LECTURE 23

Last time

- PAC learning and VC-dimension
- Sauer Lemma

Today

- PAC learning and VC-dimension
- The sample complexity of PAC learning

Project Reports are due April 24

PAC Learning [Valiant 84]

PAC means ``Probably Approximately Correct''

- $\mathcal{C}$ is a class of functions of the form $f: \Omega \to \{0,1\}$.
- $\mathcal{D}$ is a distribution over Ω.
- The learner $\boldsymbol{\mathcal{L}}$ is given parameters $\varepsilon, \delta \in (0,1)$ and a set S of m *examples* drawn i.i.d. from $\mathcal{D}$ and labeled with a function $f \in \mathcal{C}$:
$$\{(x, f(x)) : x \in S\}.$$
- Goal of $\boldsymbol{\mathcal{L}}$: to find a *hypothesis* $h \in \mathcal{C}$ with error less than ε:
$$err_{\mathcal{D}}(h) := \Pr_{x \sim \mathcal{D}}[f(x) \neq h(x)].$$
- An algorithm $\boldsymbol{\mathcal{L}}$ is a PAC-learner for a class $\mathcal{C}$ if the probability it returns a hypothesis h with $err_{\mathcal{D}}(h) \leq \varepsilon$ is at least $1 - \delta$.

The probability is taken over distribution $\mathcal{D}$ and the coins of $\boldsymbol{\mathcal{L}}$.

ERM

ERM stands for ``empirical risk minimization''

- Empirical error (or empirical risk) of a hypothesis h is
$$err_S(h) := \Pr_{x \sim S}[f(x) \neq h(x)]$$
- An empirical risk minimizer is a hypothesis that has the smallest err_S among all hypothesis in the class, i.e., it mislabels the smallest number of examples in S.
- We will see that returning an ERM hypothesis is a great strategy for the learner.

Big Question

For a given class $\mathcal{C}$,

what sample size m

is sufficient for PAC learning?

Vapnik-Chervonenkis (VC) Dimension

- We will think of functions $f \in \mathcal{C}$ as indicator functions for sets.
- For this part, we will equate them with sets. I.e., now $f \subseteq \Omega$.
- For a finite set $S \subseteq \Omega$, the projection of $\mathcal{C}$ onto S is
$$\Pi_{\mathcal{C}}(S) \coloneqq \{h \cap S : h \in \mathcal{C}\}$$

In the old notation, this is the set of possible labelings of S by hypotheses from $\mathcal{C}$

- A set S is shattered by $\mathcal{C}$ if $|\Pi_{\mathcal{C}}(S)| = 2^{|S|}$, i.e., no labeling is ruled out.
- Note that if S is shattered by $\mathcal{C}$, then so is every subset of S.
- The VC dimension of a class $\mathcal{C}$ is the size of the largest set shattered by $\mathcal{C}$:
$$VC(\mathcal{C}) \coloneqq \max\{|S| : S \text{ shattered by } \mathcal{C}\}$$
- Let $\Pi_{\mathcal{C}}(m) \coloneqq \max_{S \subseteq \Omega, |S| = m} \{|\Pi_{\mathcal{C}}(S)|\}$, i.e., the maximum size of a projection of $\mathcal{C}$ for an m-element set.

Sauer's Lemma

Sauer's Lemma [Vapnik Chervonenkis 71]

Let $\mathcal{C}$ be a class of Boolean functions and $d = VC(\mathcal{C}) < \infty$.

Then $\Pi_{\mathcal{C}}(m) \le \binom{m}{\le d} \le \left(\frac{em}{d}\right)^d = O\left(m^d\right)$

Sample complexity of PAC learning

Theorem [Vapnik Chervonenkis]

Let $\mathcal{L}$ be an algorithm that outputs an ERM hypothesis. Then $\mathcal{L}$ is a PAC learner for class $\mathcal{C}$ if the number of samples it gets satisfies

$$m \geq \frac{2}{\varepsilon}\left(\log_2\left(\Pi_{\mathcal{C}}(2m)\right) + \log_2 \frac{2}{\delta}\right).$$

Want to show: $\Pr[A] \leq \delta$.

Proof: We define several bad events.

1. Failure event of the algorithm, expressed as a function of sample $S \sim \mathcal{D}^m$:

   $A = A(S)$: $\exists h \in \mathcal{C}$ such that $err_S(h) = 0$, but $err_{\mathcal{D}}(h) > \varepsilon$

2. Another event (that, as we will show, has related probability to $\Pr[A]$), expressed as a function of two i.i.d. samples, S and S', of size m:

   $B = B(S, S')$: $\exists h \in \mathcal{C}$ such that $err_S(h) = 0$, but $err_{S'}(h) > \varepsilon/2$

3. Event B_σ parameterized by $\sigma \in \{0,1\}^m$ and expressed as a function of two i.i.d. samples, $S = \{x_1, \dots, x_m\}$ and $S' = \{x_1', \dots, x_m'\}$.

   Define $T = \{z_1, \dots, z_m\}$ and $T' = \{z_1', \dots, z_m'\}$,

   where $z_i = x_i$ and $z_i' = x_i'$ if $\sigma_i = 1$; and $z_i = x_i'$ and $z_i' = x_i$, otherwise.

   $B_\sigma = B(S, S')$: $\exists h \in \mathcal{C}$ such that $err_T(h) = 0$, but $err_{T'}(h) > \varepsilon/2$

Relating the probabilities of A and B

$$\Pr[B] \geq \Pr[B|A] \cdot \Pr[A]$$

Law of total probability

Claim

If $m \geq \frac{8}{\varepsilon}$ then $\Pr_{s,s' \sim \mathcal{D}^m}[B|A] \geq \frac{1}{2}$

Event A: $\exists h \in \mathcal{C}$ such that $err_S(h) = 0$, but $err_{\mathcal{D}}(h) > \varepsilon$.

Event B: $\exists h \in \mathcal{C}$ such that $err_S(h) = 0$, but $err_{S'}(h) > \varepsilon/2$.

Proof: Suppose A occurred.

- Consider $h \in \mathcal{C}$ such that $err_S(h) = 0$, but $err_{\mathcal{D}}(h) > \varepsilon$.
- What's the probability that $err_{S'}(h) > \varepsilon/2$?
- $err_{S'}(h)$ is the average of m i.i.d. Bernoulli random variables

$$\mathbb{E}_{S' \sim \mathcal{D}^m}[err_{S'}(h)] = err_{\mathcal{D}}(h) > \varepsilon$$

Chernoff bound for $\gamma \in (0,1)$ and the average of m i.i.d. Bernoullis with expectation μ:
$\Pr[Y < (1-\gamma)\mu] \leq e^{-\gamma^2 m\mu/2}$

- $\Pr\left[err_{S'}(h) \leq \frac{\varepsilon}{2}\right] \leq e^{-\frac{m\varepsilon}{8}}$
  $\leq e^{-1} \leq 1/2$
- $\Pr[B|A] \geq \Pr\left[err_{S'}(h) \geq \frac{\varepsilon}{2}\right] \geq \frac{1}{2}$

Relating the probabilities of A and B

$$\Pr[B] \geq \Pr[B|A] \cdot \Pr[A] \quad \geq \frac{1}{2} \cdot \Pr[A]$$

Claim

If $m \geq \frac{8}{\varepsilon}$ then $\Pr_{s,s' \sim \mathcal{D}^m}[B|A] \geq \frac{1}{2}$

$$\Pr[A] \leq 2\Pr[B]$$

Want to show: $\Pr[B] \leq \delta/2$

Bounding the Probability of B

Want to show: $\Pr[B] \leq \delta/2$

Claim 2

$$\Pr_{s,s' \sim \mathcal{D}^m}[B] = \Pr_{s,s' \sim \mathcal{D}^m, \sigma \sim \{0,1\}^m}[B_\sigma]$$

Event B: $\exists h \in \mathcal{C}$ such that $err_S(h) = 0$, but $err_{S'}(h) > \varepsilon/2$.

Event B_σ: $\exists h \in \mathcal{C}$ such that $err_T(h) = 0$, but $err_{T'}(h) > \varepsilon/2$.

Proof:

(S, S') and (T, T') have the same distribution.

Bounding the Probability of B_σ

Want to show: $\Pr_{S,S',\sigma}[B_\sigma] \le \delta/2$

We will show: $\Pr_{\sigma}[B_\sigma] \le \delta/2$ for all S, S'

Claim 3

For all $S, S' \in \Omega^m$ and all $h: \Omega \to \{0,1\}$,

$$\Pr_{\sigma\sim\{0,1\}^m}[err_T(h) = 0, \text{ but } err_{T'}(h) > \varepsilon/2] \le 2^{-\varepsilon m/2}$$

Event B_σ: $\exists h \in \mathcal{C}$ such that $err_T(h) = 0$, but $err_{T'}(h) > \varepsilon/2$.

Proof: Let's organize the answers of h on the examples in S and S' in a table.

$h(x_1)$	$h(x_2)$	…	$h(x_m)$
$h(x_1')$	$h(x_2')$	…	$h(x_m')$

- If both answers in some column are wrong, $err_T(h) = 0$ cannot occur.
- If more than $\left(1 - \frac{\varepsilon}{2}\right) m$ columns have both answers right, $err_{T'}(h) > \frac{\varepsilon}{2}$ cannot occur.

$\Pr_{\sigma}[B_\sigma] = 0$

- Assume $r \ge \varepsilon m/2$ columns have one correct and one wrong answer.
- For $err_T(h) = 0$ to occur, σ must send all the right answers from the r columns to T and all the wrong ones to T'
- The probability of this is $2^{-r} \le 2^{-\varepsilon m/2}$

Bounding the Probability of B_σ

Want to show: $\Pr_{S,S',\sigma}[B_\sigma] \le \delta/2$

Claim 3

For all $S, S' \in \Omega^m$ and all $h: \Omega \to \{0,1\}$,

$$\Pr_{\sigma\sim\{0,1\}^m}[err_T(h) = 0, \text{ but } err_{T'}(h) > \varepsilon/2] \le 2^{-\varepsilon m/2}$$

Event B_σ: $\exists h \in \mathcal{C}$ such that $err_T(h) = 0$, but $err_{T'}(h) > \varepsilon/2$.

Claim 4

For all $S, S' \in \Omega^m$,

$$\Pr_{\sigma\sim\{0,1\}^m}[B_\sigma] \le \Pi_{\mathcal{C}}(2m) \cdot 2^{-\varepsilon m/2}$$

Proof: Take a union bound over a set of hypotheses that contains one representative for each projection in $\Pi_{\mathcal{C}}(S \cup S')$, i.e., for each possible labeling of $S \cup S'$ by a hypothesis from $\mathcal{C}$.

Bounding the Probability of B_σ

Want to show: $\Pr_{S,S',\sigma}[B_\sigma] \le \delta/2$

Claim 4

For all $S, S' \in \Omega^m$,

$$\Pr_{\sigma\sim\{0,1\}^m}[B_\sigma] \le \Pi_{\mathcal{C}}(2m)\cdot 2^{-\varepsilon m/2}$$

Want: $\Pi_{\mathcal{C}}(2m)\cdot 2^{-\varepsilon m/2} \le \frac{\delta}{2}$

Equivalently, $\Pi_{\mathcal{C}}(2m)\cdot \frac{2}{\delta} \le 2^{\varepsilon m/2}$

Take the log of both sides: $\log_2\left(\Pi_{\mathcal{C}}(2m)\right) + \log_2\frac{2}{\delta} \le \frac{\varepsilon m}{2}$

Rearrange: $m \ge \frac{2}{\varepsilon}\left(\log_2\left(\Pi_{\mathcal{C}}(2m)\right) + \log_2\frac{2}{\delta}\right)$

Theorem [Vapnik Chervonenkis]

Let $\mathcal{L}$ be an algorithm that outputs an ERM hypothesis. Then $\mathcal{L}$ is a PAC learner for class $\mathcal{C}$ if the number of samples it gets satisfies

$$m \ge \frac{2}{\varepsilon}\left(\log_2\left(\Pi_{\mathcal{C}}(2m)\right) + \log_2\frac{2}{\delta}\right).$$

Rearrange: $m \ge \frac{2}{\varepsilon}\left(\log_2\left(\Pi_{\mathcal{C}}(2m)\right) + \log_2\frac{2}{\delta}\right)$

Bounding the Probability of B_σ

Want to show: $\Pr_{S,S',\sigma}[B_\sigma] \le \delta/2$

Claim 4

For all $S, S' \in \Omega^m$,

$$\Pr_{\sigma\sim\{0,1\}^m}[B_\sigma] \le \Pi_{\mathcal{C}}(2m)\cdot 2^{-\varepsilon m/2}$$

Want: $\Pi_{\mathcal{C}}(2m)\cdot 2^{-\varepsilon m/2} \le \frac{\delta}{2}$

Equivalently, $\Pi_{\mathcal{C}}(2m)\cdot \frac{2}{\delta} \le 2^{\varepsilon m/2}$

Take the log of both sides: $\log_2\left(\Pi_{\mathcal{C}}(2m)\right) + \log_2\frac{2}{\delta} \le \frac{\varepsilon m}{2}$

Rearrange: $m \ge \frac{2}{\varepsilon}\left(\log_2\left(\Pi_{\mathcal{C}}(2m)\right) + \log_2\frac{2}{\delta}\right)$

Theorem [Vapnik Chervonenkis]

Let $\mathcal{L}$ be an algorithm that outputs an ERM hypothesis. Then $\mathcal{L}$ is a PAC learner for class $\mathcal{C}$ if the number of samples it gets satisfies

$$m \ge \frac{2}{\varepsilon}\left(\log_2\left(\Pi_{\mathcal{C}}(2m)\right) + \log_2\frac{2}{\delta}\right).$$

By Sauer's lemma, $\Pi_{\mathcal{C}}(m) \le \binom{m}{\le d} \le \left(\frac{em}{d}\right)^d$

$$m \ge \frac{2}{\varepsilon}\left(d\log_2\left(\frac{2em}{d}\right) + \log_2\frac{2}{\delta}\right)$$

Algebraic manipulations give that $m = \Theta\left(\frac{1}{\varepsilon}\left(d\log\frac{1}{\varepsilon} + \log\frac{1}{\delta}\right)\right)$ **suffices.**

Sample Complexity of PAC Learning

Theorem

Let $\mathcal{C}$ be a class of functions $\Omega \to \{0,1\}$ with finite VC-dimension d.
Then, for some constants C_1 and C_2,
the sample complexity $m(\varepsilon, \delta)$ of PAC-learning $\mathcal{C}$ satisfies

$$\frac{C_1}{\varepsilon}\left(d + \log\frac{1}{\delta}\right) \le m(\varepsilon, \delta) \le \frac{C_2}{\varepsilon}\left(d \log\frac{1}{\varepsilon} + \log\frac{1}{\delta}\right).$$