SYS 6016/4582: Machine Learning		**Spring 2017**
Lecture 10		
Instructor: Quanquan Gu	*Date: Feb 21th, 2017*	*Scriber: Ruixuan He*

Last time we introduced the projection of a function class, the growth function of a function class, and the upper bound of Rademacher complexity with respect to Growth function. And then I gave two examples about how to calculate the growth function of a threshold function and Interval. At the end of last lecture, I started to introduce the VC dimension. Let's first do some revision of these concepts.

Definition 1 (Projection of a Function Class) *The projection of function class $\mathcal{F}$ onto examples $\{Z_1, Z_2, ..., Z_n\}$ is defined as:*

$$\mathcal{F}_{|_{Z_{1:n}}} = \left\{ (f(Z_1), f(Z_2), ..., f(Z_n))^\top : \ f \in \mathcal{F} \right\}.$$

Note that the projection of a function class onto a training sample is a set, where each element in the set is an n-dimensional vector. Also, we care about the projection of a function class because provided that the projections of two different function classes are the same, the Rademacher Complexity of these two different function classes are the same. Therefore, such projection can be understood as the fundamental or the kernel which determines the complexity of a function class.

Definition 2 (Growth Function, or Shattering Coefficient) *Let $\mathcal{F}$ be a function class. The growth function of $\mathcal{F}$ is the maximum number of different patterns over n training examples:*

$$\Pi(\mathcal{F}, n) = \max_{Z_{1:n}} \left| \mathcal{F}_{|_{Z_{1:n}}} \right|.$$

Note that the projection of a function class depends on specific training sample while the growth function only depends on sample size n, since the dependency of sample is eliminated by taking the maximum over different training sample of size n.

Based on the Growth Function, we can get an upper bound of the Empirical Rademacher Complexity:

$$\hat{\mathcal{R}}_n(\mathcal{F}) \leq \sqrt{\frac{2M^2 \log(\Pi(\mathcal{F}, n))}{n}}. \tag{1}$$

Since the right hand side of (1) does not depend on training samples, taking the expectations on both sides yields an upper bound of Rademacher Complexity, which is the same as above:

$$\mathcal{R}_n(\mathcal{F}) \leq \sqrt{\frac{2M^2 \log(\Pi(\mathcal{F}, n))}{n}}.$$

We also introduced the well-known concept of VC-dimension in statistical learning theory.

Definition 3 (VC-dimension) *The VC-dimension of a binary function class $\mathcal{F}$ is defined as the largest d such that the growth function of $\mathcal{F}$ with d training examples equals to 2^d, i.e.:*

$$\mathrm{VC}(\mathcal{F}) = \sup_d \left\{ d : \Pi(\mathcal{F}, d) = 2^d \right\}.$$

Before introducing examples of VC-dimension, we need the definition of shattering.

Definition 4 (Shattering) *A set $\{X_1, X_2, ..., X_d\}$ is called shattered by the function class $\mathcal{F}$, if*

$$\left| \mathcal{F}_{|_{X_{1:n}}} \right| = 2^d.$$

Note that in binary classification setting, if there are d training examples, then there are in total 2^d kinds of configurations for the label assignments. When we project the function class $\mathcal{F}$ onto these d training examples, we obtain a set consisting of d-dimensional binary vectors. As long as the cardinality of the set equals to 2^d, i.e., for any arbitrary configuration of labels on the training examples, there exists a function in $\mathcal{F}$ that can correctly classify all examples, we can say these d training examples can be shattered by $\mathcal{F}$.

In terms of shattering, the VC-dimension of a function class is the **largest** number of training examples that can be shattered by the function class.

Note that the concept of shattering is specific to a given set of examples. In contrast, the definition of VC-dimension is not specific to a particular set.

Example 1 (Threshold Function) *The function class of threshold functions is defined as*

$$f: \mathbb{R} \to \{0, 1\}, \quad \mathcal{F} = \{ f(x) = \mathbb{1}\{x > t\}, \ t \in \mathbb{R} \}.$$

In the last lecture, we have shown that its growth function $\Pi(\mathcal{F}, n) = n + 1$. Therefore we can get an upper bound of the Rademacher Complexity of $\mathcal{F}$ as

$$\mathcal{R}_n(\mathcal{F}) \leq \sqrt{\frac{2\log(n+1)}{n}} \longrightarrow 0, \quad \text{as} \quad n \to \infty.$$

Moreover, we have $\mathrm{VC}(\mathcal{F}) = 1$. First we interpret it by the definition of VC-dimension:

- When $n = 1$, $\Pi(\mathcal{F}, n) = n + 1 = 2 = 2^1$.
- When $n = 2$, $\Pi(\mathcal{F}, n) = n + 1 = 3 < 2^2$.

Hence the largest n that makes $\Pi(\mathcal{F}, n) = 2^n$ hold is $n = 1$. Therefore we have $\mathrm{VC}(\mathcal{F}) = 1$. Secondly we explain it from the view of shattering, which is demonstrated as follows: when $n = 1$, the sample can be shattered by the function class, as is shown in Figure 1. However,

Figure 1: When $n = 1$, the sample can be shattered by the threshold function class.

when $n = 2$, no matter how the examples distribute, there is always one configuration of labels that cannot be correctly predicted by any function in $\mathcal{F}$, as is shown in Figure 2. Therefore, the largest number of examples that can be shattered by $\mathcal{F}$ is 1. Hence $\mathrm{VC}(\mathcal{F}) = 1$.

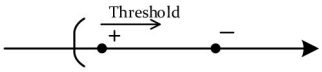

Figure 2: When $n = 2$, this configuration cannot be predicted by the threshold function class.

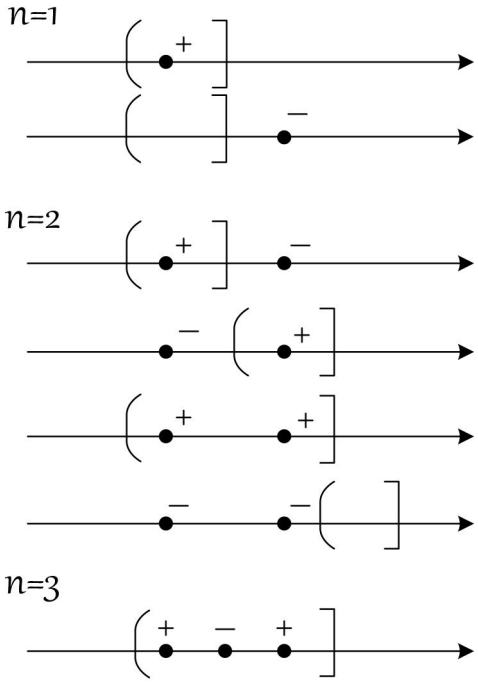

Figure 3: The interval function class can shatter 2 examples, but cannot shatter 3 examples.

Example 2 (Interval) *We define the function class of interval functions as*

$$\mathcal{F} = \{f,\ \mathbb{R} \to \{0,1\},\ f(X) = \mathbb{1}\{t_1 < X \leq t_2\},\ t_1, t_2 \in \mathbb{R}, t_1 \leq t_2\}.$$

In the last lecture, we have got its growth function as

$$\Pi(\mathcal{F}, n) = \frac{n(n+1)}{2} + 1.$$

Therefore we have an upper bound of the Rademacher Complexity

$$\mathcal{R}_n(\mathcal{F}) \leq \sqrt{\frac{2\log\left(n(n+1)/2+1\right)}{n}} = O\left(\sqrt{\frac{\log n}{n}}\right) \longrightarrow 0, \quad \text{as} \quad n \to \infty.$$

Its VC-dimension is $\mathrm{VC}(\mathcal{F}) = 2$. First we show it by the definition of VC-dimension:

- when $n = 1$, $\Pi(\mathcal{F}, n) = 1 \times (1+1)/2 + 1 = 2 = 2^1$.
- when $n = 2$, $\Pi(\mathcal{F}, n) = 2 \times (2+1)/2 + 1 = 4 = 2^2$.

- when $n = 3$, $\Pi(\mathcal{F}, n) = 3 \times (3+1)/2 + 1 = 7 < 2^3$.

Hence we have $\mathrm{VC}(\mathcal{F}) = 2$. Then we show this conclusion from the perspective of shattering. Figure 3 summarizes the analysis of whether $\mathcal{F}$ can shatter n examples when $n = 1, 2$ and 3.

As we can see from Figure 3, for $n = 1$ case, no matter the label of the example is positive or negative, we can always correctly classify it by choosing an interval which covers or does not cover it. For $n = 2$ case, no matter how the labels are configured, we can always choose an interval to correctly classify the examples, which means that the VC-dimension of $\mathcal{F}$ is at least 2. For $n = 3$ case, note that there is always one example lying between the other two examples. If the label of the example in the middle is "negative" while the labels of the examples on both sides are "positive", no interval can correctly classify all the examples.

Example 3 (Rectangle in the 2-d space) *The function class of rectangles is defined as*

$$\mathcal{F} = \left\{ f : \mathbb{R}^2 \to \{0,1\},\ f(X) = \mathbb{1}\left\{a_1 \le X_1 \le a_2, b_1 \le X_2 \le b_2\right\}, a_1 \le a_2, b_1 \le b_2 \right\}.$$

In other words, an example is classified as 1 *if it is inside or on the boundary of the rectangle surrounded by* $(a_1, b_1), (a_2, b_1), (a_2, b_2)$ *and* (a_1, b_2)*, and classified as* 0 *otherwise.*

A classifier in $\mathcal{F}$ is illustrated in Figure 4. In this case, let us directly calculate the VC-

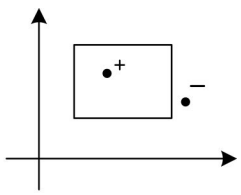

Figure 4: The illustration of a rectangular classifier.

dimension from the perspective of shattering. From Figure 5 it can be seen that: when $n = 2, 3$ or 4, there always exists a set of examples that can be shattered by $\mathcal{F}$. When $n = 5$, no matter how the examples are located, there is always a leftmost example, a rightmost example, an uppermost example and a downmost example. If a rectangle cover the four examples, it must cover the fifth example as well. Therefore, if we label the first four examples as "positive" and the fifth example "negative", no rectangle can correctly classify all of the examples. Hence the examples cannot be shattered by $\mathcal{F}$. Therefore we have $\mathrm{VC}(\mathcal{F}) = 4$.

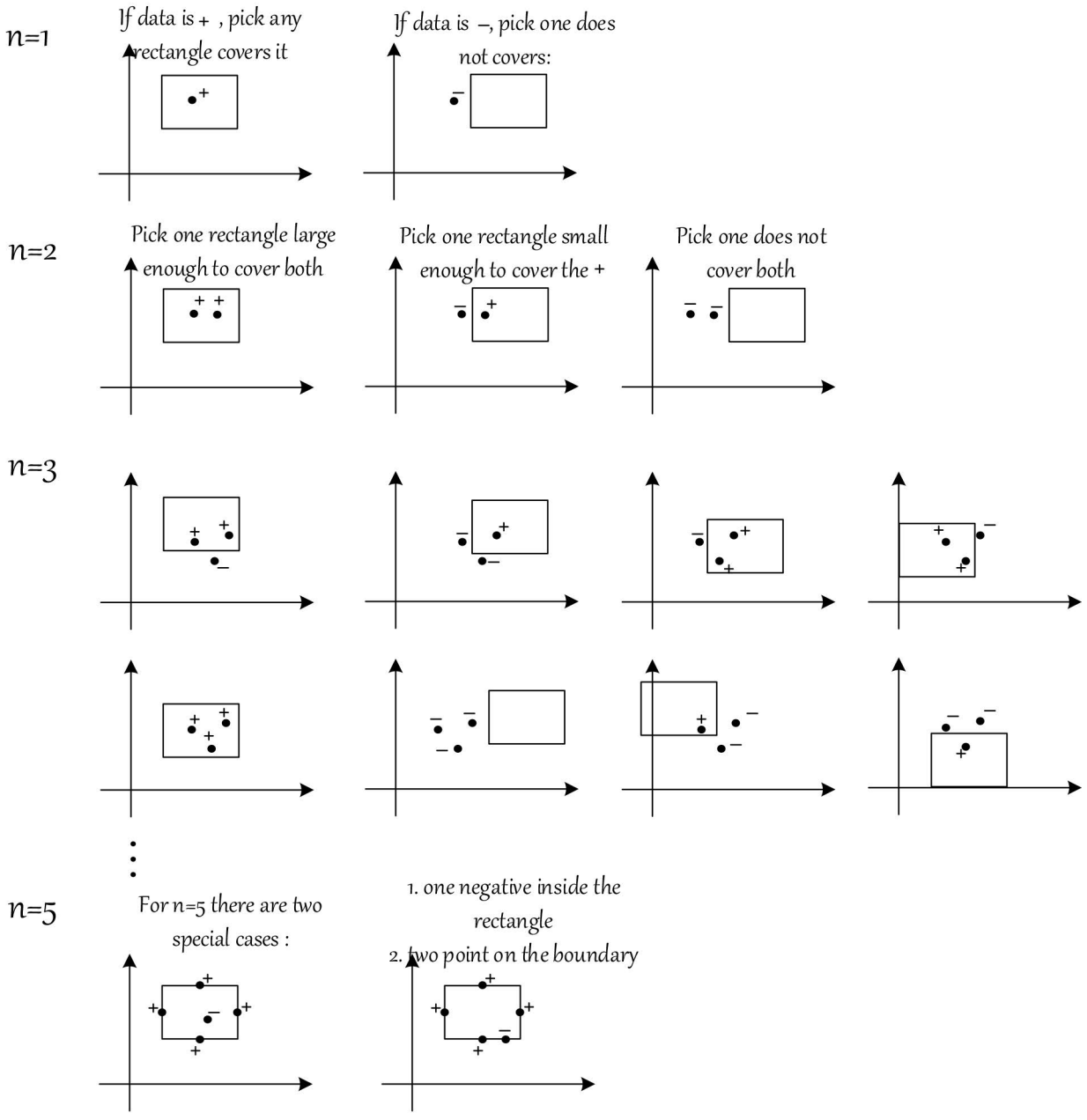

Figure 5: The function class of rectangles can shatter 2, 3 and 4 examples, but cannot shatter 5 examples.