Statistical Learning

Youwei Yu
LaTeX by Joe
Indiana University

Semester

Contents

Chapter 1

Probably Approximately Correct

1.1 Probably Approximately Correct

[KV] Sections 1.1, 1.3.

Definition 1.1.1: Probably Approximately Correct

Let $\mathcal{C}$ be a concept class over X. We say that $\mathcal{C}$ is PAC learnable if there exists an algorithm L with the following property: for every concept $c \in \mathcal{C}$, for every distribution $\mathcal{D}$ on X, and for all $0 < \epsilon < 1/2$ and $0 < \delta < 1/2$, if L is given access to $EX(c, \mathcal{D})$ and inputs ϵ and δ, then with probability at least $1 - \delta$, L outputs a hypothesis concept $h \in \mathcal{C}$ satisfying $\text{error}(h) \leq \epsilon$. This probability is taken over the random examples drawn by calls to $EX(c, \mathcal{D})$, and any internal randomization of L.

Example.

> Learning Square. We learn a tighest-fit rectangle, where the total error region is ϵ. So for each edge, the error region is $\epsilon/4$. By union bound, the error rate for m samples is $4(1-\epsilon/4)^m$. Since each draw from the error region can improve our hypothesis, this means after m samples, our hypothesis keeps ϵ error rate. As we want it bounded by δ, we shall have $m \geq \frac{4}{\epsilon} \ln \frac{4}{\delta}$.

Example.

> Learning Boolean Conjunctions. The hypothesis class is $h = x_1 \wedge \bar{x}_1 \wedge \cdots \wedge x_n \wedge \bar{x}_n$, with $2n$ literals. Similarly, each draw from the error literal can improve our hypothesis. Through the union bound, the error rate will be $2n(1-\epsilon/2n)^m$, so we shall have $m \geq \frac{2n}{\epsilon} \ln \frac{2n}{\delta}$ for $\mathbb{P}[err(h, D) > \epsilon] < \delta$ or $\mathbb{P}[err(h, D) < \epsilon] > 1 - \delta$.

1.2 Empirical Risk Minimization

[SSBD] Chapter 2

Remark.

> Domain Set $\mathcal{X}$, Label Set $\mathcal{Y}$ (usually $\{0, 1\}$), Training Data $S = \{(x_1, y_1), \ldots, (x_n, y_n)\} \subset \mathcal{X} \times \mathcal{Y}$.
> Learner's Output $h : \mathcal{X} \to \mathcal{Y}$: predictor, hypothesis, or classifier.
> Data-generation Model: probability distribution over $\mathcal{X}$ by $\mathcal{D}$, "correct" labeling function $f : \mathcal{X} \to \mathcal{Y}$ such that $y_i = f(x_i)$, and dataset S.
> Measure of Success: $L_{\mathcal{D},f}(h) := \mathbb{P}_{x \sim \mathcal{D}}[h(x) \neq f(x)] := \mathcal{D}(\{x : h(x) \neq f(x)\})$, named as general error, risk, or true error.

Remark.

> Empirical Error: $L_S(h) := \frac{|\{i\in[m]:h(x_i)\neq y_i\}|}{m}$, named as empirical error or empirical risk. Since the learner does not know what $\mathcal{D}$ and f are, a predictor h that minimizes $L_S(h)$-is called Empirical Risk Minimization (ERM).

ERM rule may lead to over-fitting, a common solution is to apply the ERM learning rule over a restricted search space, called a hypothesis class and denoted by $\mathcal{H}$. Formally,

$$\mathrm{ERM}_{\mathcal{H}} = \arg\min_{h\in\mathcal{H}} L_S(h)$$

Such restrictions are often called inductive bias.

Definition 1.2.1: Realizability

There exists $h^\star \in \mathcal{H}$ such that $L_{(\mathcal{D},f)}(h^\star) = 0$. Note that this assumption implies that with probability 1 over random samples S, where the instances of S are sampled according to $\mathcal{D}$ and are labeled by f, we have $L_S(h^\star) = 0$.

Let $\mathcal{H}_B$ be the set of "bad" hypotheses, that is, $\mathcal{H}_B = \{h \in \mathcal{H} : L_{(\mathcal{D},f)}(h) > \epsilon\}$. In addition, let $M = \{S \mid \exists h \in \mathcal{H}_B, L_S(h) = 0\}$ be the set of misleading samples. But, since the realizability assumption implies that $L_S(h_S) = 0$, it follows that the event $L_{(\mathcal{D},f)}(h_S) > \epsilon$ can only happen if for some $h \in \mathcal{H}_B$ we have $L_S(h) = 0$. Formally, we have shown that $\{S|_x : L_{(\mathcal{D},f)}(h_S) > \epsilon\} \subseteq M$. Note that $M = \bigcup_{h\in\mathcal{H}_B}\{S|_x : L_S(h) = 0\}$, hence,

$$\mathcal{D}^m\left(\{S|_x : L_{(\mathcal{D},f)}(h_S) > \epsilon\}\right) \leq \mathcal{D}^m(M) = \mathcal{D}^m\left(\bigcup_{h\in\mathcal{H}_B}\{S|_x : L_S(h) = 0\}\right).$$

Lemma 1.2.2: Union Bound

For any two sets A, B and a distribution $\mathcal{D}$, we have $\mathcal{D}(A \cup B) \leq \mathcal{D}(A) + \mathcal{D}(B)$.

Since for each sample we have $\mathcal{D}(\{x_i : h(x_i) = y_i\}) = 1 - L_{\mathcal{D},f}(h) \leq 1 - \epsilon$, for the entire samples we have $\mathcal{D}^m(\{S|_x : L_S(h) = 0\}) \leq (1-\epsilon)^m \leq e^{-\epsilon m}$. We conclude that $\mathcal{D}^m(\{S|_x : L_{(\mathcal{D},f)}(h_S) > \epsilon\}) \leq |\mathcal{H}_B|e^{-\epsilon m} \leq |\mathcal{H}|e^{-\epsilon m}$.

Corollary 1.2.3

Let $\mathcal{H}$ be a finite hypothesis class. Let $\delta \in (0,1)$ and $\epsilon > 0$, and let m be an integer that satisfies

$$m \geq \frac{\log(|\mathcal{H}|/\delta)}{\epsilon}.$$

Then, for any labeling function f, and for any distribution $\mathcal{D}$ for which the realizability assumption holds (that is, for some $h \in \mathcal{H}$, $L_{(\mathcal{D},f)}(h) = 0$), with probability of at least $1-\delta$ over the choice of an i.i.d. sample S of size m, we have that for every ERM hypothesis h_S, it holds that $L_{(\mathcal{D},f)}(h_S) \leq \epsilon$.

Chapter 2

A Formal Learning Model

2.1 Agnostic PAC

[KV] Sections 1.1, 1.3.

Now we define a more realistic model for the data-generating distribution $\mathcal{D}$, i.e., a joint distribution over domain points and labels, and we remove the realizability assumption. Since we cannot know the true error, we require that the learning algorithm will find a predictor whose error is not much larger than the best possible error of a predictor in some given benchmark hypothesis class.

Definition 2.1.1: Agnostic PAC Learnability

A hypothesis class $\mathcal{H}$ is agnostic PAC learnable if there exist a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm with the following property: For every $\epsilon, \delta \in (0,1)$ and for every distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, when running the learning algorithm on $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ i.i.d. examples generated by $\mathcal{D}$, the algorithm returns a hypothesis h such that, with probability of at least $1 - \delta$ (over the choice of the m training examples),

$$L_{\mathcal{D}}(h) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \epsilon.$$

If the realizability assumption holds, agnostic PAC learning provides the same guarantee as PAC learning. In that sense, agnostic PAC learning generalizes the definition of PAC learning. When the realizability assumption does not hold, no learner can guarantee an arbitrarily small error. Nevertheless, under the definition of agnostic PAC learning, a learner can still declare success if its error is not much larger than the best error achievable by a predictor from the class $\mathcal{H}$.

2.2 Uniform Convergence

Definition 2.2.1: ϵ-representative

A training set S is called ϵ-representative (w.r.t. domain Z, hypothesis class $\mathcal{H}$, loss function ℓ, and distribution $\mathcal{D}$) if

$$\forall h \in \mathcal{H}, \quad |L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon.$$

Lemma 2.2.2: $\frac{\epsilon}{2}$-representative

Assume that a training set S is $\frac{\epsilon}{2}$-representative (w.r.t. domain Z, hypothesis class $\mathcal{H}$, loss function ℓ, and distribution $\mathcal{D}$). Then, any output of $\mathrm{ERM}_{\mathcal{H}}(S)$, namely, any $h_S \in \arg\min_{h\in\mathcal{H}} L_S(h)$, satisfies

$$L_{\mathcal{D}}(h_S) \leq \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \epsilon.$$

Proof. For every $h \in \mathcal{H}$,

$$L_{\mathcal{D}}(h_S) \leq L_S(h_S) + \tfrac{\epsilon}{2} \leq L_S(h) + \tfrac{\epsilon}{2} \leq L_{\mathcal{D}}(h) + \tfrac{\epsilon}{2} + \tfrac{\epsilon}{2} = L_{\mathcal{D}}(h) + \epsilon.$$

□

Definition 2.2.3: Uniform Convergence

We say that a hypothesis class $\mathcal{H}$ has the uniform convergence property (w.r.t. a domain Z and a loss function ℓ) if there exists a function $m_{\mathcal{H}}^{\mathrm{UC}} : (0,1)^2 \to \mathbb{N}$ such that for every $\epsilon, \delta \in (0,1)$ and for every probability distribution $\mathcal{D}$ over Z, if S is a sample of $m \geq m_{\mathcal{H}}^{\mathrm{UC}}(\epsilon, \delta)$ examples drawn i.i.d. according to $\mathcal{D}$, then, with probability of at least $1-\delta$, S is ϵ-representative.

Corollary 2.2.4

If a class $\mathcal{H}$ has the uniform convergence property with a function $m_{\mathcal{H}}^{\mathrm{UC}}$, then the class is *agnostically PAC learnable* with the sample complexity

$$m_{\mathcal{H}}(\epsilon, \delta) \leq m_{\mathcal{H}}^{\mathrm{UC}}(\epsilon/2, \delta).$$

Furthermore, the $\mathrm{ERM}_{\mathcal{H}}$ paradigm is a successful agnostic PAC learner for $\mathcal{H}$.

2.3 Finite Agnostic PAC

We want that for all $h \in \mathcal{H}$, $|L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon$. That is, $\mathcal{D}^m(\{S : \forall h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon\}) \geq 1 - \delta$. Equivalently, we need to show that $\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) < \delta$. Writing $\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\} = \bigcup_{h\in\mathcal{H}}\{S : |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}$, and applying the union bound we obtain

$$\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) \leq \sum_{h\in\mathcal{H}} \mathcal{D}^m(\{S : |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}).$$

By the linearity of expectation, it follows that $L_{\mathcal{D}}(h)$ is also the expected value of $L_S(h)$. Hence, the quantity $|L_{\mathcal{D}}(h) - L_S(h)|$ is the deviation of the random variable $L_S(h)$ from its expectation.

Lemma 2.3.1: Hoeffding's Inequality

Let $\theta_1, \ldots, \theta_m$ be a sequence of i.i.d. random variables and assume that for all i, $\mathbb{E}[\theta_i] = \mu$ and $\mathbb{P}[a \leq \theta_i \leq b] = 1$. Then, for any $\epsilon > 0$

$$\mathbb{P}\left[\left|\frac{1}{m}\sum_{i=1}^{m}\theta_i - \mu\right| > \epsilon\right] \leq 2\exp\left(\frac{-2m\epsilon^2}{(b-a)^2}\right).$$

For number of $|\mathcal{H}|$ classes and $b = 1, a = 0$, we use the union bound on top of Hoeffding's Inequality.

Corollary 2.3.2

Let $\mathcal{H}$ be a finite hypothesis class, let Z be a domain, and let $\ell : \mathcal{H} \times Z \to [0, 1]$ be a loss function. Then, $\mathcal{H}$ enjoys the uniform convergence property with sample complexity

$$m_{\mathcal{H}}^{\mathrm{UC}}(\epsilon, \delta) \leq \left\lceil \frac{\log(2|\mathcal{H}|/\delta)}{2\epsilon^2} \right\rceil .$$

Furthermore, the class is agnostically PAC learnable using the ERM algorithm with sample complexity

$$m_{\mathcal{H}}(\epsilon, \delta) \leq m_{\mathcal{H}}^{\mathrm{UC}}(\epsilon/2, \delta) \leq \left\lceil \frac{2\log(2|\mathcal{H}|/\delta)}{\epsilon^2} \right\rceil .$$

Chapter 3

Variations

3.1 Weak and Strong Learning

Chapter 4

Nonuniform Learnability

[SSBD] Chapter 7

4.1 Nonuniform Learnability

Nonuniform "learnability" allows the sample size to be nonuniform with respect to the different hypotheses with which the learner is competing.

Definition 4.1.1: Nonuniformly Learnable

A hypothesis class $\mathcal{H}$ is nonuniformly learnable if there exist a learning algorithm, A, and a function $m_{\mathcal{H}}^{\mathrm{NUL}} : (0,1)^2 \times \mathcal{H} \to \mathbb{N}$ such that, for every $\epsilon, \delta \in (0,1)$ and for every $h \in \mathcal{H}$, if $m \geq m_{\mathcal{H}}^{\mathrm{NUL}}(\epsilon, \delta, h)$ then for every distribution $\mathcal{D}$, with probability of at least $1-\delta$ over the choice of $S \sim \mathcal{D}^m$, it holds that

$$L_{\mathcal{D}}(A(S)) \leq L_{\mathcal{D}}(h) + \epsilon.$$

The difference to agnostic PAC learnability is the question of whether the sample size m may depend on the hypothesis h to which the error of $A(S)$ is compared. Note that that nonuniform learnability is a strict relaxation of agnostic PAC learnability. That is, if a class is agnostic PAC learnable then it is also nonuniformly learnable.

Theorem 4.1.2

A hypothesis class $\mathcal{H}$ of binary classifiers is nonuniformly learnable if and only if it is a countable union of agnostic PAC learnable hypothesis classes.

Theorem 4.1.3

Let $\mathcal{H}$ be a hypothesis class that can be written as a countable union of hypothesis classes, $\mathcal{H} = \bigcup_{n\in\mathbb{N}} \mathcal{H}_n$, where each $\mathcal{H}_n$ enjoys the uniform convergence property. Then, $\mathcal{H}$ is nonuniformly learnable.

4.2 Structural Risk Minimization

Let us define the function $\epsilon_n : \mathbb{N} \times (0,1) \to (0,1)$ by

$$\epsilon_n(m, \delta) = \min\{\epsilon \in (0,1) : m_{\mathcal{H}_n}^{\mathrm{UC}}(\epsilon, \delta) \leq m\}.$$

In words, we have a fixed sample size m, and we are interested in the lowest possible upper bound on the gap between empirical and true risks achievable by using a sample of m examples.

Theorem 4.2.1

Let $w : \mathbb{N} \to [0,1]$ be a function such that $\sum_{n=1}^{\infty} w(n) \leq 1$. Let $\mathcal{H}$ be a hypothesis class that can be written as $\mathcal{H} = \bigcup_{n \in \mathbb{N}} \mathcal{H}_n$, where for each n, $\mathcal{H}_n$ satisfies the uniform convergence property with a sample complexity function $m^{\mathrm{UC}}_{\mathcal{H}_n}$. Then, for every $\delta \in (0,1)$ and distribution $\mathcal{D}$, with probability of at least $1-\delta$ over the choice of $S \sim \mathcal{D}^m$, the following bound holds (simultaneously) for every $n \in \mathbb{N}$ and $h \in \mathcal{H}_n$:

$$|L_{\mathcal{D}}(h) - L_S(h)| \leq \epsilon_n(m, w(n) \cdot \delta).$$

Therefore, for every $\delta \in (0,1)$ and distribution $\mathcal{D}$, with probability of at least $1-\delta$ it holds that

$$\forall h \in \mathcal{H}, \quad L_{\mathcal{D}}(h) \leq L_S(h) + \min_{n : h \in \mathcal{H}_n} \epsilon_n(m, w(n) \cdot \delta).$$

Let $w : \mathbb{N} \to [0,1]$ be a function such that $\sum_{n=1}^{\infty} w(n) \leq 1$. We refer to w as a *weight function* over the hypothesis classes $\mathcal{H}_1, \mathcal{H}_2, \ldots$. Such a weight function can reflect the importance that the learner attributes to each hypothesis class, or some measure of the complexity of different hypothesis classes.

Structural Risk Minimization (SRM)

prior knowledge:
$\mathcal{H} = \bigcup_n \mathcal{H}_n$ where $\mathcal{H}_n$ has uniform convergence with $m^{\mathrm{UC}}_{\mathcal{H}_n}$
$w : \mathbb{N} \to [0,1]$ where $\sum_n w(n) \leq 1$
define: ϵ_n as in Equation; $n(h)$ as in Equation
input: training set $S \sim \mathcal{D}^m$, confidence δ
output: $h \in \arg\min_{h \in \mathcal{H}} \left[L_S(h) + \epsilon_{n(h)}(m, w(n(h)) \cdot \delta) \right]$

Theorem 4.2.2

Let $\mathcal{H}$ be a hypothesis class such that $\mathcal{H} = \bigcup_{n \in \mathbb{N}} \mathcal{H}_n$, where each $\mathcal{H}_n$ has the uniform convergence property with sample complexity $m^{\mathrm{UC}}_{\mathcal{H}_n}$. Let $w : \mathbb{N} \to [0,1]$ be such that $w(n) = \frac{6}{n^2 \pi^2}$. Then, $\mathcal{H}$ is nonuniformly learnable using the *SRM* rule with rate

$$m^{\mathrm{NUL}}_{\mathcal{H}}(\epsilon, \delta, h) \leq m^{\mathrm{UC}}_{\mathcal{H}_{n(h)}} \left(\frac{\epsilon}{2}, \frac{6\delta}{(\pi n(h))^2} \right).$$

4.3 Learning 3-Term DNF Formulae

Theorem 4.3.1: Intractability of Learning 3-Term DNF Formulae

If RP $\neq$ NP, the representation class of 3-term DNF formulae is not efficiently PAC learnable.

Definition 4.3.2

Let $S = \{\langle x_1, b_1 \rangle, \ldots, \langle x_m, b_m \rangle\}$ be any labeled set of instances, where each $x_i \in X$ and each $b_i \in \{0,1\}$. Let c be a concept over X. Then we say that c is consistent with S (or equivalently, S is consistent with c) if for all $1 \leq i \leq m$, $c(x_i) = b_i$.

Theorem 4.3.3: Using 3-CNF Formulae to Avoid Intractability

The representation class of 3-CNF formulae is efficiently PAC learnable.

Definition 4.3.4: The PAC Model, Final Definition

If $\mathcal{C}$ is a concept class over X and $\mathcal{H}$ is a representation class over X, we will say that $\mathcal{C}$ is (efficiently) PAC learnable using $\mathcal{H}$ if our basic definition of PAC learning (Definition 2) is met by an algorithm that is now allowed to output a hypothesis from $\mathcal{H}$. Here we are implicitly assuming that $\mathcal{H}$ is at least as expressive as $\mathcal{C}$, and so there is a representation in $\mathcal{H}$ of every function in $\mathcal{C}$. We will refer to $\mathcal{H}$ as the hypothesis class of the PAC learning algorithm.

Definition 4.3.5

We say that the representation class $\mathcal{H}$ is polynomially evaluatable if there is an algorithm that on input any instance $x \in X_n$ and any representation $h \in \mathcal{H}_n$, outputs the value $h(x)$ in time polynomial in n and $\mathit{size}(h)$.

Theorem 4.3.6

The representation class of 1-term DNF formulae (conjunctions) is efficiently PAC learnable using 1-term DNF formulae. For any constant $k \geq 2$, the representation class of k-term DNF formulae is not efficiently PAC learnable using k-term DNF formulae (unless RP = NP), but is efficiently PAC learnable using k-CNF formulae.

4.4 Consistency Learnability

Definition 4.4.1: Consistency

Let Z be a domain set, let $\mathcal{P}$ be a set of probability distributions over Z, and let $\mathcal{H}$ be a hypothesis class. A learning rule A is *consistent* with respect to $\mathcal{H}$ and $\mathcal{P}$ if there exists a function $m_{\mathcal{H}}^{\mathrm{CON}} : (0,1)^2 \times \mathcal{H} \times \mathcal{P} \to \mathbb{N}$ such that, for every $\epsilon, \delta \in (0,1)$, every $h \in \mathcal{H}$, and every $\mathcal{D} \in \mathcal{P}$, if $m \geq m_{\mathcal{H}}^{\mathrm{NUL}}(\epsilon, \delta, h, \mathcal{D})$ then with probability of at least $1 - \delta$ over the choice of $S \sim \mathcal{D}^m$, it holds that

$$L_{\mathcal{D}}(A(S)) \leq L_{\mathcal{D}}(h) + \epsilon.$$

If $\mathcal{P}$ is the set of all distributions, we say that A is universally consistent with respect to $\mathcal{H}$.

The notion of consistency is, of course, a relaxation of our previous notion of nonuniform learnability. Clearly if an algorithm nonuniformly learns a class H it is also universally consistent for that class. The relaxation is strict in the sense that there are consistent learning rules that are not successful nonuniform learners.

Chapter 5

Vapnik-Chervonenkis Dimension

5.1 VCD

Definition 5.1.1: Behavior

For any concept class $\mathcal{C}$ over X, and any $S \subseteq X$,

$$\Pi_{\mathcal{C}}(S) = \{c \cap S : c \in \mathcal{C}\}.$$

Equivalently, if $S = \{x_1, \dots, x_m\}$, then we can think of $\Pi_{\mathcal{C}}(S)$ as the set of vectors $\Pi_{\mathcal{C}}(S) \subseteq \{0,1\}^m$:

$$\Pi_{\mathcal{C}}(S) = \{(c(x_1), \dots, c(x_m)) : c \in \mathcal{C}\}.$$

Thus, $\Pi_{\mathcal{C}}(S)$ is the set of all the behaviors or dichotomies on S that are induced or realized by $\mathcal{C}$.

Definition 5.1.2: Shattering

If $\Pi_{\mathcal{C}}(S) = \{0,1\}^m$ (where $m = |S|$), then we say that S is shattered by $\mathcal{C}$. Thus, S is shattered by $\mathcal{C}$ if $\mathcal{C}$ realizes all possible dichotomies of S.

Definition 5.1.3: VC Dimension

The Vapnik–Chervonenkis (VC) dimension of $\mathcal{C}$, denoted as VCD($\mathcal{C}$), is the cardinality of the largest set S shattered by $\mathcal{C}$. If arbitrarily large finite sets can be shattered by $\mathcal{C}$, then VCD($\mathcal{C}$) $= \infty$.

Example.

> Intervals of the real line. VCD=2
> Linear halfspaces in the plane. $\mathbb{R}^d \Rightarrow d+1$
> Axis-aligned rectangles in the plane. VCD=4

Definition 5.1.4

For any natural number m we define

$$\Pi_{\mathcal{C}}(m) = \max\{|\Pi_{\mathcal{C}}(S)| : |S| = m\}.$$

5.2 Polynomial Bound on $|\Pi_{\mathcal{C}}(S)|$

Definition 5.2.1

For any natural numbers m and d, the function $\Phi_d(m)$ is defined inductively by

$$\Phi_d(m) = \Phi_d(m-1) + \Phi_{d-1}(m-1)$$

with initial conditions $\Phi_d(0) = \Phi_0(m) = 1$.

Lemma 5.2.2

If $\mathrm{VCD}(\mathcal{C}) = d$, then for any m, $\Pi_{\mathcal{C}}(m) \leq \Phi_d(m)$.

Lemma 5.2.3

$$\Phi_d(m) = \sum_{i=0}^{d} \binom{m}{i}.$$

5.3 Polynomial Bound on the Sample Size for PAC Learning

Let us now fix the target concept $c \in \mathcal{C}$, and define the class of error regions with respect to c and $\mathcal{C}$ by $\Delta(c) = \{c\Delta c' : c' \in \mathcal{C}\}$. It is easy to show that $\mathrm{VCD}(\mathcal{C}) = \mathrm{VCD}(\Delta(c))$. To see this, for any set S we can map each element $c' \in \Pi_{\mathcal{C}}(S)$ to $c'\Delta(c \cap S) \in \Pi_{\Delta(c)}(S)$. Since this is a bijective mapping of $\Pi_{\mathcal{C}}(S)$ to $\Pi_{\Delta(c)}(S)$, $|\Pi_{\Delta(c)}(S)| = |\Pi_{\mathcal{C}}(S)|$. Since this holds for any set S, $\mathrm{VCD}(\mathcal{C}) = \mathrm{VCD}(\Delta(c))$ follows. We may further refine the definition of $\Delta(c)$ to consider only those error regions with weight at least ϵ under the fixed target distribution $\mathcal{D}$. Thus, let

$$\Delta_\epsilon(c) = \{r \in \Delta(c) : \Pr_{x\sim\mathcal{D}}[x \in r] \geq \epsilon\}.$$

Definition 5.3.1: ϵ-net

For any $\epsilon > 0$, we say that a set S is an ϵ-net for $\Delta(c)$ if every region in $\Delta_\epsilon(c)$ is "hit" by a point in S, that is, if for every $r \in \Delta_\epsilon(c)$ we have $S \cap r \neq \emptyset$.

The important property of ϵ-nets is that if the sample S drawn by a learning algorithm forms an ϵ-net for $\Delta(c)$, and the learning algorithm outputs a hypothesis $h \in \mathcal{C}$ that is consistent with S, then this hypothesis must have error less than ϵ: since $c\Delta h \in \Delta(c)$ was not hit by S (otherwise h would not be consistent with S), and S is an ϵ-net for $\Delta(c)$, we must have $c\Delta h \notin \Delta_\epsilon(c)$ and therefore $\mathrm{error}(h) \leq \epsilon$.

Thus if we can bound the probability that the random sample S fails to form an ϵ-net for $\Delta(c)$, then we have bounded the probability that a hypothesis consistent with S has error greater than ϵ. For the case of finite $\mathcal{C}$, for any fixed error region $c\Delta h \in \Delta_\epsilon(c)$, the probability that we fail to hit $c\Delta h$ in m random examples is at most $(1-\epsilon)^m$. Thus the probability that we fail to hit some $c\Delta h \in \Delta_\epsilon(c)$ is bounded above by $|\Delta(c)|(1-\epsilon)^m$, which in turn is bounded by $|\mathcal{C}|(1-\epsilon)^m$.

Lemma 5.3.2

This gives us a bound of $\Phi_d(|X|)(1-\epsilon)^m$ on the probability of failing to draw an ϵ-net for $\Delta(c)$. However, if X is infinite then $\Phi_d(|X|)$ is infinite as well.

Theorem 5.3.3

Let $\mathcal{C}$ be any concept class of VC dimension d. Let L be any algorithm that takes as input a set S of m labeled examples of a concept in $\mathcal{C}$, and produces as output a concept $h \in \mathcal{C}$ that is consistent with S. Then L is a PAC learning algorithm for $\mathcal{C}$ provided it is given a random sample of m examples from $\mathrm{EX}(c, \mathcal{D})$, where m obeys

$$m \geq c_0 \left(\frac{1}{\epsilon} \log \frac{1}{\delta} + \frac{d}{\epsilon} \log \frac{1}{\epsilon} \right)$$

for some constant $c_0 > 0$.

Theorem 5.3.4

Let $\mathcal{C}$ be any concept class. Let $\mathcal{H}$ be any representation class of VC dimension d. Let L be any algorithm that takes as input a set S of m labeled examples of a concept in $\mathcal{C}$, and produces as output a concept $h \in \mathcal{H}$ that is consistent with S. Then L is a PAC learning algorithm for $\mathcal{C}$ using $\mathcal{H}$ provided it is given a random sample of m examples from $\mathrm{EX}(c, \mathcal{D})$, where m obeys

$$m \geq c_0 \left(\frac{1}{\epsilon} \log \frac{1}{\delta} + \frac{d}{\epsilon} \log \frac{1}{\epsilon} \right)$$

for some constant $c_0 > 0$.

5.4 Sample Size Lower bounds

Chapter 6

Online Learning

6.1 Online Classification in the Realizable Case

Definition 6.1.1: Mistake Bounds, Online Learnability

Let $\mathcal{H}$ be a hypothesis class and let A be an online learning algorithm. Given any sequence $S = (x_1, h^\star(y_1)), \ldots, (x_T, h^\star(y_T))$, where T is any integer and $h^\star \in \mathcal{H}$, let $M_A(S)$ be the number of mistakes A makes on the sequence S. We denote by $M_A(\mathcal{H})$ the supremum of $M_A(S)$ over all sequences of the above form. A bound of the form $M_A(\mathcal{H}) \leq B < \infty$ is called a mistake bound. We say that a hypothesis class $\mathcal{H}$ is online learnable if there exists an algorithm A for which $M_A(\mathcal{H}) \leq B < \infty$.

```
Consistent

input: A finite hypothesis class $\mathcal{H}$
initialize: $V_1 = \mathcal{H}$
for $t = 1, 2, \ldots$
    receive $\mathbf{x}_t$
    choose any $h \in V_t$
    predict $p_t = h(\mathbf{x}_t)$
    receive true label $y_t = h^\star(\mathbf{x}_t)$
    update $V_{t+1} = \{h \in V_t : h(\mathbf{x}_t) = y_t\}$
```

Corollary 6.1.2

Let $\mathcal{H}$ be a finite hypothesis class. The Consistent algorithm enjoys the mistake bound $M_{\text{Consistent}}(\mathcal{H}) \leq |\mathcal{H}| - 1$.

```
Halving

input: A finite hypothesis class $\mathcal{H}$
initialize: $V_1 = \mathcal{H}$
for $t = 1, 2, \ldots$
    receive $\mathbf{x}_t$
    predict $p_t = \arg\max_{r \in \{0,1\}} |\{h \in V_t : h(\mathbf{x}_t) = r\}|$    (in case of a tie, predict $p_t = 1$)
    receive true label $y_t = h^\star(\mathbf{x}_t)$
    update $V_{t+1} = \{h \in V_t : h(\mathbf{x}_t) = y_t\}$
```

Corollary 6.1.3

Let $\mathcal{H}$ be a finite hypothesis class. The Halving algorithm enjoys the mistake bound $M_{\text{Halving}}(\mathcal{H}) \leq \log_2(|\mathcal{H}|)$.

6.1.1 Online Learnability

Definition 6.1.4: $\mathcal{H}$-Shattered Tree

A shattered tree of depth d is a sequence of instances $\mathbf{v}_1, \ldots, \mathbf{v}_{2^d-1}$ in $\mathcal{X}$ such that for every labeling $(y_1, \ldots, y_d) \in \{0,1\}^d$ there exists $h \in \mathcal{H}$ such that for all $t \in [d]$ we have

$$h(\mathbf{v}_{i_t}) = y_t \quad \text{where} \quad i_t = 2^{t-1} + \sum_{j=1}^{t-1} y_j 2^{t-1-j}.$$

Definition 6.1.5: Littlestone's Dimension (Ldim)

$\text{Ldim}(\mathcal{H})$ is the maximal integer T such that there exists a shattered tree of depth T, which is shattered by $\mathcal{H}$.

Lemma 6.1.6

No algorithm can have a mistake bound strictly smaller than $\text{Ldim}(\mathcal{H})$; namely, for every algorithm A, we have $M_A(\mathcal{H}) \geq \text{Ldim}(\mathcal{H})$.

Proof for Lemma

> Let $T = \text{Ldim}(\mathcal{H})$ and let $\mathbf{v}_1, \ldots, \mathbf{v}_{2^T-1}$ be a sequence that satisfies the requirements in the definition of Ldim. If the environment sets $\mathbf{x}_t = \mathbf{v}_{i_t}$ and $y_t = 1 - p_t$ for all $t \in [T]$, then the learner makes T mistakes while the definition of Ldim implies that there exists a hypothesis $h \in \mathcal{H}$ such that $y_t = h(\mathbf{x}_t)$ for all t. ■

```
Standard Optimal Algorithm (SOA)

input: A hypothesis class $\mathcal{H}$
initialize: $V_1 = \mathcal{H}$
for $t = 1, 2, \ldots$
    receive $\mathbf{x}_t$
    for $r \in \{0,1\}$ let $V_t^{(r)} = \{h \in V_t : h(\mathbf{x}_t) = r\}$
    predict $p_t = \arg\max_{r \in \{0,1\}} \text{Ldim}(V_t^{(r)})$      (in case of a tie, predict $p_t = 1$)
    receive true label $y_t$
    update $V_{t+1} = \{h \in V_t : h(\mathbf{x}_t) = y_t\}$
```

Lemma 6.1.7

SOA enjoys the mistake bound $M_{\text{SOA}}(\mathcal{H}) \leq \text{Ldim}(\mathcal{H})$.

Corollary 6.1.8

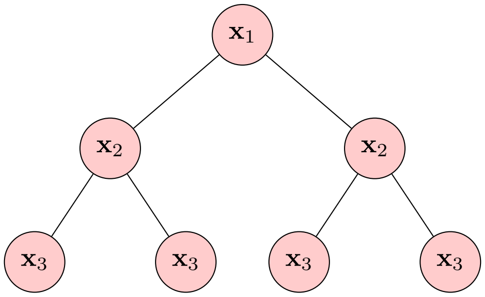
Let $\mathcal{H}$ be any hypothesis class. Then, the standard optimal algorithm enjoys the mistake bound $M_{\text{SOA}}(\mathcal{H}) = \text{Ldim}(\mathcal{H})$, and no other algorithm can have $M_A(\mathcal{H}) < \text{Ldim}(\mathcal{H})$.

6.1.2 Comparison to VC Dimension

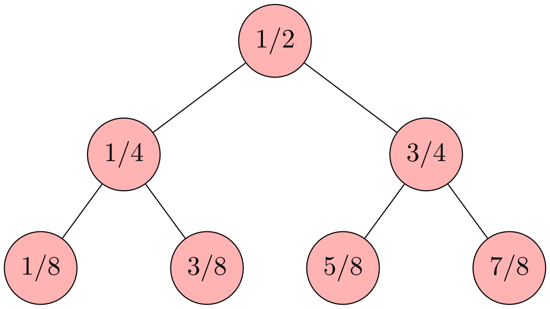
Theorem 6.1.9

For any class $\mathcal{H}$, $\text{VCdim}(\mathcal{H}) \leq \text{Ldim}(\mathcal{H})$, and there are classes for which strict inequality holds. Furthermore, the gap can be arbitrarily large.

Proof. Suppose $\text{VCdim}(\mathcal{H}) = d$ and let $\mathbf{x}_1, \ldots, \mathbf{x}_d$ be a shattered set. We now construct a complete binary tree of instances $\mathbf{v}_1, \ldots, \mathbf{v}_{2^d-1}$, where all nodes at depth i are set to be $\mathbf{x}_i$ — see the following illustration:



Now, the definition of a shattered set clearly implies that we got a valid shattered tree of depth d, and we conclude that $\text{VCdim}(\mathcal{H}) \leq \text{Ldim}(\mathcal{H})$. To show that the gap can be arbitrarily large, let $\mathcal{X} = [0, 1]$ and $\mathcal{H} = \{x \mapsto \mathbb{1}_{[x<a]} : a \in [0, 1]\}$; namely, $\mathcal{H}$ is the class of thresholds on the interval $[0, 1]$. Then, $\text{Ldim}(\mathcal{H}) = \infty$. To see this, consider the tree:



This tree is shattered by $\mathcal{H}$. And, because of the density of the reals, this tree can be made arbitrarily deep. □

Chapter 7

Convex Learning

7.1 Convexity, Lipschitzness, and Smoothness

7.1.1 Convexity

Definition 7.1.1: Convex Set

A set C in a vector space is convex if for any two vectors $\mathbf{u}, \mathbf{v}$ in C, the line segment between $\mathbf{u}$ and $\mathbf{v}$ is contained in C. That is, for any $\alpha \in [0, 1]$ we have that

$$\alpha\mathbf{u} + (1 - \alpha)\mathbf{v} \in C.$$

Definition 7.1.2: Convex Function

Let C be a convex set. A function $f : C \to \mathbb{R}$ is convex if for every $\mathbf{u}, \mathbf{v} \in C$ and $\alpha \in [0, 1]$,

$$f(\alpha\mathbf{u} + (1 - \alpha)\mathbf{v}) \le \alpha f(\mathbf{u}) + (1 - \alpha)f(\mathbf{v}).$$

The epigraph of a function f is the set

$$\text{epigraph}(f) = \{(\mathbf{x}, \beta) : f(\mathbf{x}) \le \beta\}.$$

It is easy to verify that a function f is convex if and only if its epigraph is a convex set. Every local minimum of the function is also a global minimum. For every $\mathbf{w}$ we can construct a tangent to f at $\mathbf{w}$ that lies below f everywhere. For convex differentiable functions,

$$\forall \mathbf{u}, \quad f(\mathbf{u}) \ge f(\mathbf{w}) + \langle \nabla f(\mathbf{w}), \mathbf{u} - \mathbf{w} \rangle.$$

Lemma 7.1.3

Let $f : \mathbb{R} \to \mathbb{R}$ be a scalar twice differentiable function, and let f', f'' be its first and second derivatives, respectively. Then, the following are equivalent:

1. f is convex
2. f' is monotonically nondecreasing
3. f'' is nonnegative

Claim

Assume that $f : \mathbb{R}^d \to \mathbb{R}$ can be written as $f(\mathbf{w}) = g(\langle \mathbf{w}, \mathbf{x} \rangle + y)$, for some $\mathbf{x} \in \mathbb{R}^d$, $y \in \mathbb{R}$, and $g : \mathbb{R} \to \mathbb{R}$. Then, convexity of g implies the convexity of f.

Claim

For $i = 1, \ldots, r$, let $f_i : \mathbb{R}^d \to \mathbb{R}$ be a convex function. The following functions from $\mathbb{R}^d$ to $\mathbb{R}$ are also convex:

- $g(x) = \max_{i \in [r]} f_i(x)$
- $g(x) = \sum_{i=1}^{r} w_i f_i(x)$, where for all i, $w_i \geq 0$.

7.1.2 Lipschitzness

Definition 7.1.4: Lipschitzness

Let $C \subset \mathbb{R}^d$. A function $f : \mathbb{R}^d \to \mathbb{R}^k$ is ρ-Lipschitz over C if for every $\mathbf{w}_1, \mathbf{w}_2 \in C$ we have that

$$\|f(\mathbf{w}_1) - f(\mathbf{w}_2)\| \leq \rho \|\mathbf{w}_1 - \mathbf{w}_2\|.$$

Claim

Let $f(\mathbf{x}) = g_1(g_2(\mathbf{x}))$, where g_1 is ρ_1-Lipschitz and g_2 is ρ_2-Lipschitz. Then, f is $(\rho_1 \rho_2)$-Lipschitz. In particular, if g_2 is the linear function $g_2(\mathbf{x}) = \langle \mathbf{v}, \mathbf{x} \rangle + b$, for some $\mathbf{v} \in \mathbb{R}^d$, $b \in \mathbb{R}$, then f is $(\rho_1 \|\mathbf{v}\|)$-Lipschitz.

7.1.3 Smoothness

Definition 7.1.5: Smoothness

A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is β-smooth if its gradient is β-Lipschitz; namely, for all $\mathbf{v}, \mathbf{w}$ we have

$$\|\nabla f(\mathbf{v}) - \nabla f(\mathbf{w})\| \leq \beta \|\mathbf{v} - \mathbf{w}\|.$$

Claim

Let $f(\mathbf{w}) = g(\langle \mathbf{w}, \mathbf{x} \rangle + b)$, where $g : \mathbb{R} \to \mathbb{R}$ is a β-smooth function, $\mathbf{x} \in \mathbb{R}^d$, and $b \in \mathbb{R}$. Then, f is $(\beta \|\mathbf{x}\|^2)$-smooth.

7.2 Surrogate Loss Functions

For example, in the context of learning halfspaces, we can define the so-called hinge loss as a convex surrogate for the $0 - 1$ loss, as follows:

$$\ell^{\text{hinge}}(\mathbf{w}, (\mathbf{x}, y)) \overset{\text{def}}{=} \max\{0, 1 - y\langle \mathbf{w}, \mathbf{x} \rangle\}.$$

Clearly, for all $\mathbf{w}$ and all $(\mathbf{x}, y)$, $\ell^{0-1}(\mathbf{w}, (\mathbf{x}, y)) \leq \ell^{\text{hinge}}(\mathbf{w}, (\mathbf{x}, y))$. The generalization requirement from a hinge loss learner will have the form

$$L_{\mathcal{D}}^{\text{hinge}}(A(S)) \leq \min_{\mathbf{w}\in\mathcal{H}} L_{\mathcal{D}}^{\text{hinge}}(\mathbf{w}) + \epsilon,$$

where $L_{\mathcal{D}}^{\text{hinge}}(\mathbf{w}) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell^{\text{hinge}}(\mathbf{w},(\mathbf{x},y))]$. Using the surrogate property, we can lower bound the left-hand side by $L_{\mathcal{D}}^{0-1}(A(S))$, which yields

$$L_{\mathcal{D}}^{0-1}(A(S)) \leq \min_{\mathbf{w}\in\mathcal{H}} L_{\mathcal{D}}^{\text{hinge}}(\mathbf{w}) + \epsilon.$$

We can further rewrite the upper bound as follows:

$$L_{\mathcal{D}}^{0-1}(A(S)) \leq \min_{\mathbf{w}\in\mathcal{H}} L_{\mathcal{D}}^{0-1}(\mathbf{w}) + \left(\min_{\mathbf{w}\in\mathcal{H}} L_{\mathcal{D}}^{\text{hinge}}(\mathbf{w}) - \min_{\mathbf{w}\in\mathcal{H}} L_{\mathcal{D}}^{0-1}(\mathbf{w})\right) + \epsilon.$$

That is, the $0-1$ error of the learned predictor is upper bounded by three terms:

- **Approximation error:** This is the term $\min_{\mathbf{w}\in\mathcal{H}} L_{\mathcal{D}}^{0-1}(\mathbf{w})$, which measures how well the hypothesis class performs on the distribution.
- **Estimation error:** This is the error that results from the fact that we only receive a training set and do not observe the distribution $\mathcal{D}$.
- **Optimization error:** This is the term $\left(\min_{\mathbf{w}\in\mathcal{H}} L_{\mathcal{D}}^{\text{hinge}}(\mathbf{w}) - \min_{\mathbf{w}\in\mathcal{H}} L_{\mathcal{D}}^{0-1}(\mathbf{w})\right)$, which measures the difference between the approximation error with respect to the surrogate loss and the approximation error with respect to the original loss. The optimization error is a result of our inability to minimize the training loss with respect to the original loss. The size of this error depends on the specific distribution of the data and on the specific surrogate loss we are using.

7.3 Stochastic Gradient Descent

In the previous analyses of the GD and SGD algorithms, we required that the norm of $\mathbf{w}^\star$ will be at most B, which is equivalent to requiring that $\mathbf{w}^\star$ is in the set $\mathcal{H} = \{\mathbf{w} : \|\mathbf{w}\| \leq B\}$. In terms of learning, this means restricting ourselves to a B-bounded hypothesis class. Yet any step we take in the opposite direction of the gradient (or its expected direction) might result in stepping out of this bound, and there is even no guarantee that $\bar{\mathbf{w}}$ satisfies it. The basic idea is to add a projection step; namely, we will now have a two-step update rule, where we first subtract a subgradient from the current value of $\mathbf{w}$ and then project the resulting vector onto $\mathcal{H}$. Formally,

1. $\mathbf{w}^{(t+\frac{1}{2})} = \mathbf{w}^{(t)} - \eta\mathbf{v}_t$
2. $\mathbf{w}^{(t+1)} = \arg\min_{\mathbf{w}\in\mathcal{H}} \|\mathbf{w} - \mathbf{w}^{(t+\frac{1}{2})}\|$

Lemma 7.3.1: Projection Lemma

Let $\mathcal{H}$ be a closed convex set and let $\mathbf{v}$ be the projection of $\mathbf{w}$ onto $\mathcal{H}$, namely,

$$\mathbf{v} = \arg\min_{\mathbf{x}\in\mathcal{H}} \|\mathbf{x} - \mathbf{w}\|^2.$$

Then, for every $\mathbf{u} \in \mathcal{H}$,

$$\|\mathbf{w} - \mathbf{u}\|^2 - \|\mathbf{v} - \mathbf{u}\|^2 \geq 0.$$

7.4 Tikhonov Regularization as a Stabilizer

Definition 7.4.1: Strongly Convex Functions

A function f is λ-strongly convex if for all $\mathbf{w}, \mathbf{u}$ and $\alpha \in (0,1)$ we have

$$f(\alpha\mathbf{w} + (1-\alpha)\mathbf{u}) \leq \alpha f(\mathbf{w}) + (1-\alpha)f(\mathbf{u}) - \frac{\lambda}{2}\alpha(1-\alpha)\|\mathbf{w}-\mathbf{u}\|^2.$$

Lemma 7.4.2

1. The function $f(\mathbf{w}) = \lambda\|\mathbf{w}\|^2$ is 2λ-strongly convex.
2. If f is λ-strongly convex and g is convex, then $f+g$ is λ-strongly convex.
3. If f is λ-strongly convex and $\mathbf{u}$ is a minimizer of f, then, for any $\mathbf{w}$,

$$f(\mathbf{w}) - f(\mathbf{u}) \geq \frac{\lambda}{2}\|\mathbf{w}-\mathbf{u}\|^2.$$

Theorem 7.4.3

Let $\mathcal{D}$ be a distribution. Let $S = (z_1, \ldots, z_m)$ be an i.i.d. sequence of examples and let z' be another i.i.d. example. Let $U(m)$ be the uniform distribution over $[m]$. Then, for any learning algorithm,

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}(A(S)) - L_S(A(S))\right] = \mathbb{E}_{(S,z')\sim\mathcal{D}^{m+1},\, i\sim U(m)}\left[\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i)\right].$$

When the right-hand side of Equation is small, we say that A is a stable algorithm — changing a single example in the training set does not lead to a significant change. Formally,

Definition 7.4.4: On-Average-Replace-One-Stable

Let $\epsilon : \mathbb{N} \to \mathbb{R}$ be a monotonically decreasing function. We say that a learning algorithm A is on-average-replace-one-stable with rate $\epsilon(m)$ if for every distribution $\mathcal{D}$

$$\mathbb{E}_{(S,z')\sim\mathcal{D}^{m+1},\, i\sim U(m)}\left[\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i)\right] \leq \epsilon(m).$$

A learning algorithm does not overfit if and only if it is on-average-replace-one-stable.

Corollary 7.4.5

Assume that the loss function is convex and ρ-Lipschitz. Then, the Regularized Loss Minimization (RLM, $\arg\min_{\mathbf{w}} (L_S(\mathbf{w}) + R(\mathbf{w}))$) rule with the regularizer $\lambda\|\mathbf{w}\|^2$ is on-average-replace-one-stable with rate $\frac{2\rho^2}{\lambda m}$.

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}(A(S)) - L_S(A(S))\right] \leq \frac{2\rho^2}{\lambda m}.$$

Corollary 7.4.6

Assume that the loss function is β-smooth and nonnegative. Then, the RLM rule with the regularizer $\lambda\|\mathbf{w}\|^2$, where $\lambda \geq \frac{2\beta}{m}$, satisfies

$$\mathbb{E}\Big[\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i)\Big] \leq \frac{48\beta}{\lambda m}\,\mathbb{E}[L_S(A(S))]\,.$$

Note that if for all z we have $\ell(\mathbf{0}, z) \leq C$, for some scalar $C > 0$, then for every S,

$$L_S(A(S)) \leq L_S(A(S)) + \lambda\|A(S)\|^2 \leq L_S(\mathbf{0}) + \lambda\|\mathbf{0}\|^2 = L_S(\mathbf{0}) \leq C.$$

Hence, Corollary also implies that

$$\mathbb{E}\Big[\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i)\Big] \leq \frac{48\beta C}{\lambda m}.$$

7.5 Subgradients

Lemma 7.5.1

Let S be an open convex set. A function $f : S \to \mathbb{R}$ is convex iff for every $\mathbf{w} \in S$ there exists $\mathbf{v}$ such that

$$\forall \mathbf{u} \in S, \quad f(\mathbf{u}) \geq f(\mathbf{w}) + \langle \mathbf{u} - \mathbf{w}, \mathbf{v}\rangle.$$

Definition 7.5.2: Subgradients

A vector $\mathbf{v}$ that satisfies the above Equation is called a subgradient of f at $\mathbf{w}$. The set of subgradients of f at $\mathbf{w}$ is called the differential set and denoted $\partial f(\mathbf{w})$.

7.5.1 Calculating Subgradients

Claim

If f is differentiable at $\mathbf{w}$ then $\partial f(\mathbf{w})$ contains a single element — the gradient of f at $\mathbf{w}$, $\nabla f(\mathbf{w})$.

Claim

Let $g(\mathbf{w}) = \max_{i\in[r]} g_i(\mathbf{w})$ for r convex differentiable functions $g_1, \ldots, g_r$. Given some $\mathbf{w}$, let $j \in \arg\max_i g_i(\mathbf{w})$. Then $\nabla g_j(\mathbf{w}) \in \partial g(\mathbf{w})$.

Lemma 7.5.3

Let A be a convex open set and let $f : A \to \mathbb{R}$ be a convex function. Then, f is ρ-Lipschitz over A iff for all $\mathbf{w} \in A$ and $\mathbf{v} \in \partial f(\mathbf{w})$ we have that

$$\|\mathbf{v}\| \leq \rho.$$

7.6 Learning with SGD

For simplicity, let us first consider the case of differentiable loss functions, and the risk function $L_{\mathcal{D}}$. The construction of the random vector $\mathbf{v}_t$ will be as follows: First, sample $z \sim \mathcal{D}$. Then, define $\mathbf{v}_t$ to be the gradient of the function $\ell(\mathbf{w}, z)$ with respect to $\mathbf{w}$, at the point $\mathbf{w}^{(t)}$. Then, by the linearity of the gradient we have

$$\mathbb{E}[\mathbf{v}_t \,|\, \mathbf{w}^{(t)}] = \mathbb{E}_{z\sim\mathcal{D}}[\nabla \ell(\mathbf{w}^{(t)}, z)] = \nabla \mathbb{E}_{z\sim\mathcal{D}}[\ell(\mathbf{w}^{(t)}, z)] = \nabla L_{\mathcal{D}}(\mathbf{w}^{(t)}).$$

The gradient of the loss function $\ell(\mathbf{w}, z)$ at $\mathbf{w}^{(t)}$ is therefore an unbiased estimate of the gradient of the risk function $L_{\mathcal{D}}(\mathbf{w}^{(t)})$ and is easily constructed by sampling a single fresh example $z \sim \mathcal{D}$ at each iteration t. The same argument holds for nondifferentiable loss functions. We simply let $\mathbf{v}_t$ be a subgradient of $\ell(\mathbf{w}, z)$ at $\mathbf{w}^{(t)}$. Then, for every $\mathbf{u}$ we have

$$\ell(\mathbf{u}, z) - \ell(\mathbf{w}^{(t)}, z) \geq \langle \mathbf{u} - \mathbf{w}^{(t)}, \mathbf{v}_t \rangle.$$

Taking expectation on both sides with respect to $z \sim \mathcal{D}$ and conditioned on the value of $\mathbf{w}^{(t)}$ we obtain

$$\begin{aligned} L_{\mathcal{D}}(\mathbf{u}) - L_{\mathcal{D}}(\mathbf{w}^{(t)}) &= \mathbb{E}[\ell(\mathbf{u}, z) - \ell(\mathbf{w}^{(t)}, z) \,|\, \mathbf{w}^{(t)}] \\ &\geq \mathbb{E}[\langle \mathbf{u} - \mathbf{w}^{(t)}, \mathbf{v}_t \rangle \,|\, \mathbf{w}^{(t)}] \\ &= \langle \mathbf{u} - \mathbf{w}^{(t)}, \mathbb{E}[\mathbf{v}_t \,|\, \mathbf{w}^{(t)}] \rangle. \end{aligned}$$

It follows that $\mathbb{E}[\mathbf{v}_t \,|\, \mathbf{w}^{(t)}]$ is a subgradient of $L_{\mathcal{D}}(\mathbf{w})$ at $\mathbf{w}^{(t)}$.

Stochastic Gradient Descent (SGD) for minimizing $L_{\mathcal{D}}(\mathbf{w})$

```
parameters: Scalar $\eta > 0$, integer $T > 0$
initialize: $\mathbf{w}^{(1)} = 0$
for $t = 1, 2, \ldots, T$:
    sample $z \sim \mathcal{D}$
    pick $\mathbf{v}_t \in \partial \ell(\mathbf{w}^{(t)}, z)$
    update $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \mathbf{v}_t$
output: $\bar{\mathbf{w}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{w}^{(t)}$
```

Corollary 7.6.1

Consider a convex–Lipschitz–bounded learning problem with parameters ρ, B. Then, for every $\epsilon > 0$, if we run the SGD method for minimizing $L_{\mathcal{D}}(\mathbf{w})$ with a number of iterations (i.e., number of examples)

$$T \geq \frac{B^2 \rho^2}{\epsilon^2}$$

and with $\eta = \sqrt{\frac{B^2}{\rho^2 T}}$, then the output of SGD satisfies

$$\mathbb{E}[L_{\mathcal{D}}(\bar{\mathbf{w}})] \leq \min_{\mathbf{w} \in \mathcal{H}} L_{\mathcal{D}}(\mathbf{w}) + \epsilon.$$

Note that the required sample complexity is of the same order of magnitude as the sample complexity guarantee for regularized loss minimization. In fact, the sample complexity of SGD is even better than regularized loss minimization by a factor of 8.

7.6.1 SGD for Regularized Loss Minimization

We have shown that SGD enjoys the same worst-case sample complexity bound as regularized loss minimization. However, on some distributions, regularized loss minimization may yield a better solution.

$$\min_{\mathbf{w}} \left(\frac{\lambda}{2} \|\mathbf{w}\|^2 + L_S(\mathbf{w}) \right).$$

$$\begin{aligned}
\mathbf{w}^{(t+1)} &= \mathbf{w}^{(t)} - \frac{1}{\lambda t}(\lambda \mathbf{w}^{(t)} + \mathbf{v}_t) \\
&= \left(1 - \frac{1}{t}\right) \mathbf{w}^{(t)} - \frac{1}{\lambda t}\mathbf{v}_t \\
&= \frac{t-1}{t}\mathbf{w}^{(t)} - \frac{1}{\lambda t}\mathbf{v}_t \\
&= \frac{t-1}{t}\left(\frac{t-2}{t-1}\mathbf{w}^{(t-1)} - \frac{1}{\lambda(t-1)}\mathbf{v}_{t-1}\right) - \frac{1}{\lambda t}\mathbf{v}_t \\
&= -\frac{1}{\lambda t}\sum_{i=1}^{t} \mathbf{v}_i.
\end{aligned}$$

If we assume that the loss function is ρ-Lipschitz, it follows that for all t we have $\|\mathbf{v}_t\| \leq \rho$ and therefore $\|\lambda \mathbf{w}^{(t)}\| \leq \rho$, which yields

$$\|\lambda \mathbf{w}^{(t)} + \mathbf{v}_t\| \leq 2\rho.$$

After performing T iterations we have that

$$\mathbb{E}[f(\bar{\mathbf{w}})] - f(\mathbf{w}^\star) \leq \frac{4\rho^2}{\lambda T}(1 + \log(T)).$$

7.7 Online Convex Optimization

```
Online Convex Optimization

definitions:
    hypothesis class $\mathcal{H}$ ;    domain $\mathcal{Z}$ ;    loss function $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$
assumptions:
    $\mathcal{H}$ is convex
    $\forall z \in \mathcal{Z},\ \ell(\cdot, z)$ is a convex function
for  $t = 1, 2, \ldots, T$
    learner predicts a vector $\mathbf{w}^{(t)} \in \mathcal{H}$
    environment responds with $z_t \in \mathcal{Z}$
    learner suffers loss $\ell(\mathbf{w}^{(t)}, z_t)$
```

The regret of an online algorithm with respect to a competing hypothesis, some vector $\mathbf{w}^\star \in \mathcal{H}$, is defined as

$$\text{Regret}_A(\mathbf{w}^\star, T) = \sum_{t=1}^{T} \ell(\mathbf{w}^{(t)}, z_t) - \sum_{t=1}^{T} \ell(\mathbf{w}^\star, z_t).$$

The regret of the algorithm relative to a set of competing vectors, $\mathcal{H}$, is defined as

$$\text{Regret}_A(\mathcal{H}, T) = \sup_{\mathbf{w}^\star \in \mathcal{H}} \text{Regret}_A(\mathbf{w}^\star, T).$$

Theorem 7.7.1

The Online Gradient Descent algorithm enjoys the following regret bound for every $\mathbf{w}^\star \in \mathcal{H}$,

$$\text{Regret}_A(\mathbf{w}^\star, T) \le \frac{\|\mathbf{w}^\star\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\|\mathbf{v}_t\|^2.$$

If we further assume that f_t is ρ-Lipschitz for all t, then setting $\eta = 1/\sqrt{T}$ yields

$$\text{Regret}_A(\mathbf{w}^\star, T) \le \frac{1}{2}(\|\mathbf{w}^\star\|^2 + \rho^2)\sqrt{T}.$$

If we further assume that $\mathcal{H}$ is B-bounded and we set $\eta = \frac{B}{\rho\sqrt{T}}$, then

$$\text{Regret}_A(\mathcal{H}, T) \le B\rho\sqrt{T}.$$

7.8 Online Perceptron Algorithm

```
                         Perceptron

initialize: $\mathbf{w}_1 = 0$

for  $t = 1, 2, \dots, T$
   receive $\mathbf{x}_t$
   predict $p_t = \text{sign}(\langle \mathbf{w}^{(t)}, \mathbf{x}_t \rangle)$
   if  $y_t\langle \mathbf{w}^{(t)}, \mathbf{x}_t \rangle \le 0$
      $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + y_t\mathbf{x}_t$
   else
      $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)}$
```

Theorem 7.8.1

Suppose that the Perceptron algorithm runs on a sequence $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_T, y_T)$ and let $R = \max_t \|\mathbf{x}_t\|$. Let $\mathcal{M}$ be the rounds on which the Perceptron errs and let $f_t(\mathbf{w}) = \mathbb{1}_{[t \in \mathcal{M}]}[1 - y_t\langle \mathbf{w}, \mathbf{x}_t \rangle]_+$. Then, for every $\mathbf{w}^\star$

$$|\mathcal{M}| \le \sum_t f_t(\mathbf{w}^\star) + R\|\mathbf{w}^\star\|\sqrt{\sum_t f_t(\mathbf{w}^\star) + R^2\|\mathbf{w}^\star\|^2}.$$

In particular, if there exists $\mathbf{w}^\star$ such that $y_t\langle \mathbf{w}^\star, \mathbf{x}_t \rangle \ge 1$ for all t, then

$$|\mathcal{M}| \le R^2\|\mathbf{w}^\star\|^2.$$

7.9 Strongly Convex Functions

Claim

If f is λ-strongly convex then for every $\mathbf{w}, \mathbf{u}$ and $\mathbf{v} \in \partial f(\mathbf{w})$ we have

$$\langle \mathbf{w} - \mathbf{u}, \mathbf{v} \rangle \geq f(\mathbf{w}) - f(\mathbf{u}) + \frac{\lambda}{2}\|\mathbf{w} - \mathbf{u}\|^2.$$

SGD for minimizing a λ-strongly convex function

```
**Goal:** Solve $\min_{\mathbf{w}\in\mathcal{H}} f(\mathbf{w})$
**parameter:** $T$
**initialize:** $\mathbf{w}^{(1)} = 0$
**for** $t = 1, \ldots, T$
    Choose a random vector $\mathbf{v}_t$ s.t. $\mathbb{E}[\mathbf{v}_t \,|\, \mathbf{w}^{(t)}] \in \partial f(\mathbf{w}^{(t)})$
    Set $\eta_t = 1/(\lambda t)$
    Set $\mathbf{w}^{(t+\frac{1}{2})} = \mathbf{w}^{(t)} - \eta_t \mathbf{v}_t$
    Set $\mathbf{w}^{(t+1)} = \arg\min_{\mathbf{w}\in\mathcal{H}} \|\mathbf{w} - \mathbf{w}^{(t+\frac{1}{2})}\|^2$
**output:** $\bar{\mathbf{w}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{w}^{(t)}$
```

Theorem 7.9.1

Assume that f is λ-strongly convex and that $\mathbb{E}[\|\mathbf{v}_t\|^2] \leq \rho^2$. Let $\mathbf{w}^\star \in \arg\min_{\mathbf{w}\in\mathcal{H}} f(\mathbf{w})$ be an optimal solution. Then,

$$\mathbb{E}[f(\bar{\mathbf{w}})] - f(\mathbf{w}^\star) \leq \frac{\rho^2}{2\lambda T}(1 + \log(T)).$$

Chapter 8

Rademacher Complexities

8.1 Rademacher Complexity

To simplify our notation, let us denote

$$\mathcal{F} \overset{\text{def}}{=} \ell \circ \mathcal{H} \overset{\text{def}}{=} \{\, z \mapsto \ell(h, z) : h \in \mathcal{H} \,\},$$

and given $f \in \mathcal{F}$, we define

$$L_{\mathcal{D}}(f) = \mathbb{E}_{z \sim \mathcal{D}}\big[f(z)\big], \qquad L_S(f) = \frac{1}{m} \sum_{i=1}^{m} f(z_i).$$

Definition 8.1.1: Representativeness

The representativeness of S with respect to $\mathcal{F}$ as the largest gap between the true error of a function f and its empirical error, namely,

$$\operatorname{Rep}_{\mathcal{D}}(\mathcal{F}, S) \overset{\text{def}}{=} \sup_{f \in \mathcal{F}} \big(L_{\mathcal{D}}(f) - L_S(f) \big).$$

Definition 8.1.2: Rademacher Complexity

Let $\mathcal{F} \circ S$ be the set of all possible evaluations a function $f \in \mathcal{F}$ can achieve on a sample S, namely,

$$\mathcal{F} \circ S = \{ (f(z_1), \dots, f(z_m)) : f \in \mathcal{F} \}.$$

Let the variables in σ be distributed i.i.d. according to $\Pr[\sigma_i = 1] = \Pr[\sigma_i = -1] = \frac{1}{2}$. Then, the Rademacher complexity of $\mathcal{F}$ with respect to S is defined as follows:

$$R(\mathcal{F} \circ S) \overset{\text{def}}{=} \frac{1}{m} \mathbb{E}_{\sigma \sim \{\pm 1\}^m} \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^{m} \sigma_i f(z_i) \right].$$

More generally, given a set of vectors $A \subset \mathbb{R}^m$, we define

$$R(A) \overset{\text{def}}{=} \frac{1}{m} \mathbb{E}_{\sigma} \left[\sup_{a \in A} \sum_{i=1}^{m} \sigma_i a_i \right].$$

Lemma 8.1.3

$$\mathbb{E}_{S\sim\mathcal{D}^m}\big[\mathrm{Rep}_{\mathcal{D}}(\mathcal{F},S)\big] \le 2\,\mathbb{E}_{S\sim\mathcal{D}^m}\big[R(\mathcal{F}\circ S)\big].$$

Proof. *Proof* Let $S' = \{z'_1,\dots,z'_m\}$ be another i.i.d. sample. Clearly, for all f, $L_D(f) = \mathbb{E}_{S'}[L_{S'}(f)]$. Therefore, for every $f\in\mathcal{F}$ we have

$$L_D(f) - L_S(f) = \mathbb{E}_{S'}[L_{S'}(f)] - L_S(f) = \mathbb{E}_{S'}[L_{S'}(f) - L_S(f)].$$

Taking supremum over $f\in\mathcal{F}$ of both sides, and using the fact that the supremum of expectation is smaller than expectation of the supremum we obtain

$$\sup_{f\in\mathcal{F}}\big(L_D(f) - L_S(f)\big) = \sup_{f\in\mathcal{F}}\mathbb{E}_{S'}\big[L_{S'}(f) - L_S(f)\big] \le \mathbb{E}_{S'}\Big[\sup_{f\in\mathcal{F}}\big(L_{S'}(f) - L_S(f)\big)\Big].$$

Taking expectation over S on both sides we obtain

$$\mathbb{E}_S\Big[\sup_{f\in\mathcal{F}}\big(L_D(f) - L_S(f)\big)\Big] \le \mathbb{E}_{S,S'}\Big[\sup_{f\in\mathcal{F}}\big(L_{S'}(f) - L_S(f)\big)\Big] = \frac{1}{m}\,\mathbb{E}_{S,S'}\Big[\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}\big(f(z'_i) - f(z_i)\big)\Big].$$

Next, we note that for each j, z_j and z'_j are i.i.d. variables. Therefore,

$$\mathbb{E}_{S,S'}\Big[\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}\big(f(z'_i) - f(z_i)\big)\Big] = \mathbb{E}_{S,S',\sigma}\Big[\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}\sigma_i\big(f(z'_i) - f(z_i)\big)\Big].$$

Finally,

$$\sup_{f\in\mathcal{F}}\sum_i \sigma_i\big(f(z'_i) - f(z_i)\big) \le \sup_{f\in\mathcal{F}}\sum_i \sigma_i f(z'_i) + \sup_{f\in\mathcal{F}}\sum_i(-\sigma_i)f(z_i),$$

and since the probability of σ is the same as the probability of $-\sigma$, the right-hand side of equation* can be bounded by

$$\mathbb{E}_{S,S',\sigma}\Big[\sup_{f\in\mathcal{F}}\sum_i \sigma_i f(z'_i) + \sup_{f\in\mathcal{F}}\sum_i \sigma_i f(z_i)\Big] = m\,\mathbb{E}_{S'}\big[\mathcal{R}(\mathcal{F}\circ S')\big] + m\,\mathbb{E}_S\big[\mathcal{R}(\mathcal{F}\circ S)\big] = 2m\,\mathbb{E}_S\big[\mathcal{R}(\mathcal{F}\circ S)\big].$$

□

Theorem 8.1.4

$$\mathbb{E}_{S\sim D^m}\big[L_D\big(\mathrm{ERM}_{\mathcal{H}}(S)\big) - L_S\big(\mathrm{ERM}_{\mathcal{H}}(S)\big)\big] \le 2\,\mathbb{E}_{S\sim D^m}\ \mathcal{R}(\ell\circ\mathcal{H}\circ S).$$

For any $h^\star\in\mathcal{H}$

$$\mathbb{E}_{S\sim D^m}\big[L_D\big(\mathrm{ERM}_{\mathcal{H}}(S)\big) - L_D(h^\star)\big] \le 2\,\mathbb{E}_{S\sim D^m}\ \mathcal{R}(\ell\circ\mathcal{H}\circ S).$$

If $h^\star = \arg\min_h L_D(h)$ then for each $\delta\in(0,1)$, with probability at least $1-\delta$ over the choice of S we have

$$L_D\big(\mathrm{ERM}_{\mathcal{H}}(S)\big) - L_D(h^\star) \le \frac{2\,\mathbb{E}_{S'\sim D^m}\ \mathcal{R}(\ell\circ\mathcal{H}\circ S')}{\delta}.$$

Lemma 8.1.5: McDiarmid's Inequality

Let V be some set and let $f : V^m \to \mathbb{R}$ be a function of m variables such that for some $c > 0$, for all $i \in [m]$ and for all $x_1, \ldots, x_m, x_i' \in V$ we have

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

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

$$\left| f(X_1, \ldots, X_m) - \mathbb{E}[f(X_1, \ldots, X_m)] \right| \leq c \sqrt{\frac{m}{2} \ln\left(\frac{2}{\delta}\right)}.$$

Theorem 8.1.6

Assume that for all z and $h \in \mathcal{H}$ we have that $|\ell(h, z)| \leq c$. Then,
1. With probability of at least $1 - \delta$, for all $h \in \mathcal{H}$,

$$L_D(h) - L_S(h) \;\leq\; 2\, \mathbb{E}_{S' \sim D^m}\, R(\ell \circ \mathcal{H} \circ S') \;+\; c \sqrt{\frac{2 \ln(2/\delta)}{m}}.$$

In particular, this holds for $h = \mathrm{ERM}_{\mathcal{H}}(S)$.
2. With probability of at least $1 - \delta$, for all $h \in \mathcal{H}$,

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

In particular, this holds for $h = \mathrm{ERM}_{\mathcal{H}}(S)$.
3. For any $h^\star$, with probability of at least $1 - \delta$,

$$L_D\big(\mathrm{ERM}_{\mathcal{H}}(S)\big) - L_D(h^\star) \;\leq\; 2\, R(\ell \circ \mathcal{H} \circ S) \;+\; 5c \sqrt{\frac{2 \ln(8/\delta)}{m}}.$$

8.1.1 Rademacher Calculus

Lemma 8.1.7

For any $A \subset \mathbb{R}^m$, scalar $c \in \mathbb{R}$, and vector $\mathbf{a}_0 \in \mathbb{R}^m$, we have

$$R\big(\{\, c\,\mathbf{a} + \mathbf{a}_0 : \mathbf{a} \in A \,\}\big) \leq |c|\, R(A).$$

Lemma 8.1.8

Let A be a subset of $\mathbb{R}^m$ and let

$$A' = \Big\{ \sum_{j=1}^{N} \alpha_j\, \mathbf{a}^{(j)} \;:\; N \in \mathbb{N},\ \forall j,\ \mathbf{a}^{(j)} \in A,\ \alpha_j \geq 0,\ \|\boldsymbol{\alpha}\|_1 = 1 \Big\}.$$

Then,

$$R(A') = R(A).$$

Lemma 8.1.9: Massart lemma

Let $A = \{\mathbf{a}_1, \dots, \mathbf{a}_N\}$ be a finite set of vectors in $\mathbb{R}^m$. Define $\bar{\mathbf{a}} = \frac{1}{N}\sum_{i=1}^N \mathbf{a}_i$. Then,

$$R(A) \;\leq\; \max_{\mathbf{a}\in A} \|\mathbf{a} - \bar{\mathbf{a}}\| \sqrt{\frac{2\log(N)}{m}}.$$

Lemma 8.1.10: Contraction lemma

For each $i \in [m]$, let $\phi_i : \mathbb{R} \to \mathbb{R}$ be a ρ-Lipschitz function, namely for all $\alpha, \beta \in \mathbb{R}$ we have $|\phi_i(\alpha) - \phi_i(\beta)| \leq \rho\,|\alpha - \beta|$. For $\mathbf{a} \in \mathbb{R}^m$ let $\phi(\mathbf{a})$ denote the vector $(\phi_1(a_1), \dots, \phi_m(a_m))$. Let $\phi \circ A = \{\phi(\mathbf{a}) : \mathbf{a} \in A\}$. Then,

$$R(\phi \circ A) \leq \rho\, R(A).$$

8.2 Rademacher Complexity of Linear Classes

To simplify the derivation we first define the following two classes:

$$\mathcal{H}_1 = \{\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x}\rangle : \|\mathbf{w}\|_1 \leq 1\}, \qquad \mathcal{H}_2 = \{\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x}\rangle : \|\mathbf{w}\|_2 \leq 1\}.$$

Lemma 8.2.1

Let $S = (\mathbf{x}_1, \dots, \mathbf{x}_m)$ be vectors in a Hilbert space. Define

$$\mathcal{H}_2 \circ S = \big\{(\langle \mathbf{w}, \mathbf{x}_1\rangle, \dots, \langle \mathbf{w}, \mathbf{x}_m\rangle) : \|\mathbf{w}\|_2 \leq 1\big\}.$$

Then,

$$R(\mathcal{H}_2 \circ S) \;\leq\; \frac{\max_i \|\mathbf{x}_i\|_2}{\sqrt{m}}.$$

Lemma 8.2.2

Let $S = (\mathbf{x}_1, \dots, \mathbf{x}_m)$ be vectors in $\mathbb{R}^n$. Then,

$$R(\mathcal{H}_1 \circ S) \;\leq\; \max_i \|\mathbf{x}_i\|_\infty \sqrt{\frac{2\log(2n)}{m}}.$$

8.3 Generalization Bounds for SVM

Theorem 8.3.1

Suppose that D is a distribution over $\mathcal{X} \times \mathcal{Y}$ such that with probability 1 we have that $\|\mathbf{x}\|_2 \leq R$. Let $\mathcal{H} = \{\mathbf{w} : \|\mathbf{w}\|_2 \leq B\}$ and let $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$ be a loss function of the form given in Equation (26.18) such that for all $y \in \mathcal{Y}$, $a \mapsto \phi(a, y)$ is a ρ-Lipschitz function and such that $\max_{a\in[-BR,BR]} |\phi(a, y)| \leq c$. Then, for any $\delta \in (0, 1)$, with probability of at least $1 - \delta$ over the choice of an i.i.d. sample of size m,

$$\forall\, \mathbf{w} \in \mathcal{H}, \qquad L_D(\mathbf{w}) \;\leq\; L_S(\mathbf{w}) \;+\; \frac{2\rho BR}{\sqrt{m}} \;+\; c\sqrt{\frac{2\ln(2/\delta)}{m}}\,.$$

Theorem 8.3.2

Consider a distribution D over $\mathcal{X} \times \{\pm 1\}$ such that there exists some vector $\mathbf{w}^\star$ with $\mathbb{P}_{(\mathbf{x},y)\sim D}[y \langle \mathbf{w}^\star, \mathbf{x} \rangle \geq 1] = 1$ and such that $\|\mathbf{x}\|_2 \leq R$ with probability 1. Let $\mathbf{w}_S$ be the output of Equation (26.19). Then, with probability of at least $1 - \delta$ over the choice of $S \sim D^m$, we have that

$$\mathbb{P}_{(\mathbf{x},y)\sim D}\big[y \neq \operatorname{sign}(\langle \mathbf{w}_S, \mathbf{x} \rangle)\big] \leq \frac{2R\|\mathbf{w}^\star\|}{\sqrt{m}} + \big(1 + R\|\mathbf{w}^\star\|\big)\sqrt{\frac{2\ln(2/\delta)}{m}}.$$

Theorem 8.3.3

Assume that the conditions of Theorem 26.13 hold. Then, with probability of at least $1 - \delta$ over the choice of $S \sim D^m$, we have that

$$\mathbb{P}_{(\mathbf{x},y)\sim D}\big[y \neq \operatorname{sign}(\langle \mathbf{w}_S, \mathbf{x} \rangle)\big] \;\leq\; \frac{4R\|\mathbf{w}_S\|}{\sqrt{m}} \;+\; \sqrt{\frac{\ln\left(\frac{4\log_2(\|\mathbf{w}_S\|)}{\delta}\right)}{m}}\,.$$

8.4 Generalization Bounds for Predictors with Low $\mathcal{L}_1$ Norm

Theorem 8.4.1

Suppose that D is a distribution over $\mathcal{X} \times \mathcal{Y}$ such that with probability 1 we have that $\|\mathbf{x}\|_\infty \leq R$. Let $\mathcal{H} = \{\mathbf{w} \in \mathbb{R}^d : \|\mathbf{w}\|_1 \leq B\}$ and let $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$ be a loss function of the form given in Equation (26.18) such that for all $y \in \mathcal{Y}$, $a \mapsto \phi(a, y)$ is a ρ-Lipschitz function and such that $\max_{a\in[-BR,BR]} |\phi(a, y)| \leq c$. Then, for any $\delta \in (0, 1)$, with probability of at least $1 - \delta$ over the choice of an i.i.d. sample of size m,

$$\forall\, \mathbf{w} \in \mathcal{H}, \qquad L_D(\mathbf{w}) \;\leq\; L_S(\mathbf{w}) \;+\; 2\rho BR\sqrt{\frac{2\log(2d)}{m}} \;+\; c\sqrt{\frac{2\ln(2/\delta)}{m}}\,.$$

Chapter 9

PAC-Bayes

9.1 PAC-Bayes Bounds

As in the MDL paradigm, we define a hierarchy over hypotheses in our class $\mathcal{H}$. Now, the hierarchy takes the form of a prior distribution over $\mathcal{H}$. That is, we assign a probability (or density if $\mathcal{H}$ is continuous) $P(h) \geq 0$ for each $h \in \mathcal{H}$ and refer to $P(h)$ as the prior score of h. Following the Bayesian reasoning approach, the output of the learning algorithm is not necessarily a single hypothesis. Instead, the learning process defines a posterior probability over $\mathcal{H}$, which we denote by Q. In the context of a supervised learning problem, where $\mathcal{H}$ contains functions from $\mathcal{X}$ to $\mathcal{Y}$, one can think of Q as defining a randomized prediction rule as follows. Whenever we get a new instance $\mathbf{x}$, we randomly pick a hypothesis $h \in \mathcal{H}$ according to Q and predict $h(\mathbf{x})$. We define the loss of Q on an example z to be

$$\ell(Q, z) \overset{\text{def}}{=} \mathbb{E}_{h\sim Q}, [\ell(h, z)].$$

By the linearity of expectation, the generalization loss and training loss of Q can be written as

$$L_D(Q) \overset{\text{def}}{=} \mathbb{E} h \sim Q[L_D(h)] \qquad \text{and} \qquad L_S(Q) \overset{\text{def}}{=} \mathbb{E} h \sim Q[L_S(h)].$$

The following theorem tells us that the difference between the generalization loss and the empirical loss of a posterior Q is bounded by an expression that depends on the Kullback–Leibler divergence between Q and the prior distribution P. The Kullback–Leibler is a natural measure of the distance between two distributions. The theorem suggests that if we would like to minimize the generalization loss of Q, we should jointly minimize both the empirical loss of Q and the Kullback–Leibler distance between Q and the prior distribution.

Theorem 9.1.1

Let D be an arbitrary distribution over an example domain Z. Let $\mathcal{H}$ be a hypothesis class and let $\ell : \mathcal{H} \times Z \to [0, 1]$ be a loss function. Let P be a prior distribution over $\mathcal{H}$ and let $\delta \in (0, 1)$. Then, with probability of at least $1-\delta$ over the choice of an i.i.d. training set $S = \{z_1, \ldots, z_m\}$ sampled according to D, for all distributions Q over $\mathcal{H}$ (even such that depend on S), we have

$$L_D(Q) \;\leq\; L_S(Q) \;+\; \sqrt{\frac{D(Q\|P) + \ln\bigl(m/\delta\bigr)}{2\,(m-1)}}\,,$$

where

$$D(Q\|P) \;\overset{\text{def}}{=}\; \mathbb{E}_{h\sim Q}\bigl[\ln\bigl(Q(h)/P(h)\bigr)\bigr]$$

is the Kullback–Leibler divergence.