Optimization for Machine Learning 50.579

Instructor: Ioannis Panageas Scribed by: Yanzhao Zou, Lu Xu

Lecture 8. Introduction to Statistical Learning Theory.

1 Introduction

Perceptron is a linear classifier or binary classifier, which is widely used in supervised learning to classify the given input data. The simplest perceptron is a single layer neural network, while multi-layers of perceptron are referred as neural network. Formally, the perceptron is defined as :

$$y = sign(\omega^T x - \theta), \tag{1}$$

where ω is the weight vector and θ is the threshold. And the goal is to compute a vector w that separates the two classes.

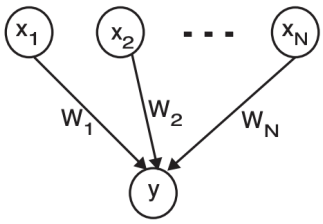

Figure 1: A simple perceptron.

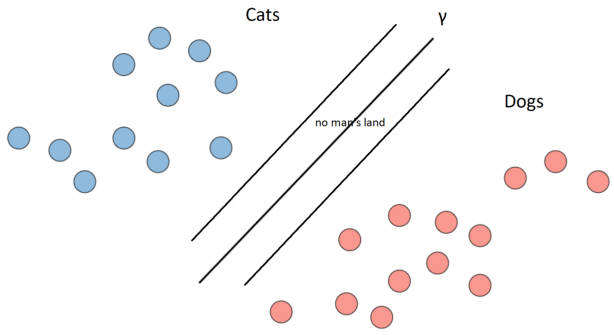

Figure 2: An example of Dogs and Cats classification.

1.1 The Perceptron Algorithm

Given $(x_1, y_1), ..., (x_T, y_T) \in X \times \{\pm 1\}$ where we assume $\| x \| = 1$ for all t. Formally γ is defined:

$$\gamma := max_{\omega:\|\omega\|=1} min_{i\in[T]}(y_i \omega^T x_i)_+, \tag{2}$$

where $(a)_+ = max(a, 0)$.

Consider the following iterative algorithm, where the goal is to iteratively update ω and optimize γ:

1. Initialize $\omega = 0$(hypothesis)
2. On round $t = 1...T$:

   Consider (x_t, y_t) and form prediction $\hat{y}_t = sign(\omega_t^T x_t)$.
   If $\hat{y}_t \neq y_t$:
   $\omega_{t+1} = \omega_t + y_t x_t$
   Else $\omega_{t+1} = \omega_t$.

1.2 Analysis of Perceptron

Theorem 1.1 *Perceptron makes at most $1/\gamma^2$ mistakes and corrections on any sequence with margin γ.*

Proof: Let m be the number of mistakes after T iterations. If a mistake is made at round t then

$$\| \omega_{t+1} \|_2^2 = \| \omega_t + y_t x_t \|_2^2 \tag{3}$$
$$\| \omega_{t+1} \|_2^2 = \| \omega_t \|_2^2 + \| x_t \|_2^2 + 2\, y_t\, x_t^T \omega_t (negative) \tag{4}$$
$$\| \omega_{t+1} \|_2^2 \leq \| \omega_t \|_2^2 + 1 \tag{5}$$

Since the update is only performed when there is a mistake, the total number of updates is equal to the number of mistakes made till step T, which is m in this case. When you sum equation 5 from 0 to m and cancel the same terms, we can get the below formula:

$$\| \omega_t \|_2^2 \leq m, \tag{6}$$

Consider a vector ω^* with margin γ, by definition of γ for all t that there is a mistake:

$$\gamma \leq y_t w^{*T} x_t = w^{*T}(w_{t+1} - w_t) \tag{7}$$

γ by definition is the the max min of $y_t w^{*T} x_t$, thus the $\leq$ relation holds. While by manipulating the iterative update step 2, we establish $y_t w^{*T} x_t = w^{*T}(w_{t+1} - w_t)$.
By adding equation 7 from 0 to m we also have that:

$$m\gamma \leq w^{*T}(w_T - w_1) = w^{*T} w_T, \tag{8}$$
$$= \| w_T \|_2 . \tag{9}$$
$$Therefore \quad m\gamma \leq \| w_T \|_2 \leq \sqrt{m} \tag{10}$$
$$Therefore \quad m \leq \frac{1}{\gamma^2}. \tag{11}$$

■

2 Random Data and 0-1 Loss Function

What we really showed is that given $(x_1, y_1), ..., (x_T, y_T) \in X \times \{\pm 1\}$, where we assume $\| x \| = 1$ for all t it holds:

$$\sum_{t=1}^{T} 1_{y_t \omega_t^T x_t \leq 0} \leq \frac{1}{\gamma^2} \tag{12}$$

Given $(x_1, y_1), ..., (x_T, y_T) \in X \times \{\pm 1\}$ IID from some distribution P. Run perceptron algorithm and consider $\omega_1, ...\omega_n$. Then choose ω.

Theorem 2.1 *IID Data: Let ω be the choice of the algorithm. It hold that:*

$$\mathbb{E}[\frac{1}{n} \sum_{i=1}^{n} 1_{y_i \omega^T x_i \leq 0}] \leq \frac{1}{n} \mathbb{E}[\frac{1}{\gamma^2}] \tag{13}$$

Proof: We have proved from before that (and taking expectation)

$$\mathbb{E}[\frac{1}{n} \sum_{i=1}^{n} 1_{y_i \omega^T x_i \leq 0}] \leq \mathbb{E}[\frac{1}{n\gamma^2}] \tag{14}$$

let $S = ((x_1, y_1), ..., (x_n, y_n))$. The LHS can be expressed as:

$$\mathbb{E}_\tau \, \mathbb{E}_S[1_{y_\tau \omega_\tau^T x_\tau \leq 0}] = \mathbb{E}_S \, \mathbb{E}_\tau[1_{y_\tau \omega_\tau^T x_\tau \leq 0}] \tag{15}$$

Observe now that ω_τ depdends only on $(x_1, y_1), ..., (x_{\tau-1}, y_{\tau-1})$, hence finally we can express the LHS in the form of a 0-1 loss function:

$$\mathbb{E}_S \, \mathbb{E}_\tau[1_{y_\tau \omega_\tau^T x_\tau \leq 0}] = \mathbb{E}_S \, \mathbb{E}_\tau \, \mathbb{E}_{(x,y)\sim P}[1_{y \omega_\tau^T x \leq 0}] = \mathbb{E}_S \, \mathbb{E}_\tau[L_{0-1}(\omega_\tau)] \tag{16}$$

where:

$$L_{0-1}(\omega) = \frac{1}{n} \sum_i 1_{y_i \omega^T x_i \leq 0}. \tag{17}$$

Note that if we keep iterating perceptron algorithm we finally get $L_{0-1}(\omega_\tau) = 0$, providing the two classes are linearly separable. ■

3 PAC Learning

Now that we have understood the definition of the algorithm and how it generalises to random data, let us talk about how we can evaluate the performance of the algorithm.
Assume we are given:

- Domain set $\mathcal{X}$, typically $\mathbb{R}^d$ or $\{0,1\}^d$. Think of 32x32 pixel images.
- Label set $\mathcal{Y}$, typically binary like $\{0,1\}$ or $\{-1,+1\}$
- A concept class $\mathcal{C} = \{h : h : \mathcal{X} \to \mathcal{Y}\}$

Given a learning problem, we analyse the performance of a learning algorithm:

- Training data $\mathcal{S} = (x_1, y_1), ..., (x_m, y_m)$, where samples $\mathcal{S}$ was generated by drawing m IID samples from the distribution $\mathcal{D}$.
- Output a hypothesis from a hypothesis class $\mathcal{H} = \{h : h : \mathcal{X} \to \mathcal{Y}\}$ of target functions.

We measure the performance through generalization error that is

$$err(h) = \mathbb{E}_{(x,y)\sim D}[l_{0-1}(h(x), y)]. \tag{18}$$

Definition 3.1 *(PAC learnable). We call a concept class $\mathcal{C}$ of target function is PAC learnable (w.r.t to $\mathcal{H}$) if there exists an algorithm A and function $m_C^A : (0,1)^2 \to \mathbb{N}$ with the following property:*

Assume $\mathcal{S} = ((x_1, y_1), ..., (x_m, y_m))$ is a smaple of IID examples generated by some arbitrary distribution D such that $y_i = h(x_i)$ for some $h \in \mathcal{C}$ almost surely. If $\mathcal{S}$ is the input of A and $m > m_C^A$ then the algorithm returns a hypothesis $h_s \in \mathcal{H}$ such that, with probability $1 - \delta$ (over the choice of the m training examples):

$$err(h_s) < \epsilon \tag{19}$$

The function m_C^A is referred to as the sample complexity of algorithm A.

To help us understand the definition of concept class, here we list two concrete examples:
Example: (Axis Aligned Rectangles). The first example of a hypothesis class will be of rectangles aligned to the axis. Here we take the domain $\mathcal{X} = \mathbb{R}^2$ and we let $\mathcal{C}$ include be defined by all rectangles that are aligned to the axis. Namely for every (z_1, z_2, z_3, z_4) consider the following function over the pane

$$f_{z_1,z_2,z_3,z_4}(x_1, x_2) = \begin{cases} 1 & z_1 \le x_1 \le z_2, z_3 \le x_2 \le z_4 \\ 0 & else \end{cases} \tag{20}$$

Then $\mathcal{C} = \{f_{z_1,z_2,z_3,z_4} : (z_1, z_2, z_3, z_4) \in \mathbb{R}^4\}$.

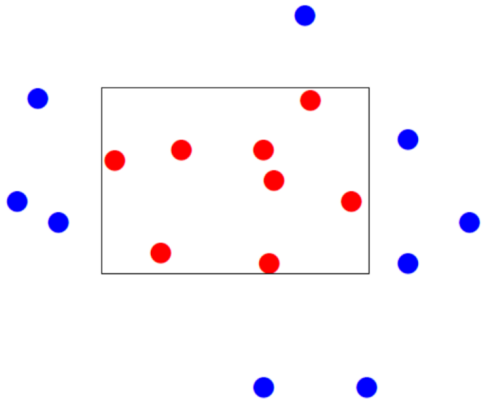

Figure 3: Concept Class of Axis Aligned Rectangles.

Example: (Half-space). A second example that is of some importance is defined by hyperplane. Here we let the domain be $\mathcal{X} = \mathbb{R}^d$ for some integer d. For every $w \in \mathbb{R}^d$, induces a half space by consider all elements $\mathcal{X}$ such that $w \cdot x \geq 0$. Thus, we may consider the class of target functions described as follows:

$$\mathcal{C} = \{f_w : w \in \mathbb{R}^d, f_w(x) = sign(w \cdot x)\} \tag{21}$$

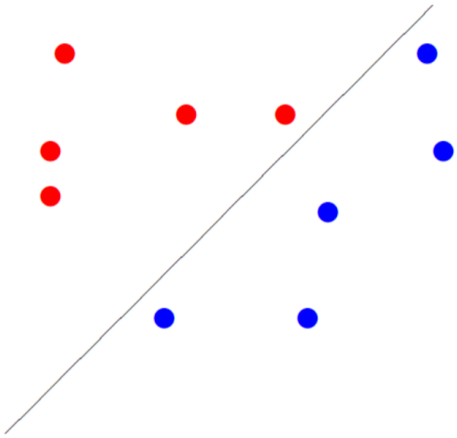

Figure 4: Concept Class of Half-Space.

4 ERM Algorithm

Now even if a concept class is PAC learnable, there might exist multiple hypothesis classes that meet the requirement. For the interest of optimization, our real focus is to find the optimal hypothesis class that gives us the minimum error given all the conditions. And that is where we Empirical Risk

Minimization (ERM) algorithm comes into play. ERM algorithm is defined as follows: Return:

$$arg\ min_{h\in\mathcal{H}}\ err_s(h), \tag{22}$$

where $err_s(h) = \frac{1}{m}\sum l_{0-1}(h(x_i), y_i)$.

Luckily, we have some nice guarantees when the concept class is finite.

Theorem 4.1 *Finite classes are PAC learnable: Consider a finite class of target function $\mathcal{H} = h_1, ..., h_t$ over a domain. Then if size of sample S is $m > \frac{2}{\epsilon^2}\log\frac{2|\mathcal{H}|}{\delta}$ then with probability $1-\delta$ we have that:*

$$max_{h\in\mathcal{H}} \mid err_S(h) - err(h) \mid < \epsilon. \tag{23}$$

Proof: Applying Hoeffding's inequality we obtain that for every S and fixed h, since $err_S(h)$ is sum of IID bernoulli with expectation $err(h)$:

$$Pr_S[\mid err_S(h) - err(h) \mid > \epsilon] \leq 2e^{-2m\epsilon^2} \tag{24}$$

Applying union bound we obtain that:

$$Pr_S[\exists h :\mid err_S(h) - err(h) \mid > \epsilon] \leq 2 \mid \mathcal{H} \mid e^{-2m\epsilon^2} \tag{25}$$

We want the RHS to be less than δ. We can achieve that With the appropriate choice of m. ■

5 VC Dimension

Now that we have see the neat result guarantee when the concept class is finite, what happens when the concept class is infinite? Does the guarantee still hold true? Let's first take a look at a motivating example.

Lemma 5.1 *Threshold Function: Consider the Hypothesis class of threshold function on the real line, that is:*

$$\mathcal{H} = h_a : a \in \mathbb{R}, \tag{26}$$

where $h_a(x) = 1_{x<a}$.$\mathcal{H}$ is PAC learnable using ERM algorithm (even if the class is infinite).

Remarks:

- Therefore, it is not necessary that the hypothesis class is of finite cardinality.
- We will show the lemma above , i.e., (ϵ, δ) - learnable using $\frac{\log\frac{2}{\delta}}{\epsilon}$ samples.

Proof: Let D be the marginal distribution over the domain and fix ϵ, δ. We need to show that taking S samples IID of size $\frac{\log\frac{2}{\delta}}{\epsilon}$ suffices so that with probability $(1-\delta$ the generalization error is at most ϵ.
Let a^* be a number such that h_{a^*} has error zero (perfect fit). Moreover, consider $a_0 < a^* < a_1$ such that:

$$Pr_{x\sim D}[x \in (a_0, a^*)] = Pr_{x\sim D}[x \in (a^*, a_1)] = \epsilon. \tag{27}$$

Observe that we might have to choose $a_0 = -\infty$ or $a_1 = +\infty$.

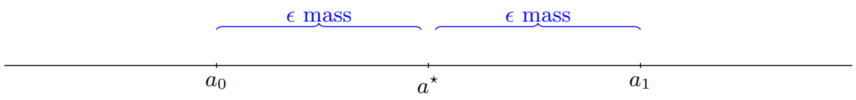

Figure 5: Concept Class of Axis Aligned Rectangles.

Let S be a set of IID samples and assume that the ERM algorithm returns a function h_S with threshold b_S.
If b_0 is the maximum x with label 1 and b_1 the minimum x with label 0 it holds that $b_S \in (b_0, b_1]$.
The error of $h_S$is at most ϵ if and only if $(b_0, b_1] \subseteq (a_0, a_1)$.
Let 's bound the probability of this event. By union bound we have:

$$Pr_{S\sim D^m}[(b_0 < a_0) \cup (b_1 > a_1)] \leq Pr_{S\sim D^m}[(b_0 < a_0)] + Pr_{S\sim D^m}[(b_1 > a_1)] \tag{28}$$

$$Pr_{S\sim D^m}[(b_0 < a_0)] \leq Pr_S[\forall x \in S, x \notin (a_0, a^*)] = (1-\epsilon)^m \leq \epsilon^{-\epsilon m} \tag{29}$$

$$Pr_{S\sim D^m}[(b_1 > a_1)] \leq Pr_S[\forall x \in S, x \notin (a^*, a_1)] = (1-\epsilon)^m \leq \epsilon^{-\epsilon m} \tag{30}$$

By adding equation 29 and 30, we conclude that the error probability is $2\epsilon^{-\epsilon m} = \delta$. Solving for m we get:

$$m = \frac{\log(\frac{2}{\delta})}{\epsilon}. \tag{31}$$

However, note that not all hypothesis classes are learnable. With the help of VC dimension in the next section, we can get more defined conditions for learnable and unlearnable cases.

■

5.1 Definition

Definition 5.1 *(Restriction). Let $\mathcal{H}$ be a class of functions from $\mathcal{X}$ to {0,1} and let $C = c_1, ...c_m$. The restriction of $\mathcal{H}$ to C is the set of functions from C to {0,1} that can be derived from $\mathcal{H}$. That is*

$$\mathcal{H}_C = \{h(c_1), ..., h(c_m)) : h \in \mathcal{H}\}, \tag{32}$$

where we represent each function from C to {0,1} as a vector in $\{0,1\}^{|C|}$

Definition 5.2 *(Shattering). A hypothesis class $\mathcal{H}$ shatters a finite set $C \subset \mathcal{X}$ if the restriction of $\mathcal{H}$ to C is the set of all functions from C to $\{0,1\}$. That is $\mid \mathcal{H} \mid = 2^{|C|}$.*

Definition 5.3 *(VC dimension). The VC-dimension hypothesis class $\mathcal{H}$, denote VCdim($\mathcal{H}$),is the maximal size of a set C that can be shattered by $\mathcal{H}$. If $\mathcal{H}$ can shatter sets of arbitrarily large size we say that $\mathcal{H}$ has infinite VC-dimension.*

Example: Let's see some examples and their intuitions:

- The class of threshold functions on real lines has VC dimension 1.
  Ans: Any point that lies on the same side of the threshold cannot be labelled as 0 and 1 at the same time.

- The class of interval functions on real line has VC dimension of 2.
  Ans: Any point that lies between two given points with the same label, cannot take the other label value.

- The class of aligned rectangle functions on the plane has VC dimension 4.
  Ans: Refering to Figure 6, any axis aligned rectangle cannot label c_5 by 0 and the rest of the points by 1.

- Any infinite class $\mathcal{H}$ hs VC dimension of at most log $\mid \mathcal{H} \mid$.
  Ans: Because by definition of shattering, $\mid \mathcal{H} \mid = 2^{|C|}$

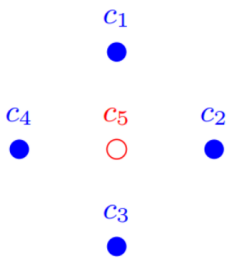

Figure 6: example of VC dimension of 4

5.2 VC Dimension of Halfspaces

Theorem 5.2 *(Halfspaces). The VC dimension of the class $\mathcal{H}$ of homogenous halfspaces in $\mathbb{R}^d$. Note that $\mathcal{H} = \{h_w(x) : h_w(x) := sign(w^T x)\}$.*

Proof: We first need to show that VC dimension is at least d by appropriately choosing a set C. Consider the set of vector $e_1, ..e_d$,where for every i the vector e_i is the all zeros vector except 1 in the i-th coordinate. This set is shattered by the class of homogenous halfspaces because for every binary vector $y_1, ...y_d$ and for $w = (y_1, ...y_d)$, we get that $h_w(e_i) = y_i$.
Next we need to show that VC dimension is less than d+1. Let $x_1, ...x_{d+1}$ be a set of d+1 vectors in $\mathbb{R}^d$.Then, there must exist real numbers $a_1, ..., a_{d+1}$, not all of them are zero, such that:

$$\sum a_i x_i = 0 \text{ linearly dependent.} \tag{33}$$

$$\text{Let } I = \{i : a_i > 0\} \text{ and } J = \{j : a_j < 0\} \tag{34}$$

If both I, J are non-empty then

$$\sum_{i \in I} a_i x_i = \sum_{j \in J} \mid a_j \mid x_j. \tag{35}$$

If $x_1, ..., x_{d+1}$ are shattered then there exists a ω such that $\omega^T x_i > 0$ for $i \in I$ and $\omega^T x_j < 0$ for $j \in J$. If the above is true, we get that:

$$0 < \sum_{i \in I} a_i \omega^T x_i = \omega^T \sum_{i \in I} a_i x_i \tag{36}$$

$$= \omega^T \sum_{j \in J} \mid a_j \mid x_j \tag{37}$$

$$= \sum_{j \in J} \mid a_j \mid \omega^T x_j < 0, \tag{38}$$

which is a contradiction. Thus we can conclude that VC dimension is less than d+1. And the theorem is thereafter proved. ■

5.3 Example of Infinite VC

As mentioned earlier before introducing the formal definition of VC dimension, it helps us define what hypothesis classes are not PAC learnable. Let's see an example of that.

Theorem 5.3 *(since has infinite VC). Consider the real line and let*

$$\mathcal{H} = \{x \to \lceil sin(\theta x) \rceil : \theta \in \mathbb{R}\}. \tag{39}$$

The VC dimension of the hypothesis class above is infinite.

Proof: We need to show that for for every d one can find d points that are shattered by $\mathcal{H}$. consider $x \in (0, 1)$ and let $0.x_1x_2x_3...$, be the binary expansion of x. Then for any natural number m, $\lceil sin(2^m \pi x) \rceil = 1 - x_m$, provided that there exists a $k \geq m$ such that $x_k = 1$.
Fix d and consider $C = \{1/2, 1/4, ..., 1/2^d\}$ and moreover choose any binary vector of labels $(y_1, ..., y_d)$. Set $x = 0.y_1...y_d1$ and use the above.
Intuitively, the sign function of sine function, is a square wave function with amplitude 1 and period given by $2\pi costan^{-1}(\theta)$. Thus by changing the value of θ, the square wave frequency can be manipulated to produce any labeling for a given set of points. Thus, its VC dimension is infinite. ■

5.4 The Importance of VC Dimension

Theorem 5.4 *Fundamental Theorem of Learnability: The following are equivalent:*

- $\mathcal{H}$ *is PAC learnable.*
- *Any ERM rule is a successful PAC learner for* $\mathcal{H}$*.*
- $\mathcal{H}$ *has finite VC dimension.*

Remarks: The number of samples needed is $O(\frac{d \log \frac{1}{\epsilon} + \log \frac{1}{\delta}}{\epsilon})$, where d is the VC dimension of the hypothesis class.