Part II — Mathematics of Machine Learning

Based on lectures by R. Adhikari

Notes taken by Joseph Tedds using Dexter Chua's header and Gilles Castel's snippets.

Lent 2020

These notes are not endorsed by the lecturers, and I have modified them (often significantly) after lectures. They are nowhere near accurate representations of what was actually lectured, and in particular, all errors are almost surely mine.

Introduction to statistical learning

Concept of risk in Machine Learning. Empirical risk minimisation. The classification problem. Bayes risk. Concept of biascomplexity tradeoff. Examples of applications. [1]

Statistical learning theory

Basic inequality. Chernoff bounds. Sub-Gaussian random variables. Hoeffdings Lemma. Azuma Hoeffding inequality. Bounded differences inequality. Symmetrisation. Bound on excess misclassification risk in terms of Rademacher complexity. Rademacher complexity for finite hypothesis class. Shattering coefficient. VC dimension and statement of SauerShelah lemma. Examples of VC classes including finite-dimensional function classes. Rademacher complexity of ‘1 and ‘2-constrained hypothesis classes. [6]

Computation for empirical risk minimisation

Properties of convex sets and functions. Projections on to convex sets. Subgradients. Surrogate losses including logistic, exponential and hinge losses. Gradient descent and stochastic gradient descent. Analysis of convergence. [4]

Popular machine learning methods

Cross-validation. Feedforward neural networks. Decision trees. Adaboost. Gradient boosting. Random forests. [5]

Contents

0 Introduction **3**

0.1 Review of conditional expectation 3

0 Introduction

Consider a pair of random elements $(X, Y) \in (\mathcal{X}, \mathcal{Y})$ where X is our input features and Y the output. (X, Y) has joint distribution P_0. We wish to predict Y from X using $h : \mathcal{X} \to \mathcal{Y}$ known as a hypothesis. Measure the quality of our prediction using loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$.
Classification setting $Y \in \{0, 1\}$ typically take ℓ to be misclassification loss

$$\ell(h(x), y)) \begin{cases} 0 & \text{if } h(x) = y \\ 1 & \text{o/w} \end{cases}.$$

In this context, we also refer to h as a classifier.
Regression setting $Y \in \mathbb{R}$. Typically take ℓ to be squared error loss.

$$\ell(h(x), y) = (h(x) - y)^2.$$

We aim to choose h with small risk

$$R(h) = \int_{(x,y)\in(\mathcal{X},\mathcal{Y})} \ell(h(x), y) \mathrm{d}P_0(x, y) = \mathbb{E}\left[\ell(h(X), Y)\right].$$

This is true for a fixed h but if we choose $h = \hat{h}$ we need to use the first definition. Take ℓ and R to be misclassification (0-1) loss and 0-1 risk respectively unless otherwise stated. $\mathcal{X} = \mathbb{R}^p$.
The classifier that minimises 0-1 risk is called a *Bayes classifier* and its risk is the Bayes risk. Define the regression function $\eta(x) = \mathbb{P}\left(Y = 1|X = x\right)$.

Proposition. A Bayes classifier is given by

$$h_0(x) = \begin{cases} 1 & \text{if } \nu(x) > \frac{1}{2} \\ -1 & \text{o/w} \end{cases}.$$

But P_0 which determines η and h_0 is unknown. Instead, suppose we have training data $(X_1, Y_1), \ldots, (X_n, Y_n)$ iid copies of (X, Y). Our task is to construct $\hat{h}$ based on the training data such that $R(\hat{h})$ is small.

Remark. $F(\hat{h})$ is a random variable depending on the random training data. $R(\hat{h}) = \mathbb{E}\left[\ell(\hat{h}(X), Y)|X_1, Y_1, \ldots, X_n, Y_n\right]$.

Suppose we are given class $\mathcal{H}$ of classifiers from which to pick $\hat{h}$

Example.
- $\mathcal{H} = \{x \mapsto \mathrm{sgn}(\mu + x^T\beta), \mu \in \mathbb{R}, \beta \in \mathbb{R}^p\}$
- $\mathcal{H} = \{x \mapsto \mathrm{sgn}(\mu + \sum_{j=1}^d \ell_j(x)\beta_j), \mu \in \mathbb{R}, \beta \in \mathbb{R}^d\}$

Technical point, for the course $\mathrm{sgn}(0) = -1$.

0.1 Review of conditional expectation

Suppose $Z \in \mathbb{R}$ and $W \in \mathbb{R}^d$ are random elements

(i) Rule of independence if Z and W are independent then $\mathbb{E}\left[Z|W\right] = \mathbb{E}\left[Z\right]$

(ii) Tower property let $f : \mathbb{R}^d \to \mathbb{R}^m$. Then $\mathbb{E}\left[\mathbb{E}\left[Z|W\right]|f(W)\right] = \mathbb{E}\left[Z|f(W)\right]$. In particular $\mathbb{E}\left[\mathbb{E}\left[Z|W\right]|W_1, \dots, W_M\right] = \mathbb{E}\left[Z|W_1, \dots, W_m\right]$ for $m \le d$ and $\mathbb{E}\left[\mathbb{E}\left[Z|W\right]\right] = \mathbb{E}\left[Z\right]$ (take $f = c$).

(iii) Taking out what is known if $\mathbb{E}\left[Z^2\right] < \infty$ and $f : \mathbb{R}^d \to \mathbb{R}$ such that $\mathbb{E}\left[f(W)^2\right] < \infty$, then $\mathbb{E}\left[Zf(W)|W\right] = f(W)\mathbb{E}\left[Z|W\right]$.

Conditional Jensen $f : \mathbb{R} \to \mathbb{R}$ convex ($\{\ \forall\ x, y \in \mathbb{R}, t \in (0,1)\ tf(x) + (1-t)f(y) \ge f(tx + (1-t)y\}$) and $\mathbb{E}\left[|f(Z)|\right] < \infty$ then

$$\mathbb{E}\left[f(Z)|W\right] \ge f(\mathbb{E}\left[Z|W\right]).$$