CSC 588: Machine learning theory **Spring 2022**

CSC 588: Machine learning theory | **Spring 2022**

Lecture 4: Agnostic PAC Learning; Hoeffding's Inequality

Lecturer: Chicheng Zhang *Scribe: Renee Zhang*

1 Review of last lecture

a. PAC learning
b. Consistency algorithm
c. Agnostic PAC model
Realizability is difficult to satisfy, so Agnostic is more realistic.
Population of two classes may have overlap;
Even they are perfectly separated, our hypothesis class might be too simple to delineate the decision boundary.

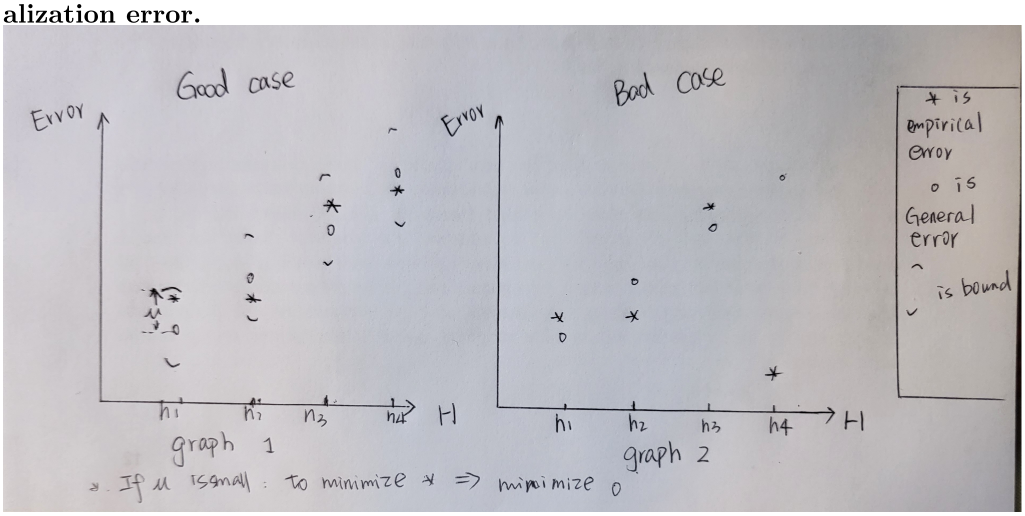
2 Leftover question from last lecture

How to design algorithm for agnostic PAC learning?
- Consistency model(introduced last lecture) is not satisfied (reason: may not have a 100 % correct classifier).
- Empirical risk (another term for training data) minimization(ERM):
Given a training data set S, return $\hat{h}$ such that it has smallest training error among all classifiers in $\mathcal{H}$:

$$\hat{h} = \operatorname*{argmin}_{h \in \mathcal{H}} \operatorname{err}(h, S)$$

3 Analysis of ERM

To analyze/evaluate ERM, we need to quantify how close the empirical error is to the generalization error.



In the above graphs, $\mathcal{H}$ has 4 classifiers, and we sort them by their generalization errors. The empirical error lies in a small range (width = μ) around it's generalization error. Graph 1 shows a good case, graph 2 shows a bad case. Our argument is: bad cases happen with low probability.

Observation: If all empirical errors are concentrated around their respective generalization errors, then ERM performs well. How well it performs depends on how close this concentration is.
Proof:
Let $h^* = argmin_{h \in \mathcal{H}} \operatorname{err}(h, D)$, and its corresponding error rate is ν.
Step 1 : $\hat{h}$ performs well in training data.

$$\operatorname{err}(\hat{h}, S) \leq \operatorname{err}(h^*, S) \leq \operatorname{err}(h^*, D) + \mu = \nu + \mu.$$

Step 2 : $\hat{h}$ performs well under the distribution D.

$$\operatorname{err}(\hat{h}, D) \leq \operatorname{err}(\hat{h}, S) + \mu \leq \nu + 2\mu.$$

3.1 Concentration of measure

Given a set of iid r.v. $Z_1, ...Z_n$, empirical mean $\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i$ concentrates around expectation of Z_1 with high (overwhelming) probability $1 - f(n)$, where $f(n) \to 0$ as $n \to \infty$.
e.g. Given classifier $h \in \mathcal{H}$, and S (training set)

$$\operatorname{err}(h, S) = \frac{1}{n}\sum_{i=1}^{n} I(h(x) \neq y),$$

we can let

$$Z_i \sim \text{Bernoulli}(\operatorname{err}(h, D)), \mu = \operatorname{err}(h, D)$$

3.2 Hoeffding's inequality

Theorem 1. *Suppose $Z_1, ...Z_n$ are i.i.d and $\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i$, let $\mu = E(Z_i)$, then for all $\epsilon > 0$,*

$$P(|\bar{Z} - \mu| > \epsilon) \leq 2\exp(-\frac{2n\epsilon^2}{(b-a)^2}).$$

Let $2\exp(-\frac{2n\epsilon^2}{(b-a)^2}) = \delta$, we have, equivalently, for any $\delta > 0$,

$$P\left(|\bar{Z} - \mu| > (b-a)\sqrt{\frac{\ln\frac{2}{\delta}}{2n}}\right) \leq \delta$$

This allows one to set δ as a small number, say $\delta = 10^{-5}$, with reasonably tight concentration results.

3.3 Chebyshev's inequality

To appreciate Hoeffding's Inequality, let us examine the consequence of applying (the familiar) Chebyshev's inequality to establish sample mean concentration.

$$P(|Y - EY| \geq a) \leq \frac{\operatorname{var}(Y)}{a^2}$$

Let $Y = \bar{Z}$, $\mathbb{E}Y = \mu$, and recall that for independent variable X_1, X_2, $\operatorname{var}(aX_1) = a^2\operatorname{var}(X_1)$, $\operatorname{var}(X_1 + X_2) = \operatorname{var}(X_1) + \operatorname{var}(X_2)$.We have,

$$\operatorname{var}(\bar{Z}) = \frac{1}{n^2}\operatorname{var}(Z_1 + ... + Z_n) = \frac{1}{n}\operatorname{var}(Z_1) = \frac{1}{n}\mathbb{E}[(Z_1 - \mu)^2] \leq \frac{(b-a)^2}{n}$$

Then:

$$P(|\bar{Z}-\mu| \geq \epsilon) \leq \frac{(b-a)^2}{n\epsilon^2}$$

Equivalently, for any $\delta > 0$,:

$$P(|\bar{Z}-\mu| \geq (b-a)\sqrt{\frac{\frac{1}{\delta}}{n}}) \leq \delta$$

Comparison:
Hoeffding's inequality decrease exponentially in sample size;
Chabyshev's inequality decrease polynomial in sample size.
Hoeffding inequality is better.

4 Proof of Hoeffding's Inequality

The proof uses moment generating functions:

Definition 2. *Consider a r.v X, define its moment generating function (mgf) to be $\phi_X(\lambda) = \mathbb{E}[e^{\lambda X}]$.*

Definition 3. *r.v X is said to be σ^2-Sub-Gaussian (SG), if for any $\lambda \in \mathbb{R}$:*

$$\mathbb{E}[e^{\lambda(x-\mu)}] \leq e^{\frac{\sigma^2}{2}}$$

Note that is X is Gaussian distributed, the above is an equality.
In order to prove Hoeffding's Inequality, we will take the following steps:
a.Prove that each $Z_i s$ are Sub-Gaussian with $\sigma^2 = \frac{(b-a)^2}{4}$;
b.Prove that independent sums of Sub-Gaussian r.v.s are Sub-Gaussian;
c.Prove that Sub-Gaussian distributions have light probability tail.
We start the proof of the easiest steps in the following subsection.

4.1 Proof of c

Lemma 4. *If X is σ^2-Sub-Gaussian then for any $\epsilon > 0$,*

$$P(|X-\mu| > \epsilon) \leq 2e^{-\frac{\epsilon^2}{2\sigma^2}}$$

Proof:

$$P(|X-\mu| > \epsilon) = P(X-\mu \leq -\epsilon) + P(X-\mu \geq \epsilon)$$

Let $\lambda > 0$, the second term:

$$P(X-\mu \geq \epsilon) = P(e^{\lambda(x-\mu)} \geq e^{\lambda\epsilon}) \leq \frac{\mathbb{E}(e^{\lambda(x-\mu)})}{e^{\lambda\epsilon}} \leq e^{-\lambda\epsilon + \frac{\sigma^2\lambda^2}{2}}$$

Chose λ to minimize the bound: $-\epsilon + \frac{\sigma^2}{2}2\lambda = 0$ We have: $\lambda = \frac{\epsilon}{\sigma^2}$ Such that:

$$P(X-\mu \geq \sigma) \leq \exp(-\frac{\epsilon^2}{2\sigma^2})$$

Use the same logic we can bound the first term.

4.2 Proof of b

Lemma 5. *If $X_1 ... X_n$ are independent, for every i, X_i is σ^2 - Sub-Gaussian for all $a_1 ... a_2 \in \mathbb{R}$ $\Sigma_{i=1}^n a_i x_i$ is $\Sigma_{i=1}^n a_i^2 \sigma_i^2$ Sub-Gaussian.*

To prove this lemma, we start with proving two special cases:
b(1). aX_1 is $a^2\sigma_1^2$ Sub-Gaussian;
b(2). $X_1 + X_2$ is $\sigma_1^2 + \sigma_2^2$ Sub-Gaussian.
Proof of b(1):
$\mathbb{E}[aX_1] = a\mu_1$, where $\mu_1 = E[X_1]$

$$\mathbb{E}[e^{\lambda(aX_1 - a\mu_1)}] = \mathbb{E}[e^{\lambda a(X_1 - \mu_1)}] \leq \exp\left(\frac{\sigma_1^2 \lambda^2 a^2}{2}\right)$$

Proof of b(2):
To be continued...

4.3 Proof of a

To be continued...