Part III — Modern Statistical Methods

Definitions

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

2 Kernel machines

2.1 Ridge regression

Definition (Ridge regression)**.** *Ridge regression* solves

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

where $\mathbf{1}$ is a vector of all 1's. Here $\lambda \geq 0$ is a tuning parameter, and it controls how much we penalize a large choice of β.

2.2 v-fold cross-validation

2.3 The kernel trick

Definition (Inner product space)**.** an inner product space is a real vector space $\mathcal{H}$ endowed with a map $\langle \cdot, \cdot \rangle : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ and obeys

- Symmetry: $\langle u, v\rangle = \langle v, u\rangle$
- Linearity: If $a, b \in \mathbb{R}$, then $\langle au + bw, v\rangle = a\langle u, v\rangle + b\langle w, v\rangle$.
- Positive definiteness: $\langle u, u\rangle \geq 0$ with $\langle u, u\rangle = 0$ iff $u = 0$.

Definition (Positive-definite kernel)**.** A *positive-definite kernel* (or simply *kernel*) is a symmetric map $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that for all $n \in \mathbb{N}$ and $x_1, \ldots, x_n \in \mathcal{X}$, the matrix $K \in \mathbb{R}^n \times \mathbb{R}^n$ with entries

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

is positive semi-definite.

Definition (Reproudcing kernel Hilbert space (RKHS))**.** A Hilbert space $\mathcal{B}$ of functions $f : \mathcal{X} \to \mathbb{R}$ is a reproducing kernel Hilbert space if for each $x \in \mathcal{X}$, there exists a $k_x \in \mathcal{B}$ such that

$$\langle k_x, f\rangle = f(x)$$

for all $x \in \mathcal{B}$.

The function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ given by

$$k(x, x') = \langle k_x, k_x'\rangle = k_x(x') = k_{x'}(x)$$

is called the *reproducing kernel* associated with $\mathcal{B}$.

2.4 Making predictions

2.5 Other kernel machines

2.6 Large-scale kernel machines

3 The Lasso and beyond

3.1 The Lasso estimator

3.2 Basic concentration inequalities

Definition (Sub-Gaussian random variable)**.** A random variable W is *sub-Gaussian* (with parameter σ) if

$$\mathbb{E}e^{\alpha(W-\mathbb{E}W)} \leq e^{\alpha^2\sigma^2/2}$$

for all $\alpha \in \mathbb{R}$.

Definition (Bernstein's condition)**.** We say that a random variable W satisfies Bernstein's condition with parameters (σ, b) where $a, b > 0$ if

$$\mathbb{E}[|W - \mathbb{E}W|^k] \leq \frac{1}{2}k!\,\sigma^2 b^{k-2}$$

for $k = 2, 3, \ldots$.

3.3 Convex analysis and optimization theory

Definition (Convex set)**.** A set $A \subseteq \mathbb{R}^d$ is convex if for any $x, y \in A$ and $t \in [0, 1]$, we have

$$(1-t)x + ty \in A.$$

Definition (Convex function)**.** A function $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ is *convex* iff

$$f((1-t)x + ty) \leq (1-t)f(x) + tf(y)$$

for all $x, y \in \mathbb{R}^d$ and $t \in (0, 1)$. Moreover, we require that $f(x) < \infty$ for at least one x.

We say it is *strictly convex* if the inequality is strict for all x, y and $t \in (0, 1)$.

Definition (Domain)**.** Define the *domain* of a function $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ to be

$$\operatorname{dom} f = \{x : f(x) < \infty\}.$$

Definition (Subgradient)**.** A vector $v \in \mathbb{R}^d$ is a *subgradient* of a convex function at x if $f(y) \geq f(x) + v^T(y - x)$ for all $y \in \mathbb{R}^d$.

The set of subgradients of f at x is denoted $\partial f(x)$, and is called the *subdifferential*.

Notation. For $x \in \mathbb{R}^d$ and $A \subseteq \{1 \ldots, d\}$, we write x_A for the sub-vector of x formed by the components of x induced by A. We write $x_{-j} = x_{\{j\}^c} = x_{\{1,\ldots,d\}\setminus j}$. Similarly, we write $x_{-jk} = x_{\{jk\}^c}$ etc.

We write

$$\operatorname{sgn}(x_i) = \begin{cases} -1 & x_i < 0 \\ 1 & x_i > 0 \\ 0 & \text{otherwise} \end{cases},$$

and $\operatorname{sgn}(x) = (\operatorname{sgn}(x_1), \ldots, \operatorname{sgn}(x_d))^T$.

3.4 Properties of Lasso solutions

Definition (Equicorrelation set)**.** Define the *equicorrelation set* $\hat{E}_\lambda$ to be the set of k such that

$$\frac{1}{n}|X_k^T(Y - X\hat{\beta}_\lambda^L)| = \lambda,$$

or equivalently, the k with $\nu_k = \pm 1$, which is well-defined since it depends only on the fitted values.

3.5 Variable selection

Definition (Compatibility factor)**.** Define the *compatibility factor* to be

$$\phi^2 = \inf_{\substack{\beta \in \mathbb{R}^p \\ \|\beta_N\|_1 \leq 3\|\beta_S\|_1 \\ \beta_S \neq 0}} \frac{\frac{1}{n}\|X\beta\|_2^2}{\frac{1}{s}\|\beta_S\|_1^2} = \inf_{\substack{\beta \in \mathbb{R}^p \\ \|\beta_S\|=1 \\ \|\beta_N\|_1 \leq 3}} \frac{s}{n}\|X_S\beta_S - X_N\beta_N\|_2^2.$$

Definition (Compatibility condition)**.** The *compatibility condition* is $\phi^2 > 0$.

3.6 Computation of Lasso solutions

3.7 Extensions of the Lasso

4 Graphical modelling

4.1 Conditional independence graphs

Definition (Graph)**.** A *graph* is a pair $\mathcal{G} = (V, E)$, where V is a set and $E \subseteq (V, V)$ such that $(v, v) \notin E$ for all $v \in V$.

Definition (Edge)**.** We say there is an *edge* between j and k and that j and k are *adjacent* if $(j, k) \in E$ or $(k, j) \in E$.

Definition (Undirected edge)**.** An edge (j, k) is *undirected* if also $(k, j) \in E$. Otherwise, it is *directed* and we write $j \to k$ to represent it. We also say that j is a *parent* of k, and write $\text{pa}(k)$ for the set of all parents of k.

Definition ((Un)directed graph)**.** A graph is *(un)directed* if all its edges are (un)directed.

Definition (Skeleton)**.** The *skeleton* of $\mathcal{G}$ is a copy of $\mathcal{G}$ with every edge replaced by an undirected edge.

Definition (Subgraph)**.** A graph $\mathcal{G}_1 = (V, E)$ is a *subgraph* of $\mathcal{G} = (V, E)$ if $V_1 \subseteq V$ and $E_1 \subseteq E$. A *proper subgraph* is one where either of the inclusions are proper inclusions.

Definition (Conditional independence)**.** Let X, Y, Z be random vectors with joint density f_{XYZ}. We say that X is *conditionally independent* of Y given Z, written $X \amalg Y \mid Z$, if

$$f_{XY|Z}(x, y \mid z) = f_{X|Z}(x \mid z) f_{Y|Z}(y \mid z).$$

Equivalently,

$$f_{X|YZ}(x \mid y, z) = f_{X|Z}(x \mid z)$$

for all y.

Definition (Conditional independence graph (CIG))**.** Let P be the law of $Z = (Z_1, \ldots, Z_p)^T$. The *conditional independent graph* (CIG) is the graph whose vertices are $\{1, \ldots, p\}$, and contains an edge between j and k iff Z_j and Z_k are conditionally dependent given Z_{-jk}.

Definition (Pairwise Markov property)**.** Let P be the law of $Z = (Z_1, \ldots, Z_p)^T$. We say P satisfies the *pairwise Markov property* with respect to a graph $\mathcal{G}$ if for any distinct, non-adjacent vertices j, k, we have $Z_j \amalg Z_k \mid Z_{-jk}$.

Definition (Separates)**.** Given a triple of (disjoint) subsets of nodes A, B, S, we say S *separates* A from B if every path from a node in A to a node in B contains a node in S.

Definition (Global Markov property)**.** We say P satisfies the *global Markov property* with respect to $\mathcal{G}$ if for any triple of disjoint subsets of V (A, B, S), if S separates A and B, then $Z_A \amalg Z_B \mid Z_S$.

Notation ($M_{A,B}$)**.** Let M be a matrix. Then $M_{A,B}$ refers to the submatrix given by the rows in A and columns in B.

4.2 Structural equation modelling

Definition (Path)**.** A *path* from j to k is a sequence $j = j_1, j_2, \ldots, j_m = k$ of (at least two) distinct vertices such that j_ℓ and $j_{\ell+1}$ are adjacent.

A path is *directed* if $j_\ell \to j_{\ell+1}$ for all ℓ.

Definition (Directed acyclic graph (DAG))**.** A *directed cycle* is (almost) a directed path but with the start and end points the same.

A *directed acyclic graph (DAG)* is a directed graph containing no directed cycles.

Definition (Structural equation model (SEM))**.** A *structural equation model* $\mathcal{S}$ for a random vector $Z \in \mathbb{R}^p$ is a collection of equations

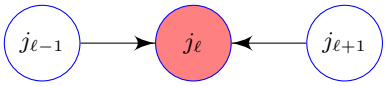
$$Z_k = h_k(Z_{p_k}, \varepsilon_k),$$

where $k = 1, \ldots, p$ and $\varepsilon_1, \ldots, \varepsilon_p$ are independent, and $p_k \subseteq \{1, \ldots, p\} \setminus \{k\}$ and such that the graph with $pa(k) = p_k$ is a directed acyclic graph.

Definition (Descendant)**.** We say k is a *descendant* of j if there is a directed path from j to k. The set of descendant of j will be denoted $\mathrm{de}(j)$.

Definition (Topological ordering)**.** Given a DAG $\mathcal{G}$ with $V = \{1, \ldots, p\}$ we say that a permutation $\pi : V \to V$ is a *topological ordering* if $\pi(j) < \pi(k)$ whenever $k \in \mathrm{de}(j)$.

Definition (Blocked)**.** In a DAG, we say a path $(j_1, \ldots, j_m)$ between j_1 and j_m is *blocked* by a set of nodes S (with neither j_1 nor j_m in S) if there is some $j_\ell \in S$ and the path is *not* of the form



or there is some j_ℓ such that the path *is* of this form, but neither j_ℓ nor any of its descendants are in S.

Definition (d-separate)**.** If $\mathcal{G}$ is a DAG, given a triple of (disjoint) subsets of nodes A, B, S, we say S *d-separates* A from B if S blocks every path from A to B.

Definition (v-structure)**.** A set of three nodes is called a *v-structure* if one node is a child of the two other nodes, and these two nodes are not adjacent.

Definition (Markov properties)**.** Let P be the distribution of Z and let f be the density. Given a DAG $\mathcal{G}$, we say P satisfies

(i) the *Markov factorization criterion* if

$$f(z_1, \ldots, z_p) = \prod_{k=1}^{p} f(z_k \mid z_{\mathrm{pa}(k)}).$$

(ii) the *global Markov property* if for all disjoint A, B, S such that A, B is d-separated by S, then $Z_A \amalg Z_B \mid Z_S$.

4.3 The PC algorithm

Definition (Causal minimality)**.** A distribution P satisfies *causal minimality* with respect to $\mathcal{G}$ but not any proper subgraph of $\mathcal{G}$.

Definition (Markov equivalence)**.** For a DAG $\mathcal{G}$, we let

$$\mathcal{M}(\mathcal{G}) = \{\text{distributions } P \text{ such that } P \text{ is Markov with respect to } \mathcal{G}\}.$$

We say two DAGs $\mathcal{G}_1, \mathcal{G}_2$ are are *Markov equivalent* if $\mathcal{M}(\mathcal{G}_1) = \mathcal{M}(\mathcal{G}_2)$.

Definition (Faithfulness)**.** We say P is *faithful* to a DAG $\mathcal{G}$ if it is Markov with respect to $\mathcal{G}$ and for all A, B, S disjoint, $Z_A \amalg Z_B \mid Z_S$ implies A, B are d-separated by S.

5 High-dimensional inference

5.1 Multiple testing

5.2 Inference in high-dimensional regression