Chapter 5
Statistical Learning Theory

supplementary slides to
Machine Learning Fundamentals
© Hui Jiang 2020
published by Cambridge University Press

August 2020
Outline

1. Formulation of Discriminative Models
2. Learnability
3. Generalization Bounds
Formulation of Discriminative Models

- input \( x \) is an \( n \)-dimensional vector from input space \( X \), e.g.
  - \( X = \mathbb{R}^n \) for unconstrained inputs
  - \( X = [0, 1]^n \) for constrained inputs

- output \( y \) from an output space \( Y \):
  - \( Y \) is finite for classification
  - \( Y \) is continuous for regression, e.g. \( Y = \mathbb{R} \).

- formulation of discriminative models
  - inputs \( x \) are random vectors: \( x \sim p(x) \) (\( \forall x \in X \))
  - \( \forall x \in X \), the corresponding output \( y \) is generated by an unknown deterministic target function function, i.e. \( y = \bar{f}(x) \)
the goal of discriminative modeling is to learn the unknown target function from a pre-specified *model space* $\mathcal{H}$

based on a training set of a finite number of samples:

$$\mathcal{D}_N = \left\{ (x_i, y_i) \mid i = 1, \cdots, N \right\}$$

where $x_i$ is an independent sample drawn from the distribution $p(x)$, i.e. $x_i \sim p(x)$, and $y_i = \bar{f}(x_i)$ for all $i = 1, 2, \cdots, N$.

we can only learn a model $y = f(x)$ from $\mathcal{H}$, i.e. $f(\cdot) \in \mathcal{H}$, which resembles the target function $\bar{f}(x)$ as much as possible
Statistical Learning Theory: Discriminative Models (II)

- introduce a loss function \( l(y, y') \) to measure the learning error
  - zero-one loss for classification: \( l(y, y') = \begin{cases} 0 & (y = y') \\ 1 & (y \neq y') \end{cases} \)
  - squared error for regression: \( l(y, y') = (y - y')^2 \)

- empirical loss (a.k.a. in-sample error) of any \( f(\cdot) \in H \): 
  \[
  R_{\text{emp}}(f|\mathcal{D}_N) = \frac{1}{N} \sum_{i=1}^{N} l(y_i, f(x_i))
  \]

- expected loss (a.k.a. generalization error) of \( f(\cdot) \in H \):
  \[
  R(f) = \mathbb{E}_{x \sim p(x)} \left[ l(\bar{f}(x), f(x)) \right] = \int_{x \in X} l(\bar{f}(x), f(x)) p(x) \, dx
  \]

\( R(f) \neq R_{\text{emp}}(f|\mathcal{D}_N) \) but \( \lim_{N \to \infty} R_{\text{emp}}(f|\mathcal{D}_N) = R(f) \)
Statistical Learning Theory: Learnability

- empirical risk minimization (ERM) aims to minimize the empirical loss in $\mathbb{H}$:

$$f^* = \arg\min_{f \in \mathbb{H}} R_{\text{emp}}(f | \mathcal{D}_N)$$

- the problem is learnable or not:
  - whether ERM can lead to a small generalization error, i.e., $R(f^*)$ is sufficiently small

- learnability depends on the following gap:

$$\left| R(f^*) - R_{\text{emp}}(f^* | \mathcal{D}_N) \right|$$

- the key to learnability: $\mathbb{H}$ must be chosen properly.
Error Bounds in Machine Learning

- assume $\bar{f}$ is the unknown target function
- assume $f^*$ is the optimal ERM solution, i.e.
  $$f^* = \arg\min_{f \in \mathcal{H}} R_{\text{emp}}(f | \mathcal{D}_N)$$
- assume $\hat{f}$ denotes the best possible model in $\mathcal{H}$, i.e.
  $$\hat{f} = \arg\min_{f \in \mathcal{H}} R(f)$$
- we can define several types of errors in machine learning:
  - generalization error $E_g$:
    $$E_g = \left| R(f^*) - R_{\text{emp}}(f^* | \mathcal{D}_N) \right| \leq B_g(N, \mathcal{H})$$
  - estimation error $E_e$:
    $$E_e = \left| R(f^*) - R(\hat{f}) \right| \leq B_e(N, \mathcal{H})$$
  - approximation error $E_a$:
    $$E_a = \left| R(\hat{f}) - R(\bar{f}) \right| = R(\hat{f}) \leq B_a(N, \mathcal{H})$$
Generalization Bounds: Hoeffding’s inequality:

Given \( \{x_1, x_2, \cdots, x_N\} \) are \( N \) i.i.d. samples of a random variable \( X \) whose distribution function is given as \( p(x) \), and \( a \leq x_i \leq b \) for every \( i \), \( \forall \epsilon > 0 \), we have

- the weak law of large numbers:

\[
\lim_{N \to \infty} \Pr \left[ \left| \mathbb{E} [X] - \frac{1}{N} \sum_{i=1}^{N} x_i \right| > \epsilon \right] = 0
\]

- Hoeffding’s inequality (one of concentration inequalities):

\[
\Pr \left[ \left| \mathbb{E} [X] - \frac{1}{N} \sum_{i=1}^{N} x_i \right| > \epsilon \right] \leq 2e^{-\frac{2N\epsilon^2}{(b-a)^2}}
\]
Generalization Bounds: $\mathcal{B}_g(N, \mathbb{H})$

- for a fixed model $f$ (assuming the zero-one loss function):
  \[ \Pr \left[ \left| R(f) - R_{\text{emp}}(f|\mathcal{D}_N) \right| > \epsilon \right] \leq 2e^{-2N\epsilon^2} \]

- the above inequality does not apply to $f^*$ since it depends on $\mathcal{D}_N$: $\mathcal{D}_N \rightarrow f^*$

- how to extend to any model $f \in \mathbb{H}$?

- consider the uniform deviation:
  \[ \mathcal{B}_g(N, \mathbb{H}) = \sup_{f \in \mathbb{H}} \left| R(f) - R_{\text{emp}}(f|\mathcal{D}_N) \right| \]

- As $f^* \in \mathbb{H}$, we have $\left| R(f^*) - R_{\text{emp}}(f^*|\mathcal{D}_N) \right| \leq \mathcal{B}_g(N, \mathbb{H})$
Finite Model Space: $|\mathbb{H}|$

- finite model space $\mathbb{H}$ consists of $|\mathbb{H}|$ distinct models, $\forall \epsilon > 0$

$$B_g(N, \mathbb{H}) > \epsilon \iff \begin{cases} |R(f_1) - R_{\text{emp}}(f_1|\mathcal{D}_N)| > \epsilon \\
|R(f_2) - R_{\text{emp}}(f_2|\mathcal{D}_N)| > \epsilon \\
\vdots \\
|R(f_{|\mathbb{H}|}) - R_{\text{emp}}(f_{|\mathbb{H}|}|\mathcal{D}_N)| > \epsilon \end{cases}$$

- union bound:

$$\Pr\left(\bigcup_i A_i\right) \leq \sum_i \Pr(A_i)$$

$$\implies \Pr\left(B_g(N, \mathbb{H}) > \epsilon\right) \leq 2|\mathbb{H}|e^{-2N\epsilon^2}$$

$$\implies \Pr\left(B_g(N, \mathbb{H}) \leq \epsilon\right) \geq 1 - 2|\mathbb{H}|e^{-2N\epsilon^2}$$


Chapter 5
Generalization Bounds for Finite Model Space

- denote $\delta = 2|\mathcal{H}|e^{-2Ne^2}$, implying $\epsilon = \sqrt{\frac{\ln |\mathcal{H}| + \ln \frac{2}{\delta}}{2N}}$
- equivalently, we can say

$$ B_g(N, \mathcal{H}) \leq \sqrt{\frac{\ln |\mathcal{H}| + \ln \frac{2}{\delta}}{2N}} $$

holds at least in probability $1 - \delta$ ($\forall \delta \in (0, 1]$).
- As $f^* \in \mathcal{H}$, we have $|R(f^*) - R_{\text{emp}}(f^*|\mathcal{D}_N)| \leq B_g(N, \mathcal{H})$.
- the first generalization bound:

$$ R(f^*) \leq R_{\text{emp}}(f^*|\mathcal{D}_N) + \sqrt{\frac{\ln |\mathcal{H}| + \ln \frac{2}{\delta}}{2N}} $$

holds at least in probability $1 - \delta$.
- $B_g(N, \mathcal{H}) \sim O\left(\sqrt{\frac{\ln |\mathcal{H}|}{N}}\right)$
what about an infinite model space $\mathcal{H}$?

given a finite number of samples, not every model makes difference in terms of separating these samples

the number of *effective models*

**VC dimension** is introduced to count the total number of effective models in an infinite model space $\mathcal{H}$

*Figure*: a 2-D linear model space, where all models within each shaded area separate these samples in the same way
VC Dimension

- VC dimension is defined based on the concept of shattering a data set
- A data set is *shattered* by $H$ iff there exists at least a model in $H$ to generate every possible label combination of all data samples
- VC dimension of $H$: the maximum number of samples that can be shattered by $H$
- VC dimension of $H$ is $H$ $\Rightarrow$
  - $H$ can shatter at least one set of $H$ points (no need to shatter all sets of $H$ points)
  - $H$ cannot any set of $H + 1$ points
- VC dimension of linear models in $\mathbb{R}^n$ is $n + 1$

Figure: A set of 3 data points are *shattered* by $H$, consisting of all 2-D linear models
Generalization Bounds for Infinite Model Space

- if VC dimension of $\mathbb{H}$ is $H$, Vapnik-Chervonenkis (VC) theory suggests the total number of effective models in $\mathbb{H}$ for a set of $N$ points is upper-bounded by
  
  \[
  \begin{cases}
  2^N & \text{if } N < H \\
  \left(\frac{eN}{H}\right)^H & \text{if } N \geq H.
  \end{cases}
  \]

- the VC generalization bound for infinite model space $\mathbb{H}$:
  \[
  R(f^*) \leq R_{\text{emp}}(f^*|\mathcal{D}_N) + \sqrt{\frac{8H(\ln \frac{2N}{H} + 1)}{N} + 8 \ln \frac{4}{\delta} \frac{4}{N}}
  \]

  holds in probability $1 - \delta$ for any large data set ($N \geq H$).

- $\mathcal{B}_g(N, \mathbb{H}) \sim O\left(\sqrt{\frac{H}{N}}\right)$
An Example of VC Bounds

1. use \( N = 1000 \) data samples (input dimension is 100) to learn a linear classifier \((H = 101)\), the training error rate is 1\% and the test error rate is 2.4\%, set \( \delta = 0.001 \)

\[
R(f^*) \leq 0.01 + 1.8123 = 182.23\% \quad (\gg 2.4\%)
\]

2. same as above except \( N = 10000 \), the test error rate is 1.1\%.

\[
R(f^*) \leq 0.01 + 0.7174 = 72.74\% \quad (\gg 1.1\%)
\]

3. same as above except input dimension is 1000 \((H = 1001)\), the test error rate is 3.8\%.

\[
R(f^*) \leq 0.01 + 3.690 = 370.0\% \quad (\gg 3.8\%)
\]

caveat: VC bounds are extremely loose