ELAD HAZAN'S UNSUPERVISED LEARNING COURSE

ERIC NASLUND

1. Introduction

Week 1, February 8th 2017.

1.1. **Supervised Learning.** Input data: Given a distribution D over $X \times Y$, where X is the domain and Y is the label set, our goal is to predict y correctly given x. Statistical/PAC learning is given a hypothesis class

$$\mathcal{H} : X \to Y$$

the learning problem $(D, X, Y, \mathcal{H})$ is learnable if and only if there exists an algorithm and a function

$$m : (\epsilon, \delta) \to m(\epsilon, \delta) \in \mathbb{N}^+$$

such that for all distributions D, after seeing $m(\epsilon, \delta)$ examples from D, $\{(x, y)\}$ returns a hypothesis h such that

$$\mathbb{E}_{(x,y)\sim D}\left[h(x) \neq y\right] \leq \min_{h^* \in \mathcal{H}} \mathbb{E}_{D\sim(x,y)}\left[h^*(x) \neq y\right] + \epsilon$$

with probability $1 - \delta$.

- Completely human independent
- Very General
- Allows for any learning algorithm, which gives room for efficient algorithms
- Very practical

For a more in depth discussion of this theory, see a course on theoretical machine learning, such as COS511.

1.2. **Unsupervised Learning.** We begin with several examples of unsupervised learning.

Example 1. Karl Pearson was studying biometric data concerning a new species of crab that he discovered. He measured their forehead breadth to body length ratios, expecting to find a normal distribution. However, he found a very skewed distribution, and deduced that this was actually the sum of two different normal distributions, and in fact, he had discovered two species of crabs.

Example 2. Anomally detection consists of finding anomalies in the data. It's not a well defined problem, and so it is a natural place to use unsupervised learning.

Example 3. Newtonian mechanics and concise formulation of Keplers laws.

This leads us to our wish list:

- $x \sim D$, $x \in \mathbb{R}^d$.
- A compact representation
- A representation that makes it easier to handle later supervised learning.
- Generalization

Date: February 15, 2017.

- Efficient algorithm based on relaxation

1.3. **Learning representations.**

- Perfect recoverability is necessary even for trivial families of future tasks
- Compression (even mild compression) is incomputable in general. There are a host of hardness results here - if you want to do compression, even detecting if there exists a compression algorithm that does better is equivalent to certain hardness properties of one way functions.

Remark 1. Many of the hardness results are for Boolean functions, and if we are working with real valued functions and allow errors, then we can obtain results.

Let's look at recoverability first.

(1) If we measure performance by a Lipschitz function, then "lossy" compression "makes sense."
(2) We will measure compression with respect to a family of hypotheses.

Definition 1. see [1] An unsupervisded problem $x \sim D$, $x \in \mathbb{R}^d$ is learnable with respect to a hypothesis class

$$\mathcal{H} = \{(f,g) : \ fX \to Y \text{ and } g : Y \to X\}$$

where $Y \subset \mathbb{R}^n$ and ℓ is a loss function

$$\ell : X \times X \to \mathbb{R}$$

if there exists

$$m(\epsilon, \delta) \to \mathbb{N}$$

such that for every ϵ, δ there exists an efficient algorithm such that after seeing $m(\epsilon, \delta)$ examples, it yields $\tilde{f}$, $\tilde{g}$ such that

$$\mathbb{E}_{x \sim D}\left[\left|x - \tilde{g} \circ \tilde{f}(x)\right| + \left|\tilde{f}(x)\right|\right] \leq \min_{(f,g) \in \mathcal{H}} \mathbb{E}_{x \sim D}\left[|x - g \circ f(x)| + |f(x)|\right] + \epsilon$$

with probability $1 - \delta$.

Example 4. Suppose that $x \in \mathbb{R}^d$, and $y \subset \mathbb{R}^k$ where $k \ll d$. Let $\ell(x, z) = |x - z|_2^2$ be the squared Euclidean distance. Define

$$\mathcal{H} = \left\{ \begin{array}{c} f(x) = Ax \text{ where } A \in \mathbb{R}^{k \times n}, \ |A|_2^2 \leq \dots \\ g(y) = By, \ B \in \mathbb{R}^{n \times k}, \ |B|_2^2 \leq \dots \end{array} \right\}.$$

This gives rise to PCA, which is

$$\begin{array}{c} \min \\ A \\ \text{rank}(A) = k \end{array} \qquad \left|x - A^t A x\right|_2^2.$$

Theorem. *For k-PCA,*

$$m(\epsilon, \delta) = O\left(\frac{k}{\epsilon^2} \log \frac{1}{\delta}\right).$$

Remark 2. The norm bound for $|A|_2^2$ and $|B|_2^2$ in the definition of $\mathcal{H}$ was not specified. This is an important sublety, and we will go into detail in future lectures. Indeed, the error in the theorem above depends on the norm bound, and so the above theorem has a specific norm bound in mind.

Example 5. *(k-means) Given $f : x \to \mu_i$,*

$$f_{\{\mu_i\}_{i=1,\ldots,k}}$$

and $g(\mu_i) = \mu_i$.

Example 6. *(Dictionary Learning)* Given $x \to y$ such that

$$f_A(x) = \Big\{ y \quad \text{where } y = \text{argmin}_{|z|_0 \leq k} |Az - x|_2^2$$

where $|z|_0$ is the number of non-zero entries, then $g_A(y) = Ay$ where $A \in \mathbb{R}^{n \times d}$.

$$m(\epsilon, \delta) = O\left(\min\left(\frac{nd}{\epsilon^2} \log \frac{1}{\delta}, \ \frac{dk \log N}{\epsilon^2} \right) \right).$$

We note that the second term in the minimum above is a stronger bound than the first, but that the first is independent of k.

1.4. **ERM and Rademacher Complexity.** *(Empirical Risk Minimization)* Take a set S of examples $\{x_i\}_{i=1}^m$ where $x_i \in \mathbb{R}^d$, $|S| = m$. Define

$$h_{ERM}(S) = \text{ERM}(S) = \text{argmin}_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^m \ell\left(h, x_i\right),$$

and let

$$\text{loss}_S(h) = \frac{1}{m} \sum_{i=1}^m \ell\left(h, x_i\right).$$

Here $\ell\left(h, x_i\right) = \text{loss}(h, x_i)$ is any Lipschitz function, but in this course we will be particularly interested in the reconstruction error

$$\ell(h, x_i) = |x - g \circ f(x)|,$$

or

$$\ell(h, x_i) = |x - g \circ f(x)| + |f(x)|$$

depending on the application. Let

$$\text{loss}(h) = \mathbb{E}_{x \sim D}\left[\text{loss}(h, x)\right].$$

Then we say that $(X, D, \mathcal{H})$ is learnable if

$$\text{loss}_S(h_{ERM}) \leq \min_{h^* \in H} \text{loss}(h^*) + \epsilon$$

with probability $1 - \delta$ for m large enough.

Definition 2. The Rademacher complexity of a hypothesis class $\mathcal{H}$ with respect to a sample S is

$$R_S(\mathcal{H}) = \mathbb{E}\left[\sum_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{,} X_i \text{loss}(h, x_i) \right]$$

where $\{X_i\}_{i=1}^m$ are independent Bernoulli random variables taking the values -1 and 1 each with probability $\frac{1}{2}$. Define

$$R_m\left(\mathcal{H}\right) = \mathbb{E}_{S \sim D^m}\left[R_S\left(\mathcal{H}\right)\right].$$

We want for every hypothesis,

$$\text{loss}_S(h) \approx \text{loss}(h).$$

The Rademacher complexity specifically captures when this condition happens.

Theorem 1. *With probability at least* $1-\delta$*, for* $m \geq C$ *for some constant* C*,*

$$\operatorname{loss}(h_{ERM}) \leq \operatorname{loss}(h^*) + R_m(\mathcal{H}) + \sqrt{\frac{\log \frac{1}{\delta}}{m}}.$$

Proof. Define

$$\Phi(S) = \sup_{h \in \mathcal{H}} \left(\operatorname{loss}(h) - \operatorname{loss}_S(h)\right).$$

Observe that

$$\left|\Phi(S) - \mathbb{E}_{S \sim D^m}\left[\Phi(S)\right]\right| \leq \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{m}}$$

with probability $1-\delta$. This follows from Markov-type results in probability theory. To be precise you have to use Martingale sequences. Next, we will show that

$$\mathbb{E}_S \Phi(S) \leq 2R_m(\mathcal{H}).$$

From this and the previous inequality, it follows that

$$\left|\operatorname{loss}_S(H) - \operatorname{loss}(h)\right| \leq R_m(\mathcal{H}) + \sqrt{\frac{\log \frac{1}{\delta}}{m}}$$

from which the result immediately follows. To prove this, notice that

$$\begin{aligned}
\mathbb{E}_S\left[\Phi(S)\right] &= \mathbb{E}_S\left[\sup_h \left(\operatorname{loss}(h) - \operatorname{loss}_S(h)\right)\right] \\
&= \mathbb{E}_S\left[\sup_h \left(\mathbb{E}_{S' \sim D^m} \operatorname{loss}_{S'}(h) - \operatorname{loss}_S(h)\right)\right] \\
&\leq \mathbb{E}_{S,S' \sim D^m}\left[\sup_h \left(\operatorname{loss}_{S'}(h) - \operatorname{loss}_S(h)\right)\right] \\
&\leq \mathbb{E}_{S,S'}\left[\sup_{h \in \mathcal{H}} \left(\frac{1}{m}\sum_{i=1} \left(\operatorname{loss}(h, x_i') - \operatorname{loss}(h, x_i)\right)\right)\right].
\end{aligned}$$

Let $\{X_i\}_{i=1}^n$ be independent uniform random variables on $\{-1, 1\}$. Then the above equals

$$\mathbb{E}_{S,S'}\left[\sup_{h \in \mathcal{H}} \left(\frac{1}{m}\sum_{i=1} X_i \left(\operatorname{loss}(h, x_i') - \operatorname{loss}(h, x_i)\right)\right)\right],$$

and this is at most

$$\mathbb{E}_{S,S'}\left[\sup_{h \in \mathcal{H}} \left(\frac{1}{m}\sum_{i=1} X_i \operatorname{loss}(h, x_i')\right)\right] + \mathbb{E}_{S,S'}\left[\sup_{h \in \mathcal{H}} \left(\frac{1}{m}\sum_{i=1} -X_i \operatorname{loss}(h, x_i)\right)\right]$$

and this equals $2R_m(\mathcal{H})$. □

Next class we will see that the Rademacher complexity can be computed easily, and is indeed an interesting function.

References

[1] Elad Hazan and Tengyu Ma, A Non-generative Framework and Convex Relaxations for Unsupervised Learning, arXiv:1610.01132, 2016.