CA684 Jia Yuan Yu

4: VC theory

DCU March 7, 2014

These notes are based on the book "A Probabilistic Theory of Pattern Recognition."

Last week, we say a preview of VC theory. Recall:

$$\mathbb{P}\left(L_n(\phi^*) - \inf_{\phi \in \mathcal{C}} L(\phi) > 2\varepsilon\right) \le 2Ke^{-2n\varepsilon^2}.$$

What does this tell us?

- Something obvious: Picking the classifier with fewest empirical errors is the way to go.
- Something less obvious and more important: it gives guarantees.
- Given the number of samples that we have, what guarantees can we give in terms of discrepancy and confidence?
- Sample complexity: Given an discrepancy threshold and a confidence threshold, how many samples do we need to find a classifier satisfying these thresholds?

1 Four notions of error

We introduce four notions of error: empirical error frequency, error probability, estimation error, and approximation error.

Our training data is $\{X_1, Y_1, \ldots, X_n, Y_n\}$. Let $\phi : \mathbb{R}^d \to \{0, 1\}$. Let $\mathcal{C}$ denote a subset of all possible classifiers. We define the empirical error frequency (or empirical error probability, or empirical risk)

$$\hat{L}_n(\phi) = \frac{1}{n}\sum_{i=1}^{n} 1_{[\phi(X_i) \neq Y_i]}.$$

Let

$$\phi_n^* \in \arg\min \hat{L}_n(\phi).$$

This approach was developed by Vapnik and Chervonenkis theory in the 1970s.

Given $X_1, Y_1, \ldots, X_n, Y_n$, we can evaluate $\hat{L}_n(\phi_n^*)$. However, the true error probability is what we are interested in:

$$L_n(\phi_n^*) = \mathbb{P}(\phi_n^*(X) \neq Y \mid X_1, Y_1, \ldots, X_n, Y_n),$$

because we want to minimize the estimation error:

$$L_n(\phi_n^*) - \inf_{\phi\in\mathcal{C}} L(\phi),$$

where $L(\phi) = \mathbb{P}(\phi(X) \neq Y)$. This quantity is also related to the Bayes error:

$$L_n(\phi_n^*) - L^* = \left(L_n(\phi_n^*) - \inf_{\phi\in\mathcal{C}} L(\phi)\right) + \left(\inf_{\phi\in\mathcal{C}} L(\phi) - L^*\right),$$

where $(\inf_{\phi\in\mathcal{C}} L(\phi) - L^*)$ is called the approximation error. Observe that the large we make$\mathcal{C}$ is, the smaller the approximation error, but the large becomes the estimation error.

2 Shatter coefficient, VC dimension

Let $\mathcal{A}$ denote a collection of measureable sets. For a fixed vector $(z_1, \ldots, z_n) \in \mathbb{R}^{d\times n}$ of n points in $\mathbb{R}^d$, let $N_{\mathcal{A}}(z_1, \ldots, z_n)$ denote the number of different sets in

$$\Big\{\{z_1, \ldots, z_n\} \cap A \mid A \in \mathcal{C}\Big\}.$$

The n-th shatter coefficient of $\mathcal{A}$ is

$$s(\mathcal{A}, n) = \max_{(z_1,\ldots,z_n)\in\mathbb{R}^{d\times n}} N_{\mathcal{A}}(z_1, \ldots, z_n),$$

which is the maximal number of different subsets of n points that can be picked out by the collection of sets $\mathcal{A}$.

Clearly $s(\mathcal{A}, n) \leq 2^n$. If

$$N_{\mathcal{A}}(z_1, \ldots, z_n) = 2^n,$$

then we say that $\mathcal{A}$ shatters the points $z_1, \ldots, z_n$. Also, if there exists k such that $s(\mathcal{A}, k) < 2^k$, then $s(\mathcal{A}, n) < 2^n$ for all $n > k$ (homework exercise).

Let $\mathcal{A}$ be a collection of at least $|\mathcal{A}| \geq 2$ sets. The VC dimension $V_{\mathcal{A}}$ of $\mathcal{A}$ is the largest integer $k \geq 1$ such that $s(\mathcal{A}, k) = 2^k$, i.e.,

$$V_{\mathcal{A}} = \max\{k \in \mathbb{N}_+ \mid s(\mathcal{A}, k) = 2^k\}.$$

It measures the complexity, size, or expressive power, of the collection $\mathcal{A}$.

Example 2.1. If $\mathcal{A}$ is the collection of all halflines of the form $(-\infty, x]$ for $x \in \mathbb{R}$, then

$$s(\mathcal{A}, 2) = 3 < 2^2,$$

since if $z_1 < z_2$, then there is no set $(-\infty, x]$ that contains z_2, but not z_1. Hence, $V_{\mathcal{A}} = 1$.

Example 2.2. If $\mathcal{A}$ is the collection of all intervals $[x, y]$ in $\mathbb{R}^1$, then

$$s(\mathcal{A}, n) = \binom{n}{0} + \binom{n}{1} + \binom{n}{2} = \frac{n(n+1)}{2} + 1,$$

since every interval containing two points z_1 and z_3 contains a third point $z_2 \in [z_1, z_3]$. Hence, $V_{\mathcal{A}} = 2$.

Example 2.3. If $\mathcal{A}$ is the collection of halfspaces in $\mathbb{R}^d$ of the form $\{x : ax \geq b, a \in \mathbb{R}^d, b \in \mathbb{R}\}$. We have

$$S(\mathcal{A}, n) \leq 2 \sum_{i=0}^{d} \binom{n-1}{i},$$

(cf. Corollary 13.1 of PTPR). Hence, $V_{\mathcal{A}} = d + 1$.

2.1 From $\mathcal{A}$ to a collection of classifiers $\mathcal{C}$

Let $\mathcal{C}$ denote a collection of classifiers. Define a collection of sets

$$\begin{aligned} \mathcal{A}_{\mathcal{C}} &= \{V(\phi) \cup W(\phi) \mid \phi \in \mathcal{C}\} \subseteq \mathbb{R}^d \times \{0, 1\}, \\ V(\phi) &= \{x : \phi(x) = 1\} \times \{0\} \subseteq \mathbb{R}^d \times \{0, 1\}, \\ W(\phi) &= \{x : \phi(x) = 0\} \times \{1\}. \end{aligned}$$

The n-th shatter coefficient of $\mathcal{C}$ is

$$\mathcal{S}(\mathcal{C}, n) = s(\mathcal{A}_{\mathcal{C}}, n).$$

The VC dimension of $\mathcal{C}$ is

$$V_{\mathcal{C}} = V_{\mathcal{A}_{\mathcal{C}}}$$

We can now introduce the VC Theorem for classifier selection.

Theorem 2.1 (VC Theorem)**.** *For every distribution of the data and for all n, we have*

$$\mathbb{P}\left(L_n(\phi_n^*) - \inf_{\phi \in \mathcal{C}} L(\phi) > \varepsilon\right) \leq 8\, \mathcal{S}(\mathcal{C}, n)\, e^{-n\varepsilon^2/128}.$$

The VC Theorem bounds the estimation error of the classifier that minimizes the empirical error frequency among a collection of classifiers.

We are not done yet. The VC theorem is useful only when $\mathcal{S}(\mathcal{C}, n)$ is relatively small (i.e., sub-exponential). It turns out that if $\mathcal{C}$ has a finite VC dimension $V_{\mathcal{C}} > 2$, we have $\mathcal{S}(\mathcal{C}, n) \leq n^{V_{\mathcal{C}}}$.

Theorem 2.2 (Shatter coefficient)**.**

$$\mathcal{S}(\mathcal{A}, n) \leq \sum_{i=0}^{V_{\mathcal{A}}} \binom{n}{i}.$$

It follows by the binomial theorem that

$$\mathcal{S}(\mathcal{A}, n) \leq (1+n)^{V_{\mathcal{A}}}.$$

Hence, we have for all n:

- either $s(\mathcal{A}, n) = 2^n$ if $V_{\mathcal{A}}$ is infinite,
- or $S(\mathcal{A}, n) \leq (1+n)^{V_{\mathcal{A}}}$ if $V_{\mathcal{A}}$ is finite.

3 VC theory applied to neural networks

Let $\mathcal{C}_k$ denote the collection of neural network classifiers with one hidden layer of k hidden nodes, and an arbitrary sigmoid function σ.

First, the approximation error can be bounded by standard approximation arguments (cf. Theorem 30.4 of PTPR).

Theorem 3.1. *For every distribution of the data, we have*

$$\lim_{k\to\infty} \inf_{\phi\in\mathcal{C}_k} L(\phi) - L^* = 0.$$

An intuition for why neural nets have small approximation error comes from the fact that any multivariate continuous function can be represented as a sum of univariate functions.

Theorem 3.2 (Kolmogorov-Lorentz)**.** *Let $f : [0,1]^d \to \mathbb{R}$ be continuous. Let $x = (x^1, \ldots, x^d)$. There exist continuous univariate functions $\Phi : \mathbb{R} \to \mathbb{R}$ and $\{\psi_{j,\ell} : \mathbb{R} \to \mathbb{R}\}$ such that*

$$f(x^1, \ldots, x^d) = \sum_{j=0}^{2d} \Phi\left(\sum_{\ell=1}^{d} \psi_{j,\ell}(x^\ell)\right). \tag{1}$$

Moreover, the functions $\{\psi_{j,\ell}\}$ do not depend on f.

The estimation error is bounded by the VC Theorem along with the following shatter coefficient bound.

Theorem 3.3 (Lower bound, cf. Theorem 30.6 of PTPR)**.**

$$\mathcal{S}(\mathcal{C}_k, n) \leq (en)^{kd+2k+1}.$$