n}\sum_{i=1}^n \frac{d_i^2}{(d_i + \lambda)^2} + \frac{\lambda}{4n} \\ &\leq \frac{\sigma^2}{n}\frac{1}{\lambda}\sum_{i=1}^n \min\left(\frac{d_i}{4}, \lambda\right) + \frac{\lambda}{4n}.\end{aligned}$$

2.5 Other kernel machines

2.6 Large-scale kernel machines

Theorem (Bochner's theorem)**.** Let $k : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ be a continuous kernel. Then k is shift-invariant if and only if there exists some distribution F on $\mathbb{R}^p$ and $c > 0$ such that if $W \sim F$, then

$$k(x, x') = c\mathbb{E}\cos((x - x')^T W).$$

3 The Lasso and beyond

3.1 The Lasso estimator

Theorem. Let $\hat{\beta}$ be the Lasso solution with

$$\lambda = A\sigma\sqrt{\frac{\log p}{n}}$$

for some A. Then with probability $1 - 2p^{-(A^2/2-1)}$, we have

$$\frac{1}{n}\|X\beta^0 - X\hat{\beta}\|_2^2 \leq 4A\sigma\sqrt{\frac{\log p}{n}}\|\beta^0\|_1.$$

3.2 Basic concentration inequalities

Lemma (Markov's inequality)**.** Let W be a non-negative random variable. Then

$$\mathbb{P}(W \geq t) \leq \frac{1}{t}\mathbb{E}W.$$

Corollary (Chernoff bound)**.** For any random variable W, we have

$$\mathbb{P}(W \geq t) \leq \inf_{\alpha>0} e^{-\alpha t}\mathbb{E}e^{\alpha W}.$$

Corollary. Any sub-Gaussian random variable W with parameter σ satisfies

$$\mathbb{P}(W \geq t) \leq e^{-t^2/2\sigma^2}. \qquad \square$$

Lemma (Hoeffding's lemma)**.** If W has mean zero and takes values in $[a, b]$, then W is sub-Gaussian with parameter $\frac{b-a}{2}$. $\square$

Proposition. Let $(W_i)_{i=1}^n$ be independent mean-zero sub-Gaussian random variables with parameters $(\sigma_i)_{i=0}^n$, and let $\gamma \in \mathbb{R}^n$. Then $\gamma^T W$ is sub-Gaussian with parameter

$$\left(\sum(\gamma_i\sigma_i)^2\right)^{1/2}.$$

Lemma. Suppose $(\varepsilon_i)_{i=1}^n$ are independent, mean-zero sub-Gaussian with common parameter σ. Let

$$\lambda = A\sigma\sqrt{\frac{\log p}{n}}.$$

Let X be a matrix whose columns all have norm $\sqrt{n}$. Then

$$\mathbb{P}\left(\frac{1}{n}\|X^T\varepsilon\|_\infty \leq \lambda\right) \geq 1 - 2p^{-(A^2/2-1)}.$$

Proposition (Bernstein's inequality)**.** Let $W_1, W_2, \ldots, W_n$ be independent random variables with $\mathbb{E}W_i = \mu$, and suppose each W_i satisfies Bernstein's condition with parameters (σ, b). Then

$$\mathbb{E}e^{\alpha(W_i-\mu)} \leq \exp\left(\frac{\alpha^2\sigma^2/2}{1-b|\alpha|}\right) \text{ for all } |\alpha| < \frac{1}{b},$$

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n W_i - \mu \geq t\right) \leq \exp\left(-\frac{nt^2}{2(\sigma^2+bt)}\right) \text{ for all } t \geq 0.$$

Lemma. Let W, Z be mean-zero sub-Gaussian random variables with parameters σ_W and σ_Z respectively. Then WZ satisfies Bernstein's condition with parameter $(8\sigma_W\sigma_Z, 4\sigma_W\sigma_Z)$.

3.3 Convex analysis and optimization theory

Proposition.

(i) Let $f_1, \dots, f_m : \mathbb{R}^d \to \bar{\mathbb{R}}$ be convex with $\operatorname{dom} f_1 \cap \cdots \cap \operatorname{dom} f_m \neq \emptyset$, and let $c_1, \dots, c_m \geq 0$. Then $c_1 + \cdots + c_m f_m$ is a convex function.

(ii) If $f : \mathbb{R}^d \to \mathbb{R}$ is twice continuously differentiable, then

(a) f is convex iff its Hessian is positive semi-definite everywhere.

(b) f is strictly convex if its Hessian positive definite everywhere. □

Proposition. Let f be convex and differentiable at $x \in \operatorname{int}(\operatorname{dom} f)$. Then $\partial f(x) = \{\nabla f(x)\}$. □

Proposition. Suppose f and g are convex with $\operatorname{int}(\operatorname{dom} f) \cap \operatorname{int}(\operatorname{dom} g) \neq \emptyset$, and $\alpha > 0$. Then

$$\partial(\alpha f)(x) = \alpha \partial f(x) = \{\alpha v : v \in \partial f(x)\}$$
$$\partial(f + g)(x) = \partial g(x) + \partial f(x).$$ □

Proposition. If f is convex, then

$$x^* \in \operatorname*{argmin}_{x \in \mathbb{R}^d} f(x) \Leftrightarrow 0 \in \partial f(x^*).$$

Proposition. For $x \in \mathbb{R}^d$ and $A \in \{j : x_j \neq 0\}$, we have

$$\partial \|x\|_1 = \{v \in \mathbb{R}^d : \|v\|_\infty \leq 1, v_A = \operatorname{sgn}(x_A)\}.$$

3.4 Properties of Lasso solutions

Proposition. $X\hat{\beta}^L_\lambda$ is unique.

3.5 Variable selection

Theorem.

(i) If $\|\Delta\|_\infty \leq 1$, or equivalently

$$\max_{k \in N} |\operatorname{sgn}(\beta^0_S)^T (X_S^T X_S)^{-1} X_S^T X_k| \leq 1,$$

and moreover

$$|\beta^0_k| > \lambda \left| \operatorname{sgn}(\beta^0_S)^T \left(\frac{1}{n} X_j^T X_j \right)^{-1}_k \right|$$

for all $k \in S$, then there exists a Lasso solution $\hat{\beta}^L_\lambda$ with $\operatorname{sgn}(\hat{\beta}^L_\lambda) = \operatorname{sgn}(\beta^0)$.

(ii) If there exists a Lasso solution with $\operatorname{sgn}(\hat{\beta}^L_\lambda) = \operatorname{sgn}(\beta^0)$, then $\|\Delta\|_\infty \leq 1$.

Theorem. Assume $\phi^2 > 0$, and let $\hat{\beta}$ be the Lasso solution with

$$\lambda = A\sigma\sqrt{\log p/n}.$$

Then with probability at least $1 - 2p^{-(A^2/8-1)}$, we have

$$\frac{1}{n}\|X(\beta^0 - \hat{\beta})\|_2^2 + \lambda\|\hat{\beta} - \beta^0\|_1 \leq \frac{16\lambda^2 s}{\phi^2} = \frac{16A^2\log p}{\phi^2}\frac{s\sigma^2}{n}.$$

Lemma. Let $\Theta, \Sigma \in \mathbb{R}^{p\times p}$. Suppose $\phi_\Theta^2(S) > 0$ and

$$\max_{j,k}|\Theta_{jk} - \Sigma_{jk}| \leq \frac{\phi_\Theta^2(S)}{32|S|}.$$

Then

$$\phi_\Sigma^2(S) \geq \frac{1}{2}\phi_\Theta^2(S).$$

Theorem. Suppose the rows of X are iid and each entry is sub-Gaussian with parameter v. Suppose $s\sqrt{\log p/n} \to 0$ as $n \to \infty$, and $\phi_{\Sigma^0,s}^2$ is bounded away from 0. Then if $\Sigma^0 = \mathbb{E}\hat{\Sigma}$, then we have

$$\mathbb{P}\left(\phi_{\hat{\Sigma},s}^2 \geq \frac{1}{2}\phi_{\Sigma^0,s}^2\right) \to 1 \text{ as } n \to \infty.$$

Corollary. Suppose the rows of X are iid mean-zero multivariate Gaussian with variance Σ^0. Suppose Σ^n has minimum eigenvalue bounded from below by $c_{min} > 0$, and suppose the diagonal entries of Σ^0 are bounded from above. If $s\sqrt{\log p/n} \to 0$, then

$$\mathbb{P}\left(\phi_{\hat{\Sigma},s}^2 \geq \frac{1}{2}c_{min}\right) \to 1 \text{ as } n \to \infty.$$

3.6 Computation of Lasso solutions

3.7 Extensions of the Lasso

4 Graphical modelling

4.1 Conditional independence graphs

Proposition. If P has a positive density, then if it satisfies the pairwise Markov property with respect to $\mathcal{G}$, then it also satisfies the global Markov property.

Proposition. Suppose $Z \sim N_p(\mu, \Sigma)$ and Σ is positive definite. Then

$$Z_A \mid Z_B = z_B \sim N_{|A|}(\mu_A + \Sigma_{A,B}\Sigma_{B,B}^{-1}(z_B - \mu_B), \Sigma_{A,A} - \Sigma_{A,B}\Sigma_{B,B}^{-1}\Sigma_{B,A}).$$

Lemma. Given k, let j' be such that $(Z_{-k})_j = Z_{j'}$. This j' is either j or $j+1$, depending on whether it comes after or before k.

If the jth component of $\Sigma_{-k,-k}^{-1}\Sigma_{-k,k}$ is 0, then $Z_k \amalg Z_{j'} \mid Z_{-kj'}$.

Lemma. Let $M \in \mathbb{R}^{p\times p}$ be positive definite, and write

$$M = \begin{pmatrix} P & Q \\ Q^T & R \end{pmatrix},$$

where P and Q are square. The *Schur complement* of R is

$$S = P - QR^{-1}Q^T.$$

Note that this has the same size as P. Then

(i) S is positive definite.

(ii)
$$M^{-1} = \begin{pmatrix} S^{-1} & -S^{-1}QR^{-1} \\ -R^{-1}Q^TS^{-1} & R^{-1} + R^{-1}Q^TS^{-1}QR^{-1} \end{pmatrix}.$$

(iii) $\det(M) = \det(S)\det(R)$

4.2 Structural equation modelling

Proposition. If P has a density with respect to a product measure, then (i) and (ii) are equivalent.

Proposition. Let P be the structural equation model with DAG $\mathcal{G}$. Then P obeys the Markov factorization property.

4.3 The PC algorithm

Proposition. Two DAGs are Markov equivalent iff they have the same skeleton and same set of v-structure.

Proposition. If nodes j and k are adjacent in a DAG $\mathcal{G}$, then no set can d-separate them.

If they are not adjacent, and π is a topological order for $\mathcal{G}$ with $\pi(j) < \pi(k)$, then they are d-separated by $\mathrm{pa}(k)$.

Proposition. Suppose we have $j - \ell - k$ in the skeleton of a DAG.

(i) If $j \to \ell \leftarrow k$, then no S that d-separates j can have $\ell \in S$.

(ii) If there exists S that d-separates j and k and $\ell \notin S$, then $j \to \ell \leftarrow k$. $\square$

5 High-dimensional inference

5.1 Multiple testing

Theorem. When using the Bonferroni correction, we have

$$\text{FWER} \leq \mathbb{E}(N) \leq \frac{m_0\alpha}{m} \leq \alpha.$$

Theorem. Closed testing makes no false rejections with probability $\geq 1-\alpha$. In particular, FWER $\leq \alpha$.

Theorem. Suppose that for each $i \in I_0$, p_i is independent of $\{p_j : j \neq i\}$. Then using the Benjamini–Hochberg procedure, the false discovery rate

$$FDR = \mathbb{E}\frac{N}{\max(R,1)} \leq \frac{\alpha M_0}{M} \leq \alpha.$$

5.2 Inference in high-dimensional regression

Theorem. Suppose the maximum eigenvalue of Σ is always at least $c_{min} > 0$ and $\max_j \Sigma_{jj} \leq 1$. Suppose further that $s_{max}\sqrt{\log(p)/n} \to 0$. Then there exists constants A_1, A_2 such that setting $\lambda = \lambda_j = A_1\sqrt{\log(p)/n}$, we have

$$\begin{aligned}\sqrt{n}(\hat{b} - \beta^0) &= W + \Delta \\ W \mid X &\sim N_p(0, \sigma^2\hat{\Theta}\hat{\Sigma}\hat{\Theta}^T),\end{aligned}$$

and as $n, p \to \infty$,

$$\mathbb{P}\left(\|\Delta\|_\infty > A_2 s\frac{\log(p)}{\sqrt{n}}\right) \to 0.$$