COMS 4771
Introduction to Machine Learning

Nakul Verma

Towards formalizing ‘learning’

What does it mean to **learn** a concept?

- Gain knowledge or experience of the concept.

The basic process of **learning**

- Observe a phenomenon
- Construct a model from observations
- Use that model to make decisions / predictions

How can we make this more precise?

A statistical machinery for learning

Phenomenon of interest:

Input space: X Output space: Y

There is an unknown distribution $\mathcal{D}$ over $(X \times Y)$

The learner **observes** m examples $(x_1, y_1), \ldots, (x_m, y_m)$ drawn from $\mathcal{D}$

Machine learning

Construct a model:

Let $\mathcal{F}$ be a collection of models, where each $f : X \to Y$ **predicts** y given x

From m observations, **select** a model $f_m \in \mathcal{F}$ which predicts **well**.

$$\mathrm{err}(f) := \mathbb{P}_{(x,y)\sim\mathcal{D}}\Big[f(x) \neq y\Big]$$ (generalization error of *f*)

We can say that we have ***learned*** the phenomenon if

$$\mathrm{err}(f_m) - \mathrm{err}(f^*) \quad \leq \epsilon \qquad\qquad f^* := \mathrm{argmin}_{f\in\mathcal{F}}\, \mathrm{err}(f)$$

for any tolerance level $\epsilon > 0$ of our choice.

PAC Learning

For all tolerance levels $\epsilon > 0$, and all confidence levels $\delta > 0$, if there exists some model selection algorithm $\mathcal{A}$ that selects $f_m^{\mathcal{A}} \in \mathcal{F}$ from *m* observations ie, $\mathcal{A} : (x_i, y_i)_{i=1}^m \mapsto f_m^{\mathcal{A}}$, and has the property:

with probability at least $1 - \delta$ over the draw of the sample,

$$\mathrm{err}(f_m^{\mathcal{A}}) - \mathrm{err}(f^*) \leq \epsilon$$

We call

- The model class $\mathcal{F}$ is PAC-learnable.
- If the m is polynomial in $\frac{1}{\epsilon}$ and $\frac{1}{\delta}$, then $\mathcal{F}$ is **efficiently** PAC-learnable

A popular algorithm:

Empirical risk minimizer (ERM) algorithm

$$f_m^{\mathrm{ERM}} := \mathrm{argmin}_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}\{f(x_i) \neq y_i\}$$

PAC learning simple model classes

Theorem (finite size $\mathcal{F}$):

Pick any tolerance level $\epsilon > 0$, and any confidence level $\delta > 0$

let $(x_1, y_1), \ldots, (x_m, y_m)$ be m examples drawn from an unknown $\mathcal{D}$

if $m \geq C \cdot \frac{1}{\epsilon^2} \ln \frac{|\mathcal{F}|}{\delta}$, then with probability at least $1 - \delta$

$$\mathrm{err}(f_m^{\mathrm{ERM}}) - \mathrm{err}(f^*) \leq \epsilon$$

$\mathcal{F}$ is efficiently PAC learnable

Occam's Razor Principle:

All things being equal, usually the simplest explanation of a phenomenon is a good hypothesis.

Simplicity = representational succinctness

Proof sketch

Define:

$$\mathrm{err}(f) := \mathbb{E}_{(x,y)\sim\mathcal{D}}\Big[\mathbf{1}\{f(x) \neq y\}\Big] \qquad \mathrm{err}_m(f) := \frac{1}{m}\sum_{i=1}^{m}\Big[\mathbf{1}\{f(x_i) \neq y_i\}\Big]$$

(generalization error of f) (sample error of f)

We need to analyze:

$$\begin{aligned}
&\mathrm{err}(f_m^{\mathrm{ERM}}) - \mathrm{err}(f^*) \\
&\quad = \quad \mathrm{err}(f_m^{\mathrm{ERM}}) - \mathrm{err}_m(f_m^{\mathrm{ERM}}) \\
&\qquad + \mathrm{err}_m(f_m^{\mathrm{ERM}}) - \mathrm{err}_m(f^*) \\
&\qquad + \mathrm{err}_m(f^*) - \mathrm{err}(f^*) \\
&\quad \leq \quad 2 \sup_{f\in\mathcal{F}} \Big|\mathrm{err}(f) - \mathrm{err}_m(f)\Big|
\end{aligned}$$

≤ 0

Uniform deviations of expectation of a random variable to the sample

Proof sketch

Fix any $f \in \mathcal{F}$ and a sample (x_i, y_i), define random variable

$$\mathbf{Z}_i^f := \mathbf{1}\{f(x_i) \neq y_i\}$$

$$\mathbb{E}[\mathbf{Z}_1^f] \qquad\qquad \frac{1}{m}\sum_{i=1}^{m}[\mathbf{Z}_i^f]$$

(generalization error of f) (sample error of f)

Lemma (Chernoff-Hoeffding bound '63):

Let $\mathbf{Z_1}, \ldots, \mathbf{Z_m}$ be *m* Bernoulli r.v. drawn independently from **B(*p*)**.

for any tolerance level $\epsilon > 0$

$$\mathbb{P}_{\mathbf{z}_i}\left[\left|\frac{1}{m}\sum_{i=1}^{m}[\mathbf{Z_i}] - \mathbb{E}[\mathbf{Z_1}]\right| > \epsilon\right] \leq 2e^{-2\epsilon^2 m}.$$

A classic result in ***concentration of measure****, proof later*

Proof sketch

Need to analyze

$$\mathbb{P}_{(x_i,y_i)}\left[\text{ exists } f \in \mathcal{F}, \ \left|\frac{1}{m}\sum_{i=1}^{m}[\mathbf{Z}_i^f] - \mathbb{E}[\mathbf{Z}_1^f]\right| > \epsilon\right]$$

$$\leq \sum_{f\in\mathcal{F}} \mathbb{P}_{(x_i,y_i)}\left[\left|\frac{1}{m}\sum_{i=1}^{m}[\mathbf{Z}_i^f] - \mathbb{E}[\mathbf{Z}_1^f]\right| > \epsilon\right]$$

$$\leq \quad 2|\mathcal{F}|e^{-2\epsilon^2 m} \quad \leq \quad \delta$$

Equivalently, by choosing $m \geq \frac{1}{2\epsilon^2}\ln\frac{2|\mathcal{F}|}{\delta}$ with probability at least $1-\delta$, for ***all*** $f \in \mathcal{F}$

$$\left|\frac{1}{m}\sum_{i=1}^{m}[\mathbf{Z}_i^f] - \mathbb{E}[\mathbf{Z}_1^f]\right| \ = \ \left|\mathrm{err}_m(f) - \mathrm{err}(f)\right| \ \leq \ \epsilon$$

■

PAC learning simple model classes

Theorem (Occam's Razor):

Pick any tolerance level $\epsilon > 0$, and any confidence level $\delta > 0$

let $(x_1, y_1), \ldots, (x_m, y_m)$ be m examples drawn from an unknown $\mathcal{D}$

if $m \geq C \cdot \frac{1}{\epsilon^2} \ln \frac{|\mathcal{F}|}{\delta}$, then with probability at least $1 - \delta$

$$\mathrm{err}(f_m^{\mathrm{ERM}}) - \mathrm{err}(f^*) \leq \epsilon$$

$\mathcal{F}$ is efficiently PAC learnable

One thing left…

Still need to prove:

Lemma (Chernoff-Hoeffding bound '63):

Let $\mathbf{Z_1}, \ldots, \mathbf{Z_m}$ be *m* Bernoulli r.v. drawn independently from **B(*p*)**.

for any tolerance level $\epsilon > 0$

$$\mathbb{P}_{\mathbf{z}_i}\left[\left|\frac{1}{m}\sum_{i=1}^{m}[\mathbf{Z_i}] - \mathbb{E}[\mathbf{Z_1}]\right| > \epsilon\right] \leq 2e^{-2\epsilon^2 m}.$$

How sample average deviates from true average

as a function of number of samples (m)

Need to analyze: How does the probability measure concentrates towards a central value (like mean)

Detour: Concentration of Measure

Let’s start with something simple:

Let *X* be a non-negative random variable.

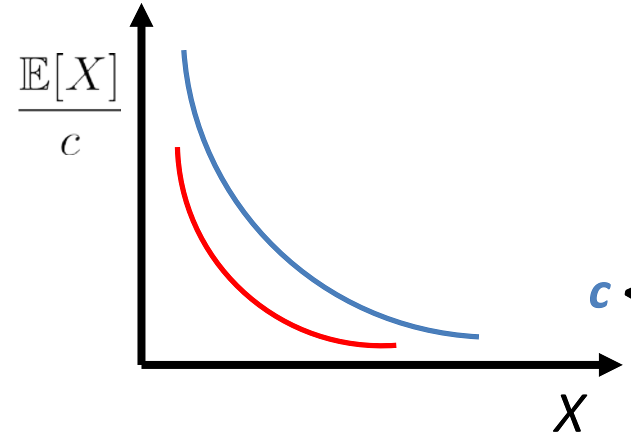
For a given constant c > 0, what is: $\mathbb{P}[X \geq c]$?

$$\mathbb{P}[X \geq c] \leq \frac{\mathbb{E}[X]}{c}$$ *Markov’s Inequality*

Why?

Observation $c \cdot \mathbf{1}[X \geq c] \leq X$

Take expectation on both sides.



Concentration of Measure

Using Markov to bound deviation from mean…

Let X be a random variable (not necessarily non-negative).

Want to examine: $\mathbb{P}[|X - \mathbb{E}X| \geq c]$ for some given constant $c > 0$

Observation:

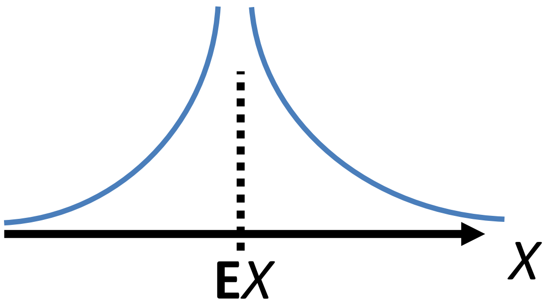
$$\mathbb{P}\big[|X - \mathbb{E}X| \geq c\big] = \mathbb{P}\big[(X - \mathbb{E}X)^2 \geq c^2\big]$$

$$\leq \frac{\mathbb{E}(X - \mathbb{E}X)^2}{c^2}$$ *by Markov's Inequality*

$$= \frac{\mathrm{Var}(X)}{c^2}$$

Chebyshev's Inequality

*True for **all** distributions!*



Concentration of Measure

Sharper estimates using an exponential!

Let X be a random variable (not necessarily non-negative).

For some given constant $c > 0$

Observation:

$$\mathbb{P}[X \geq c] = \mathbb{P}[e^{tX} \geq e^{tc}]$$ *for any t > 0*

$$\leq \frac{\mathbb{E}[e^{tX}]}{e^{tc}}$$ *by Markov's Inequality*

This is called Chernoff's bounding method

Concentration of Measure

Now, Given $X_1, \ldots, X_m$ i.i.d. random variables (assume $0 \le X_i \le 1$)

$$\mathbb{P}\Big[\frac{1}{m}\sum_{i=1}^{m} X_i - \mathbb{E}X_1 \ge c\Big] = \mathbb{P}\Big[\sum_{i=1}^{m} X_i - m\mathbb{E}X_1 \ge mc\Big]$$

*Define $Y_i := X_i$ – **E**X_i*

$$= \mathbb{P}\Big[\sum_{i=1}^{m} Y_i \ge mc\Big]$$

$$\le \frac{\mathbb{E}[e^{t(Y_1+\ldots+Y_m)}]}{e^{tmc}}$$

By Cherneoff's bounding technique

$$= \frac{1}{e^{tmc}} \prod_{i=1}^{m} \mathbb{E}[e^{tY_i}]$$

Y_i i.i.d.

$\boxed{\mathbb{E}[e^{tY_i}] \le e^{t^2/8}}$

$$\le e^{t^2 m/8 - tmc}$$

$\boxed{t = 4c}$

$$\le e^{-2c^2 m}$$

*This **implies** the Chernoff-Hoeffding bound!*

Back to Learning Theory!

Theorem (Occam's Razor):

Pick any tolerance level $\epsilon > 0$, and any confidence level $\delta > 0$

let $(x_1, y_1), \ldots, (x_m, y_m)$ be m examples drawn from an unknown $\mathcal{D}$

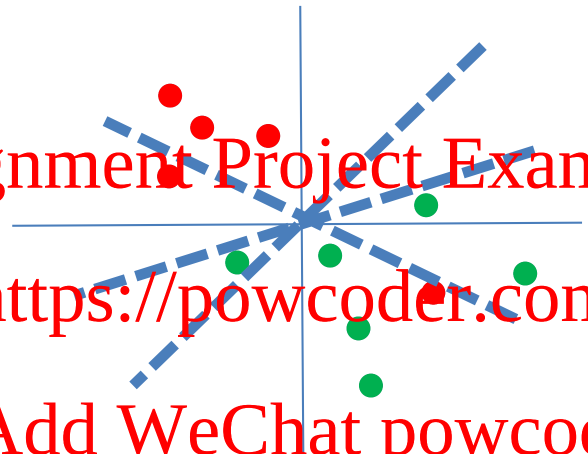
if $m \geq C \cdot \frac{1}{\epsilon^2} \ln \frac{|\mathcal{F}|}{\delta}$, then with probability at least $1 - \delta$

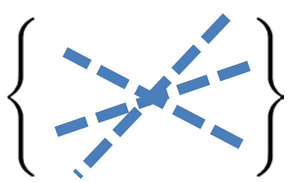
$$\operatorname{err}(f_m^{\mathrm{ERM}}) - \operatorname{err}(f^*) \leq \epsilon$$

$\mathcal{F}$ is efficiently PAC learnable

Learning general concepts

Consider linear classification



$\mathcal{F} =$ 

$|\mathcal{F}| = \infty$

Occam's Razor bound is ineffective

VC Theory

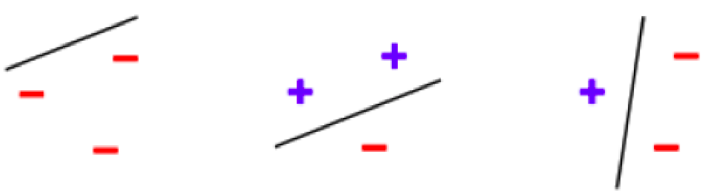
Need to capture the true richness of $\mathcal{F}$

Definition (Vapnik-Chervonenkis or VC dimension):

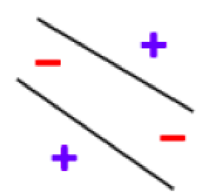
We say that a model class $\mathcal{F}$ as VC dimension *d*, if *d* is the largest set of points $x_1, \ldots, x_d \subset X$ such that for all possible labelings of $x_1, \ldots, x_d$ there exists some $f \in \mathcal{F}$ that achieves that labelling.

Example: $\mathcal{F}$ = linear classifiers in $\mathbf{R}^2$

linear classifiers can realize all possible labellings of 3 points



linear classifiers CANNOT realize all labellings of 4 points



$$\mathrm{VC}(\mathcal{F}) = 3$$

VC Dimension

Another example:

$\mathcal{F}$ = Rectangles in **R**^{2} $\qquad\qquad$ $\mathrm{VC}(\mathcal{F}) = 4$

VC dimension:

- A **combinatorial concept** to capture the true richness of $\mathcal{F}$
- Often (but not always!) proportional to the degrees-of-freedom or the number of independent parameters in $\mathcal{F}$

VC Theorem

Theorem (Vapnik-Chervonenkis '71):

Pick any tolerance level $\epsilon > 0$, and any confidence level $\delta > 0$

let $(x_1, y_1), \ldots, (x_m, y_m)$ be m examples drawn from an unknown $\mathcal{D}$

if $m \geq C \cdot \frac{\mathrm{VC}(\mathcal{F}) \ln(1/\delta)}{\epsilon^2}$, then with probability at least $1 - \delta$

$$\mathrm{err}(f_m^{\mathrm{ERM}}) - \mathrm{err}(f^*) \leq \epsilon$$

$\mathcal{F}$ is efficiently PAC learnable

VC Theorem ➔ Occam's Razor Theorem

Tightness of VC bound

Theorem (VC lower bound):

Let $\mathcal{A}$ be any model selection algorithm that given *m* samples, returns a model from $\mathcal{F}$, that is, $\mathcal{A} : (x_i, y_i)_{i=1}^m \mapsto f_m^{\mathcal{A}}$

For all tolerance levels $0 < \epsilon < 1$, and all confidence levels $0 < \delta < 1/4$,

there exists a distribution $\mathcal{D}$ such that if $m \leq C \cdot \dfrac{\mathrm{VC}(\mathcal{F})}{\epsilon^2}$

$$\mathbb{P}_{(x_i,y_i)}\left[\left|\mathrm{err}(f_m^{\mathcal{A}}) - \mathrm{err}(f^*)\right| > \epsilon\right] > \delta$$

Some implications

- VC dimension of a model class fully characterizes its learning ability!

- Results are agnostic to the underlying distribution.

One algorithm to rule them all?

From our discussion it may seem that ERM algorithm is universally consistent.

This is not the case!

Theorem (no free lunch, Devroye '82):

Pick any sample size *m*, any algorithm $\mathcal{A}$ and any $\epsilon > 0$

There exists a distribution $\mathcal{D}$ such that

$$\mathrm{err}(f_m^{\mathcal{A}}) > 1/2 - \epsilon$$

while the Bayes optimal error, $\min_f \mathrm{err}(f) = 0$

Further refinements and extensions

- How to do model class selection? Structural risk results.
- Dealing with kernels – Fat margin theory

- Incorporating priors over the models – PAC-Bayes theory

- Is it possible to get distribution dependent bound? Rademacher complexity
- How about regression? Can derive similar results for nonparametric regression.

What We Learned...

- Formalizing learning
- PAC learnability
- Occam's razor Theorem
- VC dimension and VC theorem
- VC theorem
- No Free-lunch theorem

Questions?