# Learning preferences

Federico Echenique (Berkeley)

Ridge – Montevideo Nov 2024

# What is a normal martian?

# What is a normal martian?

Each Martian has a weight $w$ and a height $h$, so you imagine them on the plane $(h, w)$.

There is a normal height interval $[h^1, h^2]$, and a normal weight interval $[w^1, w^2]$

So that a Martian $(h, w)$ is normal iff $(h, w) \in [h^1, h^2] \times [w^1, w^2]$.

You have no idea what $h^i$ and $w^i$ are.

You also have no idea what the population distribution $\mu$ is of pairs $(h, w)$.

# What is a normal martian?

You want to learn to predict when a martian is normal.

Given a data on martians, and someone to tell you which ones are normal (a Virgil who accompanies you on your journey).

Learn which ones are normal.

So when presented with a new martian drawn from $\mu$ you can with high prob classify them accurately.

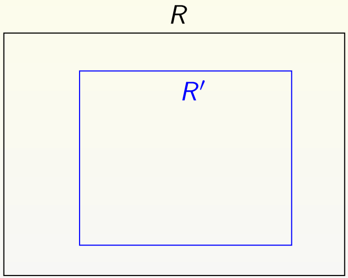
# What is a normal martian?

Now, you are presented with a finite sample of Martians $(h_i, w_i)$, $i = 1, \ldots, n$ and you are told whether each one is normal.

There is a true rectangle $R = [h^1, h^2] \times [w^1, w^2]$.

Given your sample, you construct a minimal rectangle $R'$ that exactly contains the points you have been labeled to be normal.

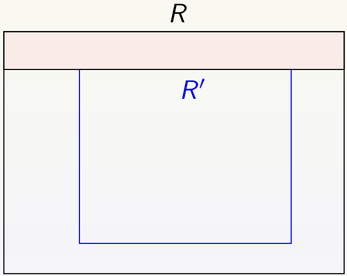
# What is a normal martian?




You want to make sure that the probability according to $\mu$ of the difference $R \setminus R'$ is smaller than $\varepsilon$.

# What is a normal martian?

Consider the difference between $R$ and $R'$ along the northern direction.



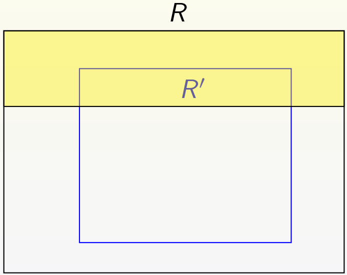
# What is a normal martian?

We want to make sure that this area has probability less than or equal to $\varepsilon/4$.

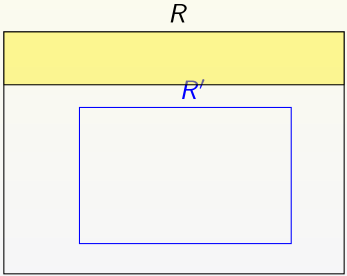
If we can ensure that this is true for the North, East, West and South direction, this means that the difference $R \setminus R'$ has probability less than or equal to $\varepsilon$ (the overcounting of the overlapping area goes in our favor).

Consider the yellow rectangle that we obtain as we sweep $R$ from its Northern boundary going south until we have an area of $\mu$-probability at most $\varepsilon/4$ (assume $\mu$ is non-atomic).

# What is a normal martian?



# What is a normal martian?



# What is a normal martian?

But for this to happen, we would have had to not observe any point in our sample in the yellow area. The probability that all $n$ sample points miss the yellow area is $(1-\varepsilon/4)^n$.

Consider the four slices (East, West, North and South).

The probability that we miss at least one of the yellow slices, each of $\mu$-weight $\varepsilon/4$, is at most (by union bound[1]) $4(1-\varepsilon/4)^n$.

For $n$ large enough we can ensure that this probability is as small as we want.

[1] $P(A \cup B) \leq P(A) + P(B)$.

# What is a normal martian?

How large must $n$ be?

Recall that $(1-\varepsilon) \leq e^{-\varepsilon}$.

Then $4(1-\varepsilon/4)^n \leq 4e^{-n\varepsilon/4}$.

Set $\delta = 4e^{-n\varepsilon/4}$.

Then we need that

$$n \geq \frac{4\ln(4/\delta)}{\varepsilon}.$$

# What is a normal martian?

This is pretty good.

The sample size grows lineary with $1/\varepsilon$ and logarithmically with $1/\delta$.

For example, if $\delta = \varepsilon = 0.05$, then we have $n \geq 80 \ln 80 \simeq 351$.

# PAC learning

# PAC learning

Given is:

- A measure space $(X, \Sigma)$, termed the instance space.
- A probability distribution $\mu$ on $(X, \Sigma)$.
- A subset $c^* \subseteq X$ is the target concept.

For ex:

- $X$ is a set of strings of text.
- $c^*$ the set of text with a particular political message.

For ex:

- $X = \mathrm{R}^d$ is the space of torax $x$-ray images (encoded as $d$-dimensional vectors).
- $c^*$ the set of images with a tumor

# PAC learning

Want to learn $c^*$ from an iid sample $S = \{x_1, \ldots, x_n\}$, taken according to $\mu$ on $X$.

Where we are told whether each $x_i \in c^*$.

In other words, each $x_i$ is labeled.

A class $\mathcal{H}$ of subsets of $X$ is called the hypothesis class.

We may or may not have $c^* \in \mathcal{H}$.

# PAC learning

Given $h \in \mathcal{H}$, the true error of the hypothesis $h$ is

$$\mathcal{E}_\mu(h) = \mu(c^* \triangle h).$$

Given a sample $S$ drawn according to $\mu$, the training error is

$$\mathcal{E}_S(h) = \frac{|S \cap (c^* \triangle H)|}{|S|}.$$

# PAC learning

Let $\varepsilon > 0$ and denote by $\mathcal{H}_\varepsilon \subseteq \mathcal{H}$ the set of all hypotheses that have true error greater than $\varepsilon$.

If $h \in \mathcal{H}_\varepsilon$, what is the probability that $h$ will have training error $= 0$ given a sample $S$?

In other words, what is the probability that $\mathcal{E}_S(h) = 0$ when $\mathcal{E}_\mu(h) \geq \varepsilon$?

This is at most

$$(1 - \varepsilon)^{|S|}.$$

# PAC learning

If $\mathcal{H}_\varepsilon$ is finite, then the probability that *at least one* $h \in \mathcal{H}_\varepsilon$ has $\mathcal{E}_S(h) = 0$ is (by union bound) at most $|\mathcal{H}_\varepsilon| (1 - \varepsilon)^{|S|}$.

We want this number to be small.

So if $\delta =$ the prob. that at least one hypothesis with true error $\geq \varepsilon$ has training error $= 0$, and we assume that $\mathcal{H}$ is finite, then:

$$\delta \leq |\mathcal{H}| \, e^{-\varepsilon |S|}$$

(using that $1 - \varepsilon \leq e^{-\varepsilon}$).

Set $n = |S|$ to be the sample size.

# PAC learning

So $\ln(\delta) \leq \ln(|\mathcal{H}|) - \varepsilon n$, or

$$\frac{\ln(1/\delta) + \ln(|\mathcal{H}|)}{\varepsilon} \geq n.$$

# PAC learning

**Theorem**

Let $\mathcal{H}$ be a finite hypothesis class. Given $\varepsilon > 0$ and $\delta \in (0, 1)$, if

$$n \geq \frac{\ln(1/\delta) + \ln(|\mathcal{H}|)}{\varepsilon}$$

then with probability at least $1 - \delta$ all hypotheses with training error $= 0$ have true error $< \varepsilon$.

# PAC learning

But what if there is no hypothesis with zero training error?

Suppose instead that we would like $\mathcal{E}_S(h)$ and $\mathcal{E}_\mu(h)$ to be close for all $h$.

This is a kind of uniform convergence results, and follows along similar lines:

**Theorem**

Let $\mathcal{H}$ be a finite hypothesis class. Given $\varepsilon > 0$ and $\delta \in (0,1)$, if

$$n \geq \frac{\ln(2/\delta) + \ln(|\mathcal{H}|)}{2\varepsilon^2}$$

then, with probability at least $1-\delta$, $|\mathcal{E}_\mu(h) - \mathcal{E}_S(h)| < \varepsilon$ for all $h \in \mathcal{H}$.

# Application: Occam's razor

We can use these ideas to formalize Occam's razor: the notion that the simplest explanations are more likely to be correct.

Suppose that $\mathcal{H}$ is described using some language that takes at most $b$ bits. The idea being that the smaller is $b$ the simpler the explanation.

Then we have that $|\mathcal{H}| \leq 2^b$.

As long as we set $n \geq \frac{1}{\varepsilon}[b \ln(2) + \ln(1/\delta)]$, then with probability $\geq 1 - \delta$, any hypothesis that can be described with $b$ bits and has a training error of zero must have true error $< \varepsilon$.

# PAC learning

What is $\mathcal{H}$ is not finite?

The previous ideas generalize.

The theory is more involved (but interesting!).

VC dimension plays the role of $|\mathcal{H}|$.

We shall see this in the context of our application.

# Learning preferences

# Learning preferences

PAC learning is about classification.

Now to economics.

What is the connection?

Well, a preference is a hypotesis.

$\succeq$ is the set of $(x, y)$ s.t. $x$ is chosen over $y$.

# Learning preferences

Let $X$ be a set of objects of choice.

For example, a set of consumption vectors ($X = \mathrm{R}^d_+$).

$\mathcal{P}$ a class of preferences on $X$.

Then each $\succeq \in \mathcal{P}$ is a subset of $X \times X$.

# Statistical model

1. In a choice problem, alternatives drawn iid according to sampling distribution $\lambda$.
2. Subjects make "mistakes."
   Upon deciding on $\{x, y\}$, a subject with preference $\succeq$ chooses $x$ over $y$ with probability $q(\succeq; x, y)$ (error probability function).
3. Only assumption: if $x \succ y$ then $q(\succeq; x, y) > 1/2$.
4. "Spatial" dependence of $q$ on $x$ and $y$ is arbitrary.

# Estimator

Kemeny-minimizing estimator: find a preference in $\mathcal{P}$ that minimizes the number of observations inconsistent with the preference.

- "Model free:" to compute estimator don't need to assume a specific $q$ or $\lambda$.
- May be computationally challenging (depending on $\mathcal{P}$).

# Assumptions

Assumption 1: $X$ is a locally compact Polish space.

Assumption 2: $\mathcal{P}$ is a closed set of locally strict preferences.

Assumption 3: $\lambda$ has full support and for all $\succeq \in \mathcal{P}$, $\{(x,y) : x \sim y\}$ has $\lambda$-probability 0.

# Second main result

**Theorem**

Under Assumptions (1), (2), (3'), if the subject's preference is $\succeq^* \in \mathcal{P}$ and $\succeq_n$ is the Kemeny-minimizing estimator for $\Sigma_n$, then, $\succeq_n \to \succeq^*$ in probability.

# Convergence rates: Digression

The VC dimension of $\mathcal{P}$ is the largest cardinality of an experiment that can always be rationalized by $\mathcal{P}$.

A measure of how flexible $\mathcal{P}$; how prone it is to overfitting.

# Convergence rates: Digression

- Think of a game between Alicia and Roberto
- Alicia defends $\mathcal{P}$; Roberto questions it.
- Given is $k$
- Alicia proposes a choice experiment of size $k$
- Roberto fills in choices adversarily.
- Alicia wins if she can rationalize the choices using $\mathcal{P}$.
- The VC dimension of $\mathcal{P}$ is the largest $k$ for which Alicia always wins.

# Convergence rates

- Let $\rho$ be a metric on preferences.

**Theorem**

Under the same assumptions as in prev. thm,

$$N(\eta, \delta) \leq \frac{2}{r(\eta)^2} \left( \sqrt{2/\delta} + C \sqrt{\mathrm{VC}(\mathcal{P})} \right)^2$$

with $C$ a universal constant.

# Convergence rates

- Let $\rho$ be a metric on preferences.
- $N(\eta, \delta)$ : smallest value of $N$ such that for all $k \geq N$, and all subject preferences $\succeq^* \in \mathcal{P}$,

$$\Pr(\rho(\succeq_k, \succeq^*) < \eta) \geq 1 - \delta.$$

**Theorem**

Under the same assumptions as in prev. thm,

$$N(\eta, \delta) \leq \frac{2}{r(\eta)^2} \left( \sqrt{2/\delta} + C\sqrt{\mathrm{VC}(\mathcal{P})} \right)^2$$

with $C$ a universal constant.

# Convergence rates

- Let $\rho$ be a metric on preferences.
- $N(\eta, \delta)$ : smallest value of $N$ such that for all $k \geq N$, and all subject preferences $\succeq^* \in \mathcal{P}$,

$$\Pr(\rho(\succeq_k, \succeq^*) < \eta) \geq 1 - \delta.$$

- $\mu(\succeq'; \succeq)$ : prob. that choice w/preference $\succeq$ is consistent w/$\succeq'$.

$$r(\eta) = \inf \left\{ \mu(\succeq; \succeq) - \mu(\succeq'; \succeq) : \succeq, \succeq' \in \mathcal{P}, \rho(\succeq, \succeq') \geq \eta \right\}.$$

**Theorem**

Under the same assumptions as in prev. thm,

$$N(\eta, \delta) \leq \frac{2}{r(\eta)^2} \left( \sqrt{2/\delta} + C\sqrt{\mathrm{VC}(\mathcal{P})} \right)^2$$

with $C$ a universal constant.

# Convergence rates

- Let $\rho$ be a metric on preferences.
- $N(\eta, \delta)$ : smallest value of $N$ such that for all $k \geq N$, and all subject preferences $\succeq^* \in \mathcal{P}$,

$$\Pr(\rho(\succeq_k, \succeq^*) < \eta) \geq 1 - \delta.$$

- $\mu(\succeq'; \succeq)$ : prob. that choice w/preference $\succeq$ is consistent w/$\succeq'$.

$$r(\eta) = \inf \left\{ \mu(\succeq; \succeq) - \mu(\succeq'; \succeq) : \succeq, \succeq' \in \mathcal{P}, \rho(\succeq, \succeq') \geq \eta \right\}.$$

- $\mathrm{VC}(\mathcal{P})$ the VC dimension of the class $\mathcal{P}$.

**Theorem**

Under the same assumptions as in prev. thm,

$$N(\eta, \delta) \leq \frac{2}{r(\eta)^2} \left( \sqrt{2/\delta} + C\sqrt{\mathrm{VC}(\mathcal{P})} \right)^2$$

with $C$ a universal constant.

# Expected utility

1. $X$ is the set of lotteries over $d$ prizes.
2. $\mathcal{P}$ is the set of **nonconstant** EU preferences: there are always lotteries $p, p'$ such as $p$ is strictly preferred to $p'$.

This preference environment satisfies Assumptions 1 and 2.

Suppose: there is $L > 0$ and $m > 0$ s.t

$$q(x, y; \succeq) \geq \frac{1}{2} + L(v \cdot x - v \cdot y)^m,$$

when $x \succeq y$ and $v$ represents $\succeq$.

# Expected utility

Under these assumptions, we can bound $r(\eta)$ and $\mathrm{VC}(\mathcal{P})$, which implies

$$N(\eta, \delta) = O\left(\frac{1}{\delta \eta^{4d-2}}\right).$$

Other examples: Cobb-Douglas, CES, and CARA subjective EU preferences, and intertemporal choice with discounted, Lipschitz-bounded utilities.

# Monotone preferences

- $K$ be a compact set in $X \equiv \mathrm{R}^d_{++}$, and fix $\theta > 0$.
- $\mathcal{P}$ has finite VC-dimension and is identified on $K$
- $\lambda$ is the uniform probability measure on $K^{\theta/2}$,
- $q$ satisfies: probability of choosing $y$ instead of $x$ when $x \succ y$ is a function of $\|x - y\|$,

**Theorem**

The Kemeny-minimizing estimator is consistent and, as $\eta \to 0$ and $\delta \to 0$,

$$N(\eta, \delta) = O\left(\frac{1}{\eta^{2d+2}} \ln \frac{1}{\delta}\right).$$

# References


- Kearns and Vazirani "An introduction to computational learning theory" MIT press (1994).
- Blum, Hopcroft and Kannan "Foundations of data science" Cambridge University Press (2020).
- Chambers, Echenique and Lambert "Recovering preferences from finite data" *Econometrica* v. 89 No. 4 (2021).