Machine Learning

From the Perspective of Statistics

H.C.

Latest Update: June 18, 2021

Mathematical Notation

Suppose that $X : \Omega \to \mathbb{R}^p$ is a random variable. Sometimes we identify X with its image $\operatorname{im} X$ for simplicity.

- $X \sim P(x)$: The random variable X follows the distribution $P(x)$.
- $X_i \overset{\text{i.i.d}}{\sim} P(x)$ $(i = 1, \cdots, N)$: The independent and identically distributed random vectors X_i follow the distribution $P(x)$.
- $\mathbf{S}_N = \{X_i\}_{i=1}^N \overset{\text{i.i.d}}{\sim} P(x)$: $\mathbf{S}_N$ is a simple random sample from the distribution $P(x)$.

Chapter 1

Introduction

1.1 Terminology and Framework

- **Data generating process**: X is a p-dimensional random variable and Y is a 1-dimensional random variable. (X, Y) follows the distribution $P(x, y)$.
  - Input vector: $X \in D \subseteq \mathbb{R}^p$.
  - Output vector: $Y \in G \subseteq \mathbb{R}$.
  - Data: Given the training sample $\mathbf{S}_N = \{(X_i, Y_i)\}_{i=1}^N \overset{\text{i.i.d}}{\sim} P(x, y)$, the training data or training set consists of the realization value of the sample, that is

    $$T_N = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}.$$

    For another sample $\mathbf{S}'_M = \{(X'_i, Y'_i)\}_{i=1}^M \overset{\text{i.i.d}}{\sim} P(x, y)$, the test sample, we define the test data or test set as the realization value of $\mathbf{S}'_M$

    $$Q_M = \{(x'_1, y'_1), (x'_2, y'_2), \cdots, (x'_M, y'_M)\}.$$

- **Objective**: Given the training set T_N, find a decision function $\hat{f} \in \mathcal{F} \subset \mathbb{R}^{(\mathbb{R}^p)}$ that produces as little expected prediction error (EPE)

  $$\mathrm{E}_{X,Y}[L(Y, f(X))]$$

  as possible.
  - Decision function: $f : \mathbb{R}^p \supset D \to \mathbb{R}, x \mapsto f(x)$ serves to make a prediction of Y, provided a specified value x of X.
  - Loss function: $L : \mathbb{R}^2 \to [0, \infty)$ is a non-negative function that satisfies

    $$L(t_1, t_2) = 0 \iff t_1 = t_2 = 0.$$

    $L(Y, f(X))$ normally has the form of

    $$L_2 = (Y - f(X))^2 \text{ or } L_1 = |Y - f(X)| \text{ or } L_I = 1_{Y \neq f(X)} \text{ or } L_C = Y \log f(X) + (1 - Y) \log(1 - f(X)).$$

  - Hypothesis space: $\mathcal{F}$ is a collection of all potential decision functions f to be selected. In some cases, we suppose that f as a candidate can be specified by some parameters. Thus $\mathcal{F} = \{f_\theta | Y = f_\theta(X),\ \theta \in \Theta\}$ can be described by the parametric space Θ.

If the probability distribution of (X, Y) was known to us, it might succeed to find the optimal solution $\tilde{f}$ of the following minimization problem

$$\tilde{f} = \arg\min_{f \in \mathcal{F}} \mathrm{E}_{X,Y}[L(Y, f(X))]$$

in virtue of the optimization theory. In other words, what need to be settled is purely an optimization problem rather than a statistical problem. In particular, in the case of $Y = f(X)$, if $f \in \mathcal{F}$, then the optimal $\tilde{f}$ coincides with the function f itself and the minimum of EPE can reach 0.

Unfortunately, the specific distribution of (X, Y) is typically inaccessible in reality. Thus we can only exploit the training data to estimate an acceptable decision function $\hat{f}$, with acceptance of the fact that $\hat{f}$ in general has a greater EPE than $\tilde{f}$.

- **Learning Algorithm**: A learning algorithm is a measurable mappings $\mathsf{A} : (D \times G)^{\mathbb{N}} \to \mathcal{F}$ sending the training set T_N to the decision function $\mathsf{A}(T_N) = \hat{f} \in \mathcal{F}$, where $T_N \in (D \times G)^N$ is identified with an element in $(D \times G)^{\mathbb{N}}$ by the natural inclusion.

  Definition 1.1 (PAC identify) Assume $\tilde{f} = \arg\min_{f \in \mathcal{F}} \mathrm{E}[L(Y, f(X))]$. If there exists a learning algorithm A such that for all $\varepsilon > 0$, $\delta > 0$ and $P(x, y)$ on $\mathbb{R}^{p+1}$, there exists a positive integer N^* such that as long as $N \geq N^*$,

  $$\mathrm{P}_{\mathbf{S}_N}\left(\mathrm{E}_{X,Y}[L(Y, (\mathsf{A}(\mathbf{S}_N))(X)) \mid \mathbf{S}_N] - \mathrm{E}_{X,Y}[L(Y, \tilde{f}(X))] \leq \varepsilon\right) \geq 1 - \delta,$$

  we say that A can PAC (probably approximately correctly) identify $\tilde{f}$ from $\mathcal{F}$ and that $\tilde{f}$ is PAC identifiable.

- **Optimization strategies**: Since the EPE minimization problem is ill-formed, we have developed two major strategies to produce a tractable optimization problem.

  – Empirical risk minimization: according to the law of large numbers, if some regular conditions[1] hold, then as $N \to \infty$, the solution of the following minimization problem will converge to the theoretically optimal solution $\tilde{f}$.

  $$\min_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} L(Y_i, f(X_i))$$

  – Structural risk minimization: In reality, the size of training set N is limited and accordingly the method of empirical risk minimization may not generate a function $\hat{f}$ which is sufficiently close to $\tilde{f}$. Later we will elaborate this phenomenon named "overfitting". However, if we add a regularizer or penalty term $\lambda J(f)$ to penalize the complexity of the decision function f as follows

  $$\min_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} L(Y_i, f(X_i)) + \lambda J(f),$$

  it is possible to lead to a better result.

We can easily generalize the assumption that Y is a 1-dimensional random variable, supposing Y is an n-dimensional random vector. For example, in the multi-class classification, Y is often assumed to be an n-dimensional random vector with components from 0 to 1. And the loss function is taken as

$$L_C = \sum_{i=1}^{n} Y_i \log(f(X)_i)$$

[1] For example, the uniform law of large numbers can be applied here. The details are consigned to the appendix.

There are some other possible objectives involving more distribution information rather than minimizing $\mathrm{E}[L(Y, f(X))]$. For instance, we can expect to solve

$$\tilde{f} = \arg\min_{f \in \mathcal{F}} \left(\mathrm{E}[L(Y, f(X))] + c_1 \mathrm{Var}[L(Y, f(X))] + c_2 \texttt{ExtremeValueRisk}[L(Y, f(X))]\right).$$

More generally, suppose $\mathrm{Div}(P_1, P_2)$ is a non-negative function measuring the difference between the two distribution P_1 and P_2. We may expect to find a conditional distribution $\hat{P}(y|x)$ to minimize

$$\mathrm{E}[\mathrm{Div}(P(Y|X), \hat{P}(Y|X))].$$

1.2 Optimal Decision Functions in Special Cases

Supposing $\mathcal{F}$ is a collection of all functions $f : D \to G$, let's calculate the forms of the optimal decision function $\tilde{f}$ for some specific loss functions.

1.2.1 Loss function for quantitative output variables: squared error loss

Let $Y \in \mathbb{R}$ be a quantitative variable. And we take the most common and convenient loss function, squared error loss

$$L(Y, f(X)) = (Y - f(X))^2.$$

In this case, the problem of minimizing the expected prediction error becomes

$$\min_{f \in \mathcal{F}} \mathrm{E}[(Y - f(X))^2] = \min_{f \in \mathcal{F}} \int (y - f(x))^2 \, \mathrm{dP}(x, y).$$

Note that

$$\mathrm{E}[(Y - f(X))^2] = \mathrm{E}[\mathrm{E}[(Y - f(X))^2 |\, X]] = \int \mathrm{E}[(Y - f(x))^2 |\, X = x] \, \mathrm{dP}(x).$$

It suffices to minimize EPE pointwise, that is,

$$\begin{aligned}
&\min_{f(x) \in \mathbb{R}} \mathrm{E}[(Y - f(x))^2 |\, X = x] \\
&= \min_{c \in \mathbb{R}} \mathrm{E}[(Y - c)^2 |\, X = x] \\
&= \min_{c \in \mathbb{R}} \mathrm{Var}[Y - c |\, X = x] + (\mathrm{E}[(Y - c) |\, X = x])^2 \\
&= \min_{c \in \mathbb{R}} \mathrm{Var}[Y |\, X = x] + (\mathrm{E}[Y |\, X = x] - c)^2.
\end{aligned}$$

We see the optimal solution is

$$\tilde{f}(x) = \mathrm{E}[Y |\, X = x],$$

Thus the best prediction of Y at any point $X = x$ is the conditional expectation, when best is measured by average squared error.
Later in this book we are to develop effective methods to estimate the conditional expectation $\mathrm{E}[Y |\, X = x]$.

1.2.2 Loss function for categorical output variable: 0-1 indicator

Assume that $Y \in G$ is a categorical variable and that the set of possible categories $G = \{G_1, G_2, \cdots, G_K\}$. This time the 0–1 loss function

$$L(Y, f(X)) = 1_{Y \neq f(X)} = \begin{cases} 0, & Y = f(X), \\ 1, & Y \neq f(X), \end{cases}$$

is adopted for simplification. Likewise it suffices to minimize EPE pointwise.

$$\begin{aligned} &\min_{f(x) \in G} \mathrm{E}[1_{Y \neq f(x)} |\ X = x] \\ &= \min_{g \in G} \mathrm{E}[1 - 1_{Y=g} |\ X = x] \\ &= \min_{g \in G} 1 - \mathrm{P}(Y = g |\ X = x) \end{aligned}$$

And the optimal solution is

$$\tilde{f}(x) = \max_{g \in G} \mathrm{P}(Y = g |\ X = x)$$

1.3 Generalization Error Bound

As is mentioned before, we always hope that $\hat{f}$ has as small EPE as possible. EPE is also called the generalization error, indicating it gauges the performance of the selected function $\hat{f}$ in a general sense. We also emphasize that without knowing the probability distribution of (X, Y), there is no way to calculate the expectation of $L(Y, f(X))$. In practice, it is typical to analyze the upper bound of generalization error to describe the generalization ability of the selected function $\hat{f}$.

For the convenience of description, let's denote the expected prediction error and the empirical error respectively by

$$R(f) = \mathrm{E}[L(Y, f(X))] \quad \text{and} \quad \widehat{R}(f) = \frac{1}{N} \sum_{i=1}^{N} L\left(Y_i, f\left(X_i\right)\right).$$

In the case of binary classifications with a finite hypothesis space, deriving the generalization error bound is relatively simple. Actually, we have the following result.

Theorem 1.1 Suppose that the training sample $\mathbf{S}_N = \{(X_1, Y_1), (X_2, Y_2), \cdots, (X_N, Y_N)\}$ where $X_i \in D \subseteq \mathbb{R}^p$ and $Y_i \in \{+1, -1\}$ is independently generated from the distribution $P(x, y)$. If the hypothesis space is a finite set $\mathcal{F} = \{f_1, f_2, \cdots, f_d\}$ and the 0-1 loss function is taken, then with a probability of not less than $1 - \delta$ we have for all $f \in \mathcal{F}$,

$$\widehat{R}(f) - \sqrt{\frac{1}{2N}\left(\ln d + \ln \frac{1}{\delta}\right)} \leq R(f) \leq \widehat{R}(f) + \sqrt{\frac{1}{2N}\left(\ln d + \ln \frac{1}{\delta}\right)}.$$

Proof. The Hoeffding's inequality states that if $X_1, \cdots, X_n$ are independent random variables bounded by the interval $[0, 1]$: $0 \leq X_i \leq 1$, then $\overline{X} = \sum_{i=1}^{n} X_i / n$ satisfies

$$\mathrm{P}(\mathrm{E}[\overline{X}] - \overline{X} \geq t) \leq e^{-2nt^2}, \quad \mathrm{P}(\overline{X} - \mathrm{E}[\overline{X}] \geq t) \leq e^{-2nt^2}.$$

for all $t > 0$. Applying this inequality we get

$$\mathrm{P}\left(R(f_i) - \widehat{R}(f_i) \geq t\right) \leq e^{-2Nt^2}, \quad \mathrm{P}\left(\widehat{R}(f_i) - R(f_i) \geq t\right) \leq e^{-2Nt^2} \quad (i = 1, 2, \cdots, d).$$

Let A_i be the event which refers to $R(f_i) - \widehat{R}(f_i) \geq t$. We have

$$\mathrm{P}\left(\bigcup_{i=1}^{d} A_i\right) \leq \sum_{i=1}^{d} \mathrm{P}\left(A_i\right) \leq de^{-2Nt^2} \implies \mathrm{P}\left(\bigcap_{i=1}^{d} A_i^c\right) \geq 1 - de^{-2Nt^2}.$$

In other words, for any $f \in \mathcal{F}$, with a probability of not less than $1 - de^{-2Nt^2}$ we have

$$R(f) - \widehat{R}(f) < t \quad \text{or} \quad R(f) < \widehat{R}(f) + t.$$

Likewise, we can let B_i be the event which refers to $\widehat{R}(f_i) - R(f_i) \geq t$ and deduce

$$\widehat{R}(f) - R(f) < t \quad \text{or} \quad \widehat{R}(f) - t < R(f).$$

Take

$$t = \sqrt{\frac{1}{2N}\left(\ln d + \ln \frac{1}{\delta}\right)}$$

and we completes our proof.

□

It would be natural to consider binary classification problems with a infinite hypothesis space, which seems to be a more realistic assumption. To characterize the complexity of the infinite hypothesis space, we have to introduce some new concepts.

Definition 1.2 (dichotomies) The *dichotomies* generated by $\mathcal{F} \subseteq \{+1, -1\}^D$ on the points $x_1, \cdots, x_n \in D$ are defined by

$$\mathcal{F}\left(x_1, \cdots, x_n\right) = \left\{\left(f\left(x_1\right), \cdots, f\left(x_n\right)\right) | x_1, \cdots, x_n \in D,\ f \in \mathcal{F}\right\},$$

where $\{+1, -1\}^D$ consists of all mappings from D to $\{+1, -1\}$.

f sends each x_i to a specific value $+1$ or -1 as if it endows each x_i with a label. Thus we see $x_1, \cdots, x_n$ can be classified into two categories by these labels and that different f may result in different ways of classification.

Definition 1.3 (growth function) The *growth function* is defined for a hypothesis space $\mathcal{F}$ by

$$\Pi_{\mathcal{F}}(n) = \max_{\{x_1, \cdots, x_n\} \subseteq D} \left|\mathcal{F}\left(x_1, \cdots, x_n\right)\right|,$$

where $|\cdot|$ denotes the cardinality (number of elements) of a set.

$\Pi_{\mathcal{F}}(n)$ is the maximum number of dichotomies that can be generated by $\mathcal{F}$ on any n points. Since $\mathcal{F}\left(x_1, \cdots, x_n\right) \subseteq \{+1, -1\}^n$, it clearly holds that

$$\Pi_{\mathcal{F}}(n) \leq 2^n.$$

If $\mathcal{F}$ is capable of generating all possible dichotomies on $\{x_1, \cdots, x_n\}$, then $\mathcal{F}\left(x_1, \cdots, x_n\right) = \{+1, -1\}^n$ and we say that $\mathcal{F}$ can *shatter* $x_1, \cdots, x_n$. This signifies that $\mathcal{F}$ is as diverse as can be on this particular sample.

Definition 1.4 (VC dimension) The Vapnik-Chervonenkis dimension of a hypothesis space $\mathcal{F}$, denoted by $d_{VC}(\mathcal{F})$ or simply d_{VC}, is the largest value of N for which $\Pi_{\mathcal{F}}(n) = 2^n$. That is,

$$d_{VC}(\mathcal{F}) = \max\left\{m \mid \Pi_{\mathcal{F}}(m) = 2^m\right\}$$

If $\Pi_{\mathcal{F}}(n) = 2^n$ for all n, then $d_{VC}(\mathcal{F}) = \infty$.

It is straightforward to show that if $\Pi_{\mathcal{F}}(m) = 2^m$ then $\Pi_{\mathcal{F}}(n) = 2^n$ for all $n < m$, since $\mathcal{F}(x_1, \cdots, x_m) = \{+1, -1\}^m \implies \mathcal{F}(x_1, \cdots, x_{m-1}) = \{+1, -1\}^{m-1}$. Thus we have

$$\begin{cases} \Pi_{\mathcal{F}}(n) = 2^n, & \text{if } n \le d_{VC}(\mathcal{F}), \\ \Pi_{\mathcal{F}}(n) < 2^n, & \text{if } n > d_{VC}(\mathcal{F}). \end{cases}$$

If $\Pi_{\mathcal{F}}(k) < 2^k$, we say k is a *break point* for $\mathcal{F}$. Next we are to find a more accurate bound for $\Pi_{\mathcal{F}}(n)$ at the break point.

Definition 1.5 Let $\mathfrak{G} = \{\mathcal{F} \subseteq \{+1, -1\}^D \mid \Pi_{\mathcal{F}}(k) < 2^k\}$. Define

$$B(n, k) = \max_{\mathcal{F} \in \mathfrak{G},\, \{x_1, \cdots, x_n\} \subseteq D} |\mathcal{F}(x_1, \cdots, x_n)|.$$

$B(n, k)$ is the maximum number of dichotomies which can be generated by some hypothesis space on n points in D such that these dichotomies cannot shatter any subsets of size k of the n points. If k is a break point for $\mathcal{F}$ and $n \ge k$, then we can see $\Pi_{\mathcal{F}}(n) \le B(n, k)$ just from their definitions.

To evaluate $B(n, k)$, we start with the two boundary conditions $k = 1$ and $n = 1$.

$$\begin{aligned} &B(n, 1) = 1, \\ &B(1, k) = 2 \text{ for } k > 1. \end{aligned}$$

$B(n, 1) = 1$ for all n since if no subset of size 1 can be shattered, then only one dichotomy can be allowed. A second different dichotomy must differ on at least one point and then that subset of size 1 would be shattered. $B(l, k) = 2$ for $k > 1$ since in this case there do not even exist subsets of size k; the constraint is vacuously true and we have 2 possible dichotomies ($+1$ and -1) on the one point .

Theorem 1.2 If $d_{VC} < \infty$, then for all $n \ge 1$,

$$\Pi_{\mathcal{F}}(n) \le \sum_{i=0}^{d_{VC}} \binom{n}{i} \le n^{d_{VC}} + 1.$$

Corollary 1.1 If $d_{VC} < \infty$, then for all $n \ge d_{vc}$,

$$\Pi_{\mathcal{F}}(n) \le \left(\frac{\mathrm{e} \cdot n}{d_{vc}}\right)^{d_{vc}}$$

Theorem 1.3 Suppose that the training sample $\mathbf{S}_N = \{(X_1, Y_1), (X_2, Y_2), \cdots, (X_N, Y_N)\}$ where $X_i \in D \subseteq \mathbb{R}^p$ and $Y_i \in \{+1, -1\}$ is independently generated from the distribution $P(x, y)$. If the 0-1 loss function is taken, then for any $\mathcal{F} \subseteq \{+1, -1\}^D$, $f \in \mathcal{F}$, tolerance $\delta > 0$,

$$R(f) \le \widehat{R}(f) + \sqrt{\frac{8}{N} \ln \frac{4\Pi_{\mathcal{F}}(2N)}{\delta}} \le \widehat{R}(f) + \sqrt{\frac{8}{N} \ln \frac{4(2N)^{d_{VC}} + 4}{\delta}}$$

with probability $\ge 1 - \delta$.

Corollary 1.2 Suppose that the training sample $\mathbf{S}_N = \{(X_1, Y_1), (X_2, Y_2), \cdots, (X_N, Y_N)\}$ where $X_i \in D \subseteq \mathbb{R}^p$ and $Y_i \in \{+1, -1\}$ is independently generated from the distribution $P(x, y)$. If the 0-1 loss function is taken, then for any $\mathcal{F} \subseteq \{+1, -1\}^D$, $f \in \mathcal{F}$, $N \ge d_{VC}$, tolerance $\delta > 0$,

$$R(f) \le \widehat{R}(f) + \sqrt{\frac{8 d_{VC} \ln \frac{2\mathrm{e}N}{d_{VC}} + 8 \ln \frac{4}{\delta}}{N}}$$

with probability $\ge 1 - \delta$.

Theorem 1.4 If $d_{VC} < \infty$, empirical risk minimization algorithm

$$\mathsf{A}_E : T_N \longmapsto \hat{f} = \arg\min_{f \in \mathcal{F}} \widehat{R}(f)$$

can PAC (probably approximately correctly) identify $\tilde{f}$ from $\mathcal{F}$.

Proof. Given any $\varepsilon > 0$, $\delta > 0$, $P(x, y)$ on $\mathbb{R}^{p+1}$, according to theorem 1.1 we take $\mathcal{F} = \{\tilde{f}\}$ and

$$\frac{\varepsilon}{2} = \sqrt{\frac{\ln(2/\delta)}{2N_1}},$$

which leads to

$$P\left(\widehat{R}(\tilde{f}) - R(\tilde{f}) > \frac{\varepsilon}{2}\right) \leq \frac{\delta}{2}.$$

According to theorem 1.3, let

$$\frac{\varepsilon}{2} = \sqrt{\frac{8}{N_2} \ln \frac{8(2N_2)^{d_{VC}} + 8}{\delta}}$$

and then we have

$$P\left(R(\hat{f}) - \widehat{R}(\hat{f}) > \frac{\varepsilon}{2}\right) \leq \frac{\delta}{2}.$$

Hence there must be that

$$\begin{aligned}
\mathrm{P}\left(R(\hat{f}) - R(\tilde{f}) > \varepsilon\right) &\leq \mathrm{P}\left(R(\hat{f}) - R(\tilde{f}) + \widehat{R}(\tilde{f}) - \widehat{R}(\hat{f}) > \varepsilon\right) \qquad \left(\text{ since } \widehat{R}(\tilde{f}) - \widehat{R}(\hat{f}) \geq 0 \right) \\
&\leq \mathrm{P}\left(\widehat{R}(\tilde{f}) - R(\tilde{f}) > \frac{\varepsilon}{2} \text{ or } R(\hat{f}) - \widehat{R}(\hat{f}) > \frac{\varepsilon}{2}\right) \\
&\leq \mathrm{P}\left(\widehat{R}(\tilde{f}) - R(\tilde{f}) > \frac{\varepsilon}{2}\right) + \mathrm{P}\left(R(\hat{f}) - \widehat{R}(\hat{f}) > \frac{\varepsilon}{2}\right) \\
&\leq \frac{\delta}{2} + \frac{\delta}{2} \\
&\leq \delta.
\end{aligned}$$

Fix ε and solve for δ

$$\delta = \frac{2}{\mathrm{e}^{N_1\varepsilon^2/2}} = \frac{8(2N_2)^{d_{VC}} + 8}{\mathrm{e}^{N_2\varepsilon^2/32}}.$$

Note that when N_1, N_2 are sufficiently large these inequalities hold for even smaller $\delta > 0$. We conclude that there exists a positive integer N^* such that for all $N \geq N^*$,

$$\mathrm{P}\left(\mathrm{E}[L(Y, \hat{f}(X))] - \mathrm{E}[L(Y, \tilde{f}(X))] \leq \varepsilon\right) \geq 1 - \delta.$$

□

1.4 tradeoff

The generalization error bound $\widehat{R}(f) + \varepsilon(d_{VC}, N, \delta)$ helps up understand what is "overfitting" and why we can use structural risk minimization to get over it.

Assume N is fixed and $\hat{f} = \arg\min_{f \in \mathcal{F}} \widehat{R}(f)$. By increasing d_{VC} endlessly, we can extend the hypothesis space $\mathcal{F}$ and consequently reduce or at least maintain the empirical error $\widehat{R}(\hat{f})$ all along. In other words, $\widehat{R}(\hat{f})$ is non-increasing in d_{VC}. However, the cost it brings is a greater $\varepsilon(d_{VC}, N, \delta)$, since $\varepsilon(d_{VC}, N, \delta)$ is strictly increasing in d_{VC}. Noticing $\widehat{R}(\hat{f})$ cannot be less than 0 and $\varepsilon(d_{VC}, N, \delta) \to \infty$ as $d_{VC} \to \infty$, we can assert that when d_{VC} is sufficiently large, generalization error bound will be also

increasing in d_{VC}. As a result, although we can find a function $\hat{f}$ with a perfect performance on the training set, the generalization ability of $\hat{f}$ can be very poor. That is to say, $\hat{f}$ is likely to perform badly on the sample out of the training set. The term "overfitting" exactly refers to such a case in which the selected function $\hat{f}$ has a quite small empirical error along with a tremendous generalization error.

Therefore, we have a tradeoff: more complex models help reduce $\widehat{R}(f)$ and hurt $\varepsilon(d_{VC}, N, \delta)$. The optimal model is a compromise that minimizes a combination of the two terms. Now it is clear to see why a penalty term is introduced into the structural risk minimization. It takes the size of hypothesis space $\mathcal{F}$ or equivalently the complexity of candidate functions into consideration. Thus when N is limited, this method promisingly leads to a smaller generalization error.

Chapter 2

Linear Model

2.1 Finite Sample Linear Model

2.1.1 Statistic model setup

Linear model supposes the data generating process is

$$Y = X^T\beta + \varepsilon,$$

where $X = (1, X^{(1)}, \cdots, X^{(p)}) \in \mathbb{R}^{p+1}$, $Y \in \mathbb{R}$, $\beta \in \mathbb{R}^{p+1}$ is an unknown parameter and ε is an error term which cannot be directly observed. Without loss of generality, we can always assume that $\mathrm{E}[\varepsilon|\ X] = 0$. Given a finite sample $(\mathbf{X}, \mathbf{y})$ of size n, the linear Model indicates

$$\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\varepsilon},$$

where

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{pmatrix},\ \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{N1} & \cdots & x_{Np} \end{pmatrix} = \begin{pmatrix} X_1^T \\ X_2^T \\ \vdots \\ X_N^T \end{pmatrix},\ \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix},\ \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{pmatrix}.$$

2.1.2 Ordinary least square estimate

If we take squared error loss, then the best prediction of Y is

$$\tilde{f}(X) = \mathrm{E}[Y|\ X] = X'\beta.$$

We can use least square method to estimate the parameter β in the linear model, by minimizing the residual sum-of-squares

$$\hat{\beta} = \arg\min_{\theta \in \mathbb{R}^{p+1}} RSS(\theta) = \arg\min_{\theta \in \mathbb{R}^{p+1}} \sum_{i=1}^{N} (Y_i - X_i^T\theta)^2.$$

If $\mathbf{X}$ is of full column rank, the optimization problem has a unique solution

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

The statistic properties of the OLS estimator $\hat{\beta}$ is remarkable. According to Gauss-Markov theorem, Under Assumptions 1.1-1.4

1.1 linearity: $Y_i = X_i^T \beta + \varepsilon_i$, $(i = 1, 2, \cdots, N)$,

1.2 strict exogeneity: $\mathrm{E}[\boldsymbol{\varepsilon}|\ \mathbf{X}] = 0$,

1.3 no multicollinearity: $\mathrm{P}(\mathrm{rank}(\mathbf{X}) = p + 1) = 1$,

1.4 spherical error variance: $\mathrm{Var}[\boldsymbol{\varepsilon}|\ \mathbf{X}] = \sigma^2 > 0$,

$\hat{\beta}$ is the best linear unbiased estimator(BLUE). That is, for any unbiased estimator $\hat{\theta}$ that is linear in $\mathbf{Y}$,

$$\mathrm{Var}[\hat{\theta}|\ \mathbf{X}] \geq \mathrm{Var}[\hat{\beta}|\ \mathbf{X}]$$

in the matrix sense.

Chapter 3

SVM

3.1 linear SVM

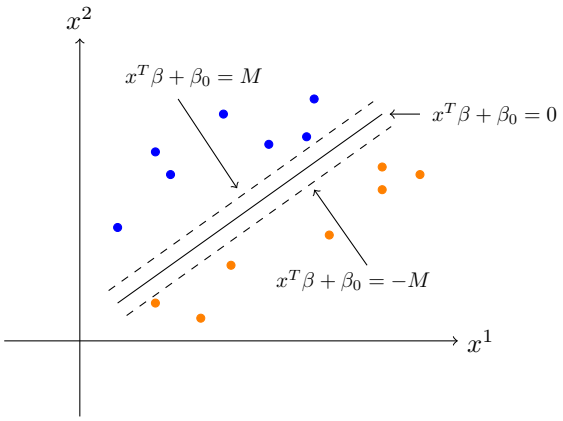
Assume that $X \in \mathbb{R}^p$ and that $Y \in G$ is a categorical variable with the set of possible categories $G = \{+1, -1\}$. Linear support vector machine uses the following kind of decision functions to model the relationship between X and Y:

$$Y = \text{sign}\left(X^T\beta + \beta_0\right).$$

Suppose we have already had some training data $T_N = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$. If we can find a superplane $x^T\beta + \beta_0 = 0$ such that

$$x_i^T\beta + \beta_0 > 0 \text{ if } y_i = +1,$$
$$x_i^T\beta + \beta_0 < 0 \text{ if } y_i = -1,$$

then the empirical loss of the decision function $\hat{f}(X) = \text{sign}\left(X^T\beta + \beta_0\right)$ on the training data T_N can reach 0. In this case, we say the two classes are linearly separable and our aim is to maximize the "margin" shown as follows.



The distance between the superplane $P_0 : x^T\beta + \beta_0 = 0$ and $P_1 : x^T\beta + \beta_0 = M$ is $\frac{M}{\|\beta\|}$, which can be seen from

$$M = (x - y)^T\beta \leq \|x - y\| \cdot \|\beta\|, \ \forall x \in P_1, \forall y \in P_0.$$

Thus the maximization problem can be stated as

$$\begin{aligned} &\max_{\beta,\beta_0,M} \frac{M}{\|\beta\|} \\ \text{s.t.} \quad &y_i\left(x_i^T\beta+\beta_0\right) \geq M,\ i=1,\ldots,N. \end{aligned}$$

In general, classes usually overlap in feature space. One way to deal with the overlap is to still maximize $\frac{M}{\|\beta\|}$, but allow for some points to be on the wrong side of the margin. By defining the slack variables $\xi = (\xi_1, \xi_2, \ldots, \xi_N)$, we can modify the problem.

$$\begin{aligned} &\max_{\beta,\beta_0,M} \frac{M}{\|\beta\|} \\ \text{s.t.} \quad &y_i\left(x_i^T\beta+\beta_0\right) \geq M(1-\xi_i),\ i=1,\ldots,N, \\ &\xi_i \geq 0, \\ &\sum_{i=1}^{N}\xi_i \leq U. \end{aligned}$$

If we set $M = 1$, it yields the following maximization problem

$$\begin{aligned} &\max_{\beta,\beta_0} \frac{1}{\|\beta\|} \\ \text{s.t.} \quad &y_i\left(x_i^T\beta+\beta_0\right) \geq 1,\ i=1,\ldots,N, \end{aligned}$$

which is equivalent to the following quadratic programming problem

$$\begin{aligned} &\min_{\beta,\beta_0} \frac{1}{2}\|\beta\|^2 \\ \text{s.t.} \quad &y_i\left(x_i^T\beta+\beta_0\right) \geq 1,\ i=1,\ldots,N. \end{aligned}$$

The optimal solutions of the problem must satisfy some necessary conditions (Karush-Kuhn-Tucker conditions). Consider the Lagrangian function

$$L(\beta,\beta_0,\lambda) = \frac{1}{2}\|\beta\|^2 - \sum_{i=1}^{N}\lambda_i\left(y_i\left(x_i^T\beta+\beta_0\right)-1\right),$$

we have

$$\begin{aligned} &\frac{\partial L}{\partial \beta} = \beta^T - \sum_{i=1}^{N}\lambda_i y_i x_i^T = 0, \\ &\frac{\partial L}{\partial \beta_0} = \sum_{i=1}^{N}\lambda_i y_i = 0, \\ &\lambda_i\left(y_i\left(x_i^T\beta+\beta_0\right)-1\right) = 0, \\ &\lambda_i \geq 0, \\ &y_i\left(x_i^T\beta+\beta_0\right)-1 \geq 0. \end{aligned}$$

Cancel β, β_0 and we obtain the dual problem

$$\max_{\lambda} \; -\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j y_i y_j x_i^T x_j + \sum_{i=1}^{N}\lambda_i$$
$$\text{s.t.} \quad \sum_{i=1}^{N}\lambda_i y_i = 0,$$
$$\lambda_i \geq 0.$$

3.2 linear SVM

Chapter 4

KNN

4.1 Nearest-neighbor methods

Nearest-neighbor methods use those observations in the training set T closest in input space to x to estimate the aforementioned conditional expectation $\mathrm{E}[Y|\ X = x]$. Specifically, the k-nearest neighbor fit for $\hat{Y}$ is defined as follows:

- Quantitative output
$$\hat{Y} = \hat{f}(x) = \frac{1}{k} \sum_{i:x_i \in N_k(x)} y_i,$$
- Categorical output
$$\hat{Y} = \hat{f}(x) = \arg\max_{g \in G} \sum_{i:x_i \in N_k(x)} 1_{y_i = g},$$

where $N_k(x)$ is the neighborhood of x defined by the k closest points x_i in the training sample. Closeness implies a metric, which for the moment we assume is Euclidean distance

$$\|x_1 - x_2\|_2 = (x_1 - x_2)^2.$$

.

Chapter 5

Decision Tree

Tree-based methods partition the feature space into a set of rectangles, and then fit a simple model (like a constant) in each one. We first describe a popular tree-based method called CART (classification and regression tree), and later contrast it with C4.5, a major competitor. CART requires a specific hypothesis space $\mathcal{F}$ which consists of functions of such forms:

$$f(x) = \sum_{m=1}^{M} c_m 1\left[x \in R_m\right],$$

where p-dimensional rectangular regions $R_1, R_2, \cdots, R_M$ constitute a recursive partition of the feature space $\mathbb{R}^p$. That is, if a partition $P_J = (R_1, R_2, \cdots, R_J)$ has been obtained, we just cut some rectangular region $R_m (1 \leq m \leq J)$ into two parts by a $p-1$-dimensional hyperplane $x^{(i)} = \ell_J (1 \leq i \leq p)$ to produce a new partition $P_{J+1} = (R'_1, R'_2, \cdots, R'_{J+1})$

5.1 Regression Trees

Assume the output Y is a continuous variable. Then it leads to the concept of regression trees. Conventionally we choose the squared error loss and accordingly obtain the minimization problem

$$\begin{aligned}
\min_{f\in\mathcal{F}} \mathrm{E}[(Y - f(X))^2] &= \min_{f\in\mathcal{F}} \mathrm{E}\left[\left(\sum_{m=1}^{M} (Y - c_m) 1_{X\in R_m}\right)^2\right] \\
&= \min_{f\in\mathcal{F}} \mathrm{E}\left[\sum_{m=1}^{M} (Y - c_m)^2 1_{X\in R_m}\right] \\
&= \min_{f\in\mathcal{F}} \int \mathrm{E}\left[\sum_{m=1}^{M} (Y - c_m)^2 1_{x\in R_m} \middle| X = x\right] \mathrm{dP}(x) \\
&= \min_{f\in\mathcal{F}} \sum_{m=1}^{M} \int_{R_m} \mathrm{E}\left[(Y - c_m)^2 \middle| X = x\right] \mathrm{dP}(x).
\end{aligned}$$

It is easy to see whichever partition is specified, the optimal c_m remains

$$\tilde{c}_m = \mathrm{E}[Y | \, X \in R_m].$$

Now let's consider the empirical risk minimization problem

$$\min_{f\in\mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} L(Y_i, f(X_i)) = \min_{f\in\mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} \sum_{m=1}^{M} (Y_i - c_m)^2 1_{X_i\in R_m}.$$

Thus the proper estimate of $\hat{c}_m$ is just the average of Y_i in region R_m

$$\hat{c}_m = \text{ave}\left(Y_i | X_i \in R_m\right).$$

However, finding the best binary partition in terms of minimum sum of squares is generally computationally infeasible. Hence we proceed with a greedy algorithm. Starting with all of the data, consider a splitting variable indexed by $r \in \{1, 2, \cdots, p\}$ and split point $s \in \mathbb{R}$, and define the pair of half-planes

$$R_1(r,s) = \left\{X \in \mathbb{R}^p | X^{(r)} \leq s\right\} \text{ and } R_2(r,s) = \left\{X \in \mathbb{R}^p | X^{(r)} > s\right\}.$$

Then we seek the splitting variable r and split point s that solve

$$\begin{aligned} &\min_{r,s}\left[\min_{c_1}\sum_{X_i \in R_1(r,s)}\left(Y_i - c_1\right)^2 + \min_{c_2}\sum_{X_i \in R_2(r,s)}\left(Y_i - c_2\right)^2\right] \\ = &\min_{r,s}\sum_{X_i \in R_1(r,s)}\left(Y_i - \hat{c}_1\right)^2 + \sum_{X_i \in R_2(r,s)}\left(Y_i - \hat{c}_2\right)^2. \end{aligned}$$

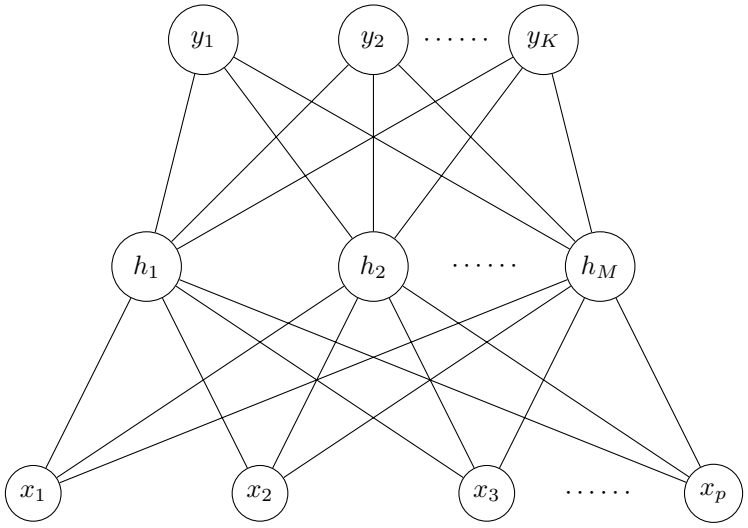
Having found the best split, we partition the data into the two resulting regions and repeat the splitting process on each of the two regions. Then this process is repeated on all of the resulting regions.

Clearly a very large tree might overfit the data until some minimum node size (say 5) is reached.

Chapter 6

Neural Networks

6.1 Single hidden layer back-propagation network



6.1.1 Binary classification

In the matrix form

$$y = \sigma(W^{(2)}\mathbf{h} + \mathbf{b}^{(2)}), \mathbf{h} = \sigma(W^{(1)}\mathbf{x} + \mathbf{b}^{(1)}),$$

where $\sigma(x_i) = \frac{1}{1+e^{-x_i}}$.

6.1.2 Multi classification

In the matrix form

$$\mathbf{y} = \texttt{softmax}(W^{(2)}\mathbf{h} + \mathbf{b}^{(2)}), \mathbf{h} = \sigma(W^{(1)}\mathbf{x} + \mathbf{b}^{(1)}),$$

where

$$\texttt{softmax}_k(\mathbf{x}) = \frac{e^{x_k}}{\sum_{i=1}^{K} e^{x_i}}.$$

6.2 Convolutional neural network

$$\texttt{Conv2d}(X_{C_{\text{in}},H,W}) = b_{C_{\text{out}}} \otimes 1_{H,W} + \text{Weight}_{C_{\text{out}},S*S} \star X_{C_{\text{in}},H,W}$$

Appendix

1. Hoeffding's inequality

Lemma 6.1 Let X be a random variable. Then

$$\mathrm{P}(X > \epsilon) \leq \inf_{t \geq 0} \frac{\mathrm{E}\left[e^{tX}\right]}{e^{t\epsilon}}.$$

Proof. For any $t > 0$, we have

$$\mathrm{P}(X > \epsilon) = \mathrm{P}\left(e^{tX} > e^{t\epsilon}\right) = \mathrm{P}\left(\frac{e^{tX}}{e^{t\epsilon}} > 1\right) = \mathrm{E}\left(\mathbb{1}\left\{\frac{e^{tX}}{e^{t\epsilon}} > 1\right\}\right) \leq \frac{\mathrm{E}\left[e^{tX}\right]}{e^{t\epsilon}}.$$

Notice that if $t = 0$,

$$\frac{\mathrm{E}\left[e^{tX}\right]}{e^{t\epsilon}} = 1 \geq \mathrm{P}(X > \epsilon).$$

Thus we complete the proof. □

Lemma 6.2 Suppose X is a random variable such that $a \leq X \leq b$. Then for $t \in \mathbb{R}$,

$$\mathrm{E}\left[e^{tX}\right] \leq e^{t\mu} e^{\frac{t^2(b-a)^2}{8}}$$

where $\mu = \mathbb{E}[X]$.

Proof. Since $f(x) = e^{tx}$ is convex, by Jensen's inequality we get for $x \in [a, b]$

$$e^{tx} \leq \frac{b-x}{b-a} e^{ta} + \frac{x-a}{b-a} e^{tb}.$$

Let Y be a random variable such that $\mathrm{E}[Y] = 0$ and let $\lambda = -a/(b-a)$, we have

$$\mathrm{E}[e^{tY}] \leq \mathrm{E}\left[\frac{b-Y}{b-a} e^{ta} + \frac{Y-a}{b-a} e^{tb}\right] = \frac{be^{ta} - ae^{tb}}{b-a} = (1-\lambda)e^{-t\lambda(b-a)} + \lambda e^{t(1-\lambda)(b-a)}.$$

Suppose $u = t(b-a)$ and

$$g(u) = \ln\left((1-\lambda)e^{-\lambda u} + \lambda e^{(1-\lambda)u}\right) = -\lambda u + \ln\left(1 - \lambda + \lambda e^{u}\right), \qquad u > 0.$$

Taylor's theorem implies that there exists $\xi \in [0, u]$ such that

$$g(u) = g(0) + g'(0)u + \frac{g''(\xi)}{2!}u^2 = \frac{(1-\lambda)\lambda e^{\xi}}{2(1-\lambda+\lambda e^{\xi})^2}u^2.$$

Since

$$(1-\lambda+\lambda e^{\xi})^2 \geq 4(1-\lambda)\lambda e^{\xi} \implies \frac{4(1-\lambda)\lambda e^{\xi}}{(1-\lambda+\lambda e^{\xi})^2} \leq 1,$$

we have $g(u) \leq \frac{1}{8}u^2$. Thus we have

$$\mathrm{E}[e^{tY}] \leq e^{g(u)} \leq e^{\frac{1}{8}u^2} = e^{\frac{t^2(b-a)^2}{8}}.$$

Take $Y = X - \mu$, we can show that

$$\mathrm{E}\left[e^{tX}\right] \leq e^{t\mu} e^{\frac{t^2(b-a)^2}{8}}.$$

□

Theorem 6.1 (Hoeffding's Inequality) Let $X_1, \ldots, X_n$ be i.i.d.random variables such that $\mathbb{E}(X_i) = \mu$ and $a \leq X_i \leq b$. Then, for any $\epsilon > 0$

$$\mathrm{P}\left(\left|\bar{X}_n - \mu\right| \geq \epsilon\right) \leq 2e^{-2n\epsilon^2/(b-a)^2}$$

Proof. Since $a \leq X_i \leq b$, we have $a - \mu \leq X_i - \mu \leq b - \mu$ and

$$\mathrm{P}(X_i - \mu \geq \epsilon) \leq \inf_{t \geq 0} \frac{\mathrm{E}\left[e^{t(X_i-\mu)}\right]}{e^{t\epsilon}} \leq \inf_{t \geq 0} \frac{e^{\frac{t^2(b-a)^2}{8}}}{e^{t\epsilon}} = e^{\frac{-4\epsilon}{(b-a)^2}}.$$

□