CS 560: Statistical Machine Learning

Instructor: Jie Shen | Lecture 4
Scribe: Humphrey De Guzman, Daniyal Mufti and Karthik Karanam | Date: 2/14/2022

1 Representative Sets

1.1 Introduction

In the previous lecture we talked about PAC learning and in our example reference some oracle who could give us some samples of data from some distribution D. What exactly is said distribution? If D is supposed to be some all-encompassing distribution that is the *"true"* statistical representation of all real world cases, how do we measure this empirically? This is where representative sets come into play.

1.2 Mathematical Definition

We define a set S ϵ-representative if the following conditions are true

$$\text{Rep}_D(H, S) \triangleq \ \sup_{\forall h \in H} |\text{Cost}_D(h) - \text{Cost}_S(h)| \leq \epsilon$$

Where $\text{Rep}_D(H, S)$ is our representative set with respect to distribution D and H, S refers to our universe of hypothesis classes and empirical set respectively. $\text{Cost}_X(h)$ refers to the average loss of all training samples from set $S = (z_1, ..., z_m) \in \mathbb{R}^m$ that come from distribution X with respect to hypothesis h.

$$\text{Cost}_X(h) \triangleq \frac{1}{m} \sum_{i=1}^{m} l(h, z_i)$$

Note that we use $\text{Cost}_D(h)$ in our notation here but it is sometimes denoted as $L_D(h)$. To avoid confusion with the Lagrange dual equation [who also is denoted as $L_D(h)$] and be more granular with our example, we use $\text{Cost}_D(h)$.

1.3 What does this all mean?

We use ϵ-representative sets as a way to ensure our empirical data sets are quantitatively close to whatever real world distribution we are trying to encompass in our training data. By ensuring that our two sets of data have very similar costs/loss values, we can be comfortable with the notion that our set S can be used to accurately portray the real phenomenon at play, hence the name *representative sets*. However there is one glaring issue, how can we calculate the cost function of the real distribution D, after all it is unfeasible to literally collect all of the data in the world for whatever we are trying to model. This is where we can make use of another concept called Rademacher Complexity to solve our problem.

2 Rademacher Complexity

2.1 Introduction

The Rademacher complexity takes the expectation of two disjoint subsets S_1 and S_2 which represent the spitting out of our initial set S where S_1 represents the validation set and the S_2 represents the training set. Formally defined, we consider $F \circ S$ be a set of all possible evaluations of $f \in F$ achieved on a sample S(mathematical definition below) and let σ be a i.i.d distributed with P[σ_i = -1] = P[σ_i = 1] = 0.5. Then the Rademacher complexity is defined as stated below(mathematical definition below). Finally we can express a lemma that can bound the representativeness of S two times the expected Rademacher complexity(mathematical definition below).

2.2 Mathematical Definition

This union S = $S_1 \cup S_2$ can be represented as:

$$LS_1(h) - LS_2(h)$$

$$= \sum_{S_1} l(h_1 z_i) - \sum_{S_2} l(h_1 z_j)$$

$$= \sum_{S} \sigma_i l(h_1 z_i)$$

$$\sigma_i = \{+1\ Z_i E S_1\ w.p.\ 0.5,\ -1\ Z_i E S_2\ w.p.\ 0.5\}$$

Rademacher complexity is:

$$R(F \circ S) \overset{def}{=} \frac{1}{m} \underset{\sigma^2\{\underset{-}{+}\}^m}{E} [\underset{f \in F}{sup} \sum_{i=1}^{m} \sigma_i f(z_i)]$$

Representativeness bounding lemma is:

$$\underset{S \sim D^m}{E}[Rep_D(F,S)] \ \leq \ 2 \underset{S \sim D^m}{E}[R(F \circ S)]$$

3 McDiarmid's Inequality

3.1 Introduction

From Rademacher Complexity we have discussed some features about generalization error and how many samples we need to draw to learn error rate ϵ. McDiarmid's Inequality helps us to develop generalization bounds that depend upon the number of samples m(where m depended on ϵ and δ).this inequality tells Let us take V be some set and and have a function f which maps from V to real value i.e., *f:* $V^m \rightarrow R$ be a function of m variable such that for some $c >$ *0*, $\forall\ i \in [m]$ and $\forall\ x_1$,x_2, $\ldots$, x_m , $x_i^{'} \in V$ we have

$$f(x_1, \ldots, x_m) - f(x_1 \ldots x_{i-1}, x_i, x_{i+1}^{'}, \ldots, x_m| \leq c.$$

i.e. McDiarmid's Inequality tells us changing one input to the function f does not change its value by much.

Let $X_1, \ldots, X_m$ be m independent random variables taking values in V. Then, with probability of at least $1 - \delta$ we have

$$|f(X_1, \ldots, X_m) - E|f(X_1, \ldots, X_m|| \leq c\sqrt{ln\tfrac{2}{\delta}m/2}$$

On the basis of McDiarmid inequality we can derive generalization bounds with a better dependence on the confidence parameter.

3.2 Theorem

Assume that for all z and h ∈H we have that $|l(h, z) \leq c|$.
Then,

1. With probability of at least $1 - \delta$, for all h ∈ H,

   $$L_D(h) - L_S(h) \leq 2 \underset{S^i \approx D^m}{E} R(l \circ H \circ S^{'}) + c\sqrt{\frac{2ln(2/\delta)}{m}}$$

   The above equation directly follows from McDiarmid's Inequality, because the function L has the property it is a summation of a lot of l. if we change one l on the single example but the absolute value of the function L does not change too much and we guaranteed that the function value is concentrated around its expectation.

   $$\text{LHS: } L_D(h) - L_S(h) = Rep_D(F, S)$$

   $$Rep_D(F, S) \leq E[Rep_D(F, S)] + c\sqrt{\frac{2ln(2/\delta)}{m}}$$

   $$\text{But, } \underset{S^i \approx D^m}{E} Rep_D(F, S) \leq 2 \underset{S^i \approx D^m}{E} R(l \circ H \circ S)$$

   From above equations we get:

   $$L_D(h) - L_S(h) \leq 2 \underset{S^i \approx D^m}{E} R(l \circ H \circ S^{'}) + c\sqrt{\frac{2ln(2/\delta)}{m}}$$

2. With probability of at least $1 - \delta$, for all h ∈ H,

   $$L_D(h) - L_S(h) \leq 2R(l \circ H \circ S) + 4c\sqrt{\frac{2ln(2/\delta)}{m}}$$

   If we measure $L_S(h)$ then it will not differs from the true error rates i.e. true lose function $L_D(h)$ too much because the difference between them is bounded by Rademacher Complexity and generalization term which depends on c, δ and m

   the above inequality we note that the random variable $R(l \circ H \circ S^{'})$ also satisfies the bounded differences condition of McDiarmid's Inequality with a constant 2c/m. Therefore, the second inequality follows from the first inequality, McDiarmid's Inequality, and the union bound.

   keynotes: if RHS is small then LHS should be very small. So this inequality tells us if the Rademacher Complexity of the sample sets is small then it means it generalizes well
   In particular, this holds for $h = ERM_H(S)$.

3. For any h* with probability of at least $1-\delta$,

$$L_D(h) - L_D(h^*) \leq 2R(l \circ H \circ S) + 5c\sqrt{\frac{2ln(2/\delta)}{m}}$$

we can re-write above equation as

$$(L_D(h) - L_S(h)) + (L_S(h) - L_S(h^*)) + (L_S(h^*) - L_D(h^*))$$

But, we know $(L_S(h) - L_S(h^*)) \leq 0$

Therefore, $L_D(h) - L_D(h^*) \leq L_D(h) - L_S(h)) + (L_S(h^*) - L_D(h^*)$

From 2nd Inequality we got: $L_D(h) - L_S(h) \leq 2R(l \circ H \circ S) + 4c\sqrt{\frac{2ln(2/\delta)}{m}} \rightarrow 1$

By using Hoeffding's inequality we obtain: $(L_S(h^*) - L_D(h^*) \leq c\sqrt{\frac{2ln(2/\delta)}{m}} \rightarrow 2$

from adding 1 and 2 we got: $L_D(h) - L_D(h^*) \leq 2R(l \circ H \circ S) + 5c\sqrt{\frac{2ln(2/\delta)}{m}}$

4 Rademacher Complexity of Linear Classes

4.1 Introduction

In this part we analysis the Rademacher complexity of linear classes. We start with defining a lemma which bounds the Rademacher complexity of a class H and then use provide the proof using the Cauchy-Schwartz inequality and then using the Jensen's inequality. The mathematical definitions are provided below.

4.2 Mathematical Definition

Start with defining a class:

$$\underset{\triangle}{H} = \{x \implies < w, x >: \|w\|_2 \leq B\}$$

Then let:

$$S = \{x_1, ..., x_m\}$$

and:

$$H \circ S = \{< w, x_1 >, ... < w, x_m >\}$$

then Lemma is defined as:

$$R(H \circ S) \leq \frac{Bmax_i \|x_i\|_2}{\sqrt{m}}$$

Proof:
First using Cauchy-Schwartz inequality we have:

$$mR(H \circ S) = \underset{\sigma}{E}[\underset{\mathbf{a} \in H \circ S}{sup} \sum_{i=1}^{m} \sigma_i a_i]$$

$$= \underset{\sigma}{E}[\underset{\mathbf{a}\in H\circ S}{sup}\overset{m}{\underset{i=1}{\Sigma}}\sigma_i < w, xi >]$$

$$= \underset{\sigma}{E}[sup < w, \overset{m}{\underset{i=1}{\Sigma}}\sigma_i x_i >]$$

$$\le \underset{\sigma}{E}[\underset{\|w\|_2\le B}{sup}\|w\|_2 * \|\overset{m}{\underset{i=1}{\Sigma}}\sigma_i x_i\|_2]$$

$$= \underset{\|w\|\|_2\le B}{sup}\|w\|_2 * \underset{\sigma}{E}[\|\overset{m}{\underset{i=1}{\Sigma}}\sigma_i x_i\|_2]$$

Now using Jensen's inequality we can deduce:

$$= \underset{\sigma}{E}[(\|\overset{m}{\underset{i=1}{\Sigma}}\sigma_i x_i\|_2^2)^{\frac{1}{2}}] \le (\underset{\sigma}{E}[\|\overset{m}{\underset{i=1}{\Sigma}}\sigma_i x_i\|_2^2])^{\frac{1}{2}}$$

$$= \underset{\sigma}{E}[\|\overset{m}{\underset{i=1}{\Sigma}}\sigma_i x_i\|_2^2] = \underset{\sigma}{E}[\overset{m}{\underset{i=1}{\Sigma}}\sigma_i x_i * \overset{m}{\underset{i=1}{\Sigma}}\sigma_i x_i]$$

$$= \underset{\sigma}{E}[\underset{i,j}{\Sigma}\sigma_i\sigma_j(x_i, x_j)]$$

$$= \underset{i\neq j}{\Sigma} < x_i, x_j > \underset{\sigma}{E}[\sigma_i\sigma_j] + \overset{m}{\underset{i=1}{\Sigma}} < x_i, x_i > \underset{\sigma}{E}[\sigma_i^2]$$

$$= \overset{m}{\underset{i=1}{\Sigma}}\|x_i\|_2^2 \le m\ \underset{i}{max}\|x_i\|_2^2$$

5 Lipschitz Continuity

5.1 Introduction

Lipschitz continuity is a fairly basic concept with the basis that pertains to the boundedness of some function f at any of its given points. Specifically the idea is that for any given point in our function f, the derivative of said function at all its points are less than or equal to some value, typically denoted as ρ. Visually speaking, it is possible to move a double cone through all of the points in f and only the current point the double cone is centered around will be inside of it. Of course the real specification of Lipschitz continuity is more specific than this, so let us see the real mathematical definition to get a deeper understanding.

5.2 Mathematical Definition

Let us consider some space $C \in \mathbb{R}^d$ and a function $f : \mathbb{R}^d \mapsto \mathbb{R}^k$. We can consider said function ρ-Lipschitz continuous in our space C if for every $\boldsymbol{w}_1, \boldsymbol{w}_2 \in C$ we have the following

$$||f(\boldsymbol{w}_1) - f(\boldsymbol{w}_2)|| \le \rho||\boldsymbol{w}_1 - \boldsymbol{w}_2||$$

Alternatively if our function f maps $\mathbb{R} \mapsto \mathbb{R}$ we can also find the following from the mean value theorem. If we consider some u between w_1 and w_2 then we can find that the following equality is true:

$$f(w_1) - f(w_2) = f'(u)(w_1 - w_2)$$

5.3 Examples

As seen from the last statement, it is easy to see that all linear functions are globally Lipschitz continuous, after all the derivative of all linear functions are constant. If our function f takes the form $f(x) = \boldsymbol{w}x + b$ the inequality looks like $|\langle \boldsymbol{w}, \boldsymbol{w_1} - \boldsymbol{w_2} \rangle| \leq ||\boldsymbol{w}||\,||\boldsymbol{w_1} - \boldsymbol{w_2}||$. Though is our function is equal to something like $f(x) = x^2$ we can see that this is not Lipschitz continuous at a global scale. However, we must remember that Lipschitz continuity specifies a space C! This means that the function $f(x) = x^2$ is Lipschitz continuous as long as our $\rho/2 \leq |x|$ as $f'(x) = 2x = \rho$

6 Contraction Lemma

6.1 Introduction

Now that we have understood Lipschitz continuity and Rademacher complexity, we can make use of both to formulate another useful lemma for us. Let us consider some set A such that $A = \{\boldsymbol{a_1}, ..., \boldsymbol{a_m}\}$, then if we take the composite of A with any Lipschitz continuous functions, according to the Contraction Lemma, the Rademacher complexity of the composite of A and the function is at most the Rademacher complexity of A scaled by our lipschitz bound ρ.

6.2 Mathematical Definition

Let $\phi_i : \mathbb{R} \mapsto \mathbb{R}$ be some ρ-Lipschitz function s.t. $\forall \alpha, \beta \in \mathbb{R}$ the lipschitz condition holds, $||\phi_i(\alpha) - \phi_i(\beta)|| \leq \rho||\alpha - \beta||$. If we then create some set $\boldsymbol{\phi} = \{\phi_i(a_i), ..., \phi_m(a_m)\}$ where all phi_i meet the lipschitz condition and each a_i come from the set A. We can then define the following:

$$\boldsymbol{\phi} \circ A \triangleq \boldsymbol{\phi(a)} : a \in A$$
$$R(\boldsymbol{\phi} \circ A) \leq \rho R(A)$$