Machine Learning: Scribe Note 4

Lecturer: Wang Liwei
He Shuncheng Yi Minzhen

October 16, 2015

1 VC Theory

1.1 Review

This lecture, we continued the topic of VC Theory. We perceive VC Dimension as a kind of **uniform convergence** over a set of indicator functions. Let us review the definition of VC Dimension.

Definition:

Say d is the VC-dim of Φ (Φ is a set of indicator functions), if $\exists z_1, z_2, \cdots, z_d$ such that

$$|(\phi(z_1), \cdots, \phi(z_d), \phi \in \Phi)| = 2^d$$

and there are **no** $z_1, \cdots, z_d, z_{d+1}$ such that

$$|(\phi(z_1), \cdots, \phi(z_{d+1}), \phi \in \Phi)| = 2^{d+1}$$

Using Chernoff Ineq., we can get (no proof here)

$$P(\sup_{\phi \in \Phi} |E\phi(z) - \frac{1}{n}\sum_{i=1}^{n} \phi(z)| \geq \varepsilon) \leq 4e^{-n\varepsilon^2/8}(\frac{en}{d})^d$$

Furthermore, we obtain a bound of $E\phi(z)$ by solving the inequality above. $\forall \delta > 0$, with probability $1 - \delta$ over the random draw of $z_1, z_2, \cdots, z_n$,

$$E\phi(z) \leq \frac{1}{n}\sum_{i=1}^{n} \phi(z_i) + \mathcal{O}\left(\sqrt{\frac{d\log(\frac{n}{d}) + \log(\frac{1}{\delta})}{n}}\right)$$

holds true simultaneously for all $\phi \in \Phi$.

1.2 Empirical Risk Minimization (ERM)

Given hypothesis space $\mathcal{H}$, find $f \in \mathcal{H}$ to minimize training error. This procedure is so-called ERM.

Theorem:

Let $\mathcal{H}$ be a hypothesis space $y = \{\pm 1\}$. Assume $VC(\mathcal{H}) = d$. For any learning problem (i.e., any underlying distribution D of the data), the classifier $\hat{f}$ returned by the ERM learning alg. satisfies with prob. $1-\delta$ over the random draw of a training set S of size n,

$$P_D(y \neq \hat{f}(x)) \leq P_S(y \neq \hat{f}(x)) + \mathcal{O}\left(\sqrt{\frac{d \log(\frac{n}{d}) + \log(\frac{1}{\delta})}{n}}\right)$$

Note that $P_D(y \neq \hat{f}(x))$ refers to the generalization error and $P_S(y \neq \hat{f}(x))$ refers to the training error. If we denote

$$f^* = \operatorname*{argmin}_{f \in \mathcal{H}} P_D(y \neq f(x))$$

we get

$$P_D(y \neq \hat{f}(x)) \leq P_D(y \neq f^*(x)) + \mathcal{O}\left(\sqrt{\frac{d \log(\frac{n}{d}) + \log(\frac{1}{\delta})}{n}}\right)$$

2 Practical Learning Algorithms

2.1 Linear Classifier

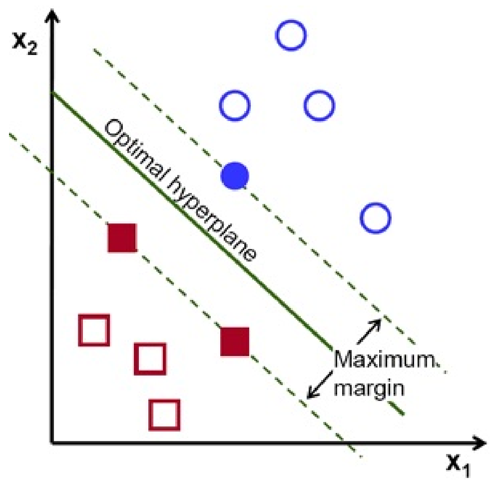
The problem of linear classification is described as follow:

Input: $x \in \mathcal{R}^d$
Output: $y = \{\pm 1\}$
Hypothesis Space: $\mathcal{H} = (w, b) | w \in \mathcal{R}^d, b \in \mathcal{R}$
Classifier: $f(x) = sgn(w^T x + b)$

It is trivial to infer that $VC(\mathcal{H}) = d + 1$ in this problem. The task to find a hyperplane is equivalent to this optimization problem:

$$\begin{aligned} \max_{w,b,t} \quad & t \\ s.t. \quad & y_i(w^T x_i + b) \geq t, \forall i \in [n] \\ & ||w|| = 1 \end{aligned}$$

A hyperplane that classifies all the training samples to the correct class exists when the solution of the optimization problem gives $t \geq 0$. In fact, it is a **large margin classifier** (see the figure below for the definition of margin).



Since we have algorithms to solve Linear Programming, Quadratic Programming, Convex Optimization and Semi-definite Programming, it is crucial to describe the problem in anothor way. The original programming is equivalent to a quadratic programming problem (its proof is left as exercise).

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2}||w||^2 \\ s.t. \quad & y_i(w^T x_i + b) \geq 1, \forall i \in [n] \end{aligned}$$

Before we introduce the algorithm of this problem, the knowledge of Duality Theory is required.

2.2 Minimax Theorem and Duality

2.2.1 Matrix Game

Pure strategy

Consider a matrix $M = \{(m_{ij}, \tilde{m}_{ij})\}_{s\times t}$ and two players, Row player and Column player. The Row player chooses one row first, the ith row for exapmle, and the Column player chooses one column (the jth column) seeing the Row player's move. Then, the Row player should pay m_{ij} to Column and Column should pay $\tilde{m}_{ij}$ to Row. If matrix $M = \{m_{ij}\}_{s\times t}$, it is a zero-sum game, which means that Row should pay m_{ij} to Column.

When Row takes the first move, they will reach

$$\min_i \max_j m_{ij}$$

When Row takes the second move, they will reach

$$\max_j \min_i m_{ij}$$

We have a conclusion that

$$\min_i \max_j m_{ij} \geq \max_j \min_i m_{ij}$$

Mixed strategy

Row player chooses a distribution p over $[s]$, and Column player, after seeing Row player's p, chooses a distribution q over $[t]$.

When Row takes the first move, they will reach

$$\min_p \max_q p^T M q$$

When Row takes the second move, they will reach

$$\max_q \min_p p^T M q$$

And we have a theorem (Von Neuman Minimax Theorem)

$$\min_p \max_q p^T M q = \max_q \min_p p^T M q$$

Theorem 1:
$\forall M = \{m_{ij}\}_{s \times t}$, $\exists p^*, q^*$, such that $\forall p, q$

$$p^{*T} M q \leq p^{*T} M q^* \leq p^T M q^*$$

Theorem 2:
Function $f(x, y)$, $\forall y$, $f(\cdot, y)$ is convex, and $\forall x$, $f(x, \cdot)$ is concave

$$\min_x \max_y f(x, y) = \max_y \min_x f(x, y)$$

and (x^*, y^*) is saddle point.

2.2.2 Lagrangian Duality

A primal problem:

Primal:

$$\begin{aligned} \min_x \quad & f(x) \\ s.t. \quad & g_i(x) \leq 0, \forall i \\ & h_j(x) = 0, \forall j \end{aligned}$$

f, g_i are convex fuctions and h_j are linear functions.

Proposition 1:

$$\textbf{Primal} \Leftrightarrow \min_x \max_{\lambda,\mu,\lambda\geq 0} f(x) + \sum_i \lambda_i g_i(x) + \sum_j \mu_j h_j(x)$$

Denote the target function with $L(x; \lambda, \mu)$

Proposition 2:

$$\textbf{Primal} \Leftrightarrow \max_{\lambda,\mu,\lambda\geq 0} f(\varphi) + \sum_i \lambda_i g_i(\varphi) + \sum_j \mu_j h_j(\varphi)$$

And

$$\left.\frac{\partial L}{\partial x}\right|_{x=x^*} \Rightarrow x^* = \varphi(\lambda, \mu)$$

---

3 Exercise

3.1 Ex 1

Give a proof of

$$VC(\mathcal{H}) = VC(\Phi)$$

3.2 Ex 2

Prove the equivalence of the two optimization problems

$$
\begin{array}{ll}
\max_{w,b,t} & t \\
s.t. & y_i(w^T x_i + b) \geq t, \forall i \in [n] \\
 & ||w|| = 1
\end{array}
$$

and

$$
\begin{array}{ll}
\min_{w,b} & \frac{1}{2}||w||^2 \\
s.t. & y_i(w^T x_i + b) \geq 1, \forall i \in [n]
\end{array}
$$

3.3 Ex 3

Give the dual problem of the latter one of **Ex 2**, namely the dual problem of

$$
\begin{array}{ll}
\min_{w,b} & \frac{1}{2}||w||^2 \\
s.t. & y_i(w^T x_i + b) \geq 1, \forall i \in [n]
\end{array}
$$