Statistical and computational foundations of machine learning

- What is machine learning?

- Study of computer algorithms that improve through experience and use of data.

- As scientific discipline machine learning is part of
  1) Statistics
  2) Optimization (Operations research)
  3) Computer science

It is part of AI.

These days it is almost synonymous with AI.

Applications

- Security applications
  - Spam filtering (emails, discussion forums)
  - Credit card fraud detection
  - Network intrusion detection
- Content optimization
  - Web search
  - Recommendation systems (Netflix, Amazon, ...)

- Online advertising

- Computer vision
  - OCR
  - Face detection in cameras
  - Face recognition
  - Object recognition
- Natural language processing
  - Speech-to-text
  - Translation
  - Chatbots

- Reinforcement learning
  - Robotics
  - Trading

Various types of ML

- Supervised vs. Unsupervised
  - Are there labels that need be predicted?
  - No obvious labels (clustering, topic modelling)
- Active vs. Passive (How are labels collected?)
  - Passive: For random data examples?
  - Active: For carefully and adaptively chosen examples
- Online vs. Batch
  - Learning and prediction

- is interleaved
- Learn, then predict

- Adversarial vs. IID vs. Helpful
  - How are examples constructed

We will focus on supervised learning, passive learning.

Mostly in batch setting with some excursions to online. Mostly i.i.d. with some excursions to adversarial.

We will almost exclusively look at binary classification.

Papayas

- Tropical fruit.
- Can observe color and softness
- Predict tasty / non-tasty

Statistical model

- Domain X
  (Domain = non-empty set)
  (Feature space)
  - Papaya's $X = \mathbb{R}^2$

- Emails $X = \{a, b, \ldots, z\}^*$

  ⋮

- Unlabeled example

  $x \in X$

- Set of possible labels

  $Y = \{0, 1\}$ or $Y = \{+1, -1\}$

  (We are dealing with binary classification only.)

- Labeled example

  $(x, y) \in X \times Y$

- Unlabeled sample

$$S = (x_1, x_2, \ldots, x_m) \in X^*$$

- Labeled sample

$$S = ((x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)) \in (X \times Y)^*$$

- Classifier

$$h : X \to Y$$

- Predictor $h \in Y^X$
- Hypothesis
- Concept
- Model
- Set $h \leftrightarrow h^{-1}(1)$

- Learning algorithm

  = mapping from labeled samples to classifiers

$$A : \underbrace{(X \times Y)^*}_{\text{set of all possible}} \rightarrow \underbrace{Y^X}_{\text{set of all possible}}$$

labeled samples → Learner / Training algorithm → classifiers

- Mistake / error of a classifier h on labeled example (x, y)

  iff $h(x) \neq y$

- Probality distribution D over $X \times Y$

- D implicitly induces a distribution over labeled samples of size m:

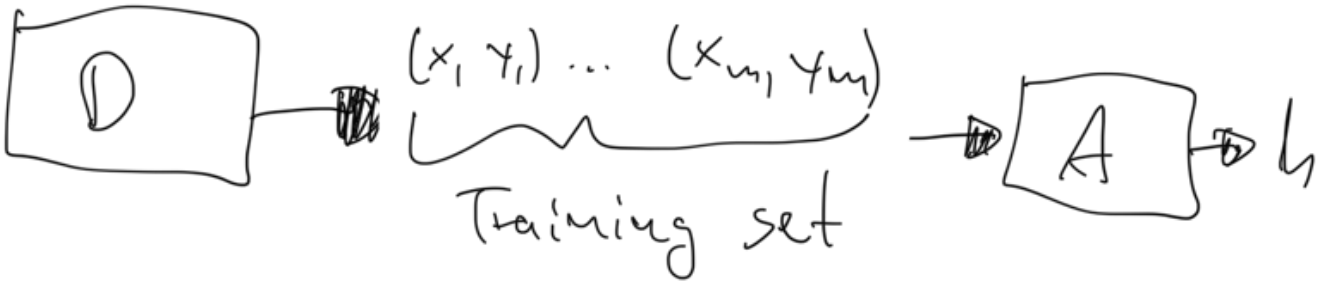
  D^m

$(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)$

is an i.i.d. sequence sampled from D.

- $(x_i, y_i) \sim D$
- $(x_1, y_1) \ldots (x_m, y_m)$ are independent
  - Beware x_i and y_i might be dependent!
- IID = independently identically distributed

System diagram



D models both the distribution of unlabeled examples and the "correct" labels.

Papaya example

- Collect bunch of papayas at random
- Document their features
- Taste them all
- Produce labeled sample

Email spam example

Learning spam filter

- Collect bunch of emails at random
- Manually determine for each if its spam or not
- The resulting collection is a labeled sample

Success criterion

- Generalization error of a predictor $h : X \to Y$

$$err_D(h) = Pr[h(x) \neq y]$$

where $(x, y) \sim D$.

$$err_S(h) = \frac{\#[h(x) \neq y]}{m}$$

$L_D(h) = \mathbb{E}[\ell(h(x), y)]$

More generally we can look at

$$L_D(h) = \mathbb{E}[\underbrace{\ell(h(x), y)}_{\text{loss of } h \text{ on } (x,y)}]$$

$L_D(h)$ / $\text{err}_D(h)$ is generalization error, expected loss risk.

Note for fixed h generalization error is a fixed number. It is <u>not</u> random.

The problem however is

that D is often unknown and thus $err_D(h)$ is unknown as well.

Optimal Bayes classifier

Suppose you know D.

The best classifier you can construct is called "Bayes optimal classifier".

It is defined as

$$h^*(x) = \begin{cases} 1 & \text{if } Pr[y=1 \mid x] \geq 1/2 \\ 0 & \text{if } Pr[y=1 \mid x] < 1/2 \end{cases}$$

- When $\Pr[y=1|x] = 1/2$ we can predict 0 or 1. It does not matter.

- If labels are $\{0,1\}$ we can also express it as

$$h^*(x) = \begin{cases} 1 & \text{if } \mathbb{E}[y|x] \geq 1/2 \\ 0 & \text{if } \mathbb{E}[y|x] < 1/2 \end{cases}$$

- If labels are $\{+1,-1\}$ we can express if as

$$h^*(x) = \begin{cases} +1 & \text{if } \mathbb{E}[y|x] \geq 0 \end{cases}$$

$$= -1 \quad \text{if } \mathbb{E}[Y|X] < 0$$

$$= \text{sign}\left(\mathbb{E}[Y|X]\right)$$

where

$$\text{sign}(z) = \begin{cases} +1 & \text{if } z \geq 0 \\ -1 & \text{if } z < 0 \end{cases}$$

(This version of sign function is very common in ML. In some other fields

$$\begin{cases} +1 & \text{if } z > 0 \end{cases}$$

$$\text{sign}(z) = \begin{cases} 0 & \text{if } z = 0 \\ -1 & \text{if } z < 0 \end{cases}$$

- The generalization error of Bayes optimal classifier is the lowest among all possible functions:

Consider any $h : X \to Y$

Then

$$\text{err}_D(h) = \Pr[h(x) \neq y]$$

$$= \mathbb{E}[\mathbb{1}[h(x) \neq y]]$$

$$= \mathbb{E}\Big[\underbrace{\mathbb{E}\big[\mathbb{1}[h(x) \neq y] \mid x\big]}\Big]$$

$$\geq \mathbb{E}\big[\mathbb{1}[h^*(x) \neq y] \mid y\big]$$

$$\Pr[h(x) \neq y \mid x] \geq \Pr[h^*(x) \neq y \mid x]$$

If $h(x) = h^*(x)$ then both sides are equal. Suppose $h(x) \neq h^*(x)$

$$\Pr[h(x) \neq y \mid x] = \begin{cases} \Pr[y=0 \mid x] & \text{if } h(x)=1 \\ \Pr[y=1 \mid x] & \text{if } h(x)=0 \end{cases}$$

$$\Pr[h^*(x) \neq y \mid x] = \begin{cases} \Pr[y=0 \mid x] & \text{if } h^*(x)=1 \\ \Pr[y=1 \mid x] & \text{if } h^*(x)=0 \end{cases}$$

$\geq \Pr[y=1|x]$ if $h^*(x)=1$

Case 1: $h(x)=0$, $h^*(x)=1$

$h^*(x)=1$, $\Pr[h^*(x) \neq y | x] = \Pr[y=0|x]$

$h(x)=0$, $\Pr[h(x) \neq y | x] = \Pr[y=1|x]$

$h^*(x)=1 \Rightarrow \Pr[y=1|x] \geq 1/2$

$\Pr[y=0|x] \leq 1/2$

Case 2: $h(x)=1$, $h^*(x)=0$

Symmetric.

- Note that

$\text{err}(h^*) \leq 1/2$

$err_D(M) = 1 -$

- Note that $err_D(h^*)$ can be positive, and even if can be pretty high.

  This is the fundamental limit on generalization of <u>any</u> classifier.

ERM

- ERM = empirical risk minimization
- Also called SEM = sample error minimization
- Pick a class predictors $H \subset Y^X$

$H \subseteq$ /

- Given labeled sample

$$S = \{(x_1, y_1), (x_2, y_2) \cdots (x_m, y_m)\}$$

define empirical error or sample error or training error

$$\widehat{err}_S(h) = \frac{|\{i : 1 \le i \le m, h(x_i) \neq y_i\}|}{m}$$

fraction of mistakes h makes on S.

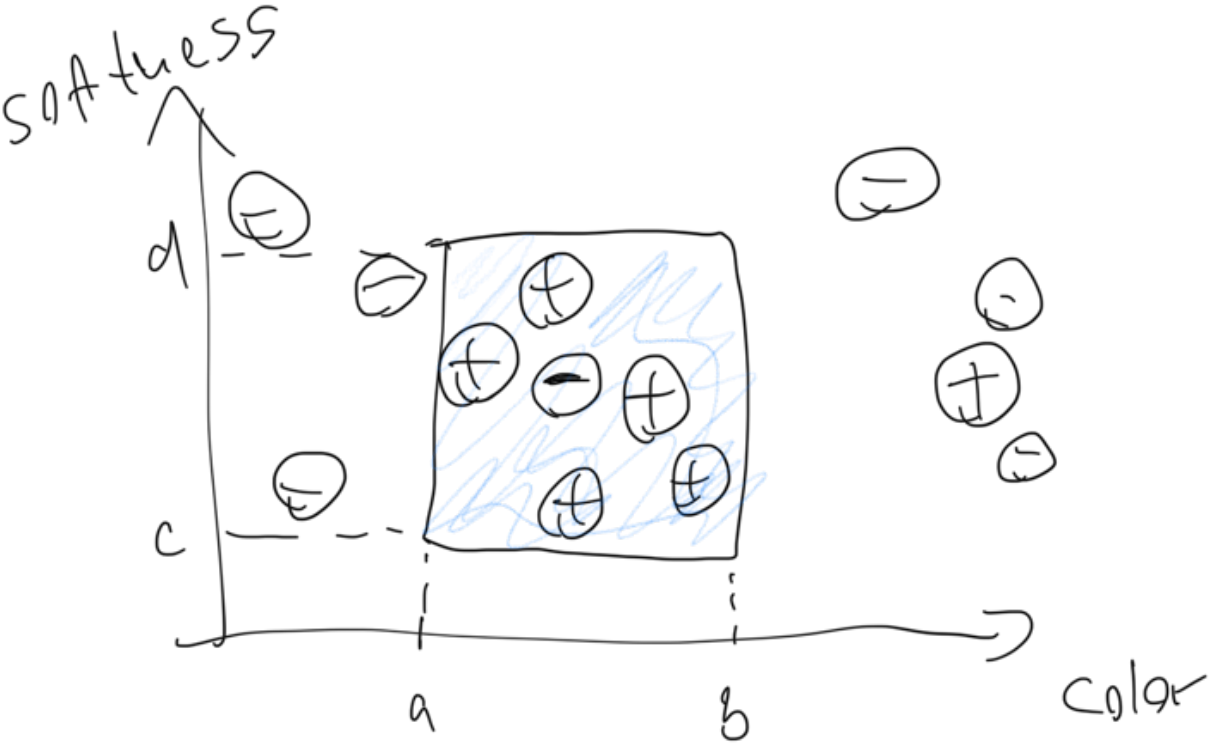
- ERM algorithm:

• Find $h \in H$ with

* Find $h \in H$ with the smallest possible empirical error.

$$ERM(S) = \underset{h \in H}{\operatorname{argmin}} \widehat{err}_S(h)$$

Papaya example + Rectangles



$$h_{a,b,c,d}(x) = \begin{cases} 1 & \text{if } a \le x_1 \le b,\ c \le x_2 \le d \\ 0 & \text{otherwise} \end{cases}$$

ERM can go horribly

Imagine that U is uniform distribution over $[0,1)$.

Define D over $[0,1) \times \{0,1\}$ as follows $(x,y) \sim D$

$x \sim U$

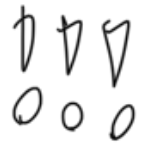
$y = 1$

$ERM(S) = h$ where

$$h(x) = \begin{cases} 1 & \text{if } (x,1) \in S \\ 0 & \text{otherwise} \end{cases}$$

$\widehat{err}_S(h) = 0$

$err_D(h) = 1$



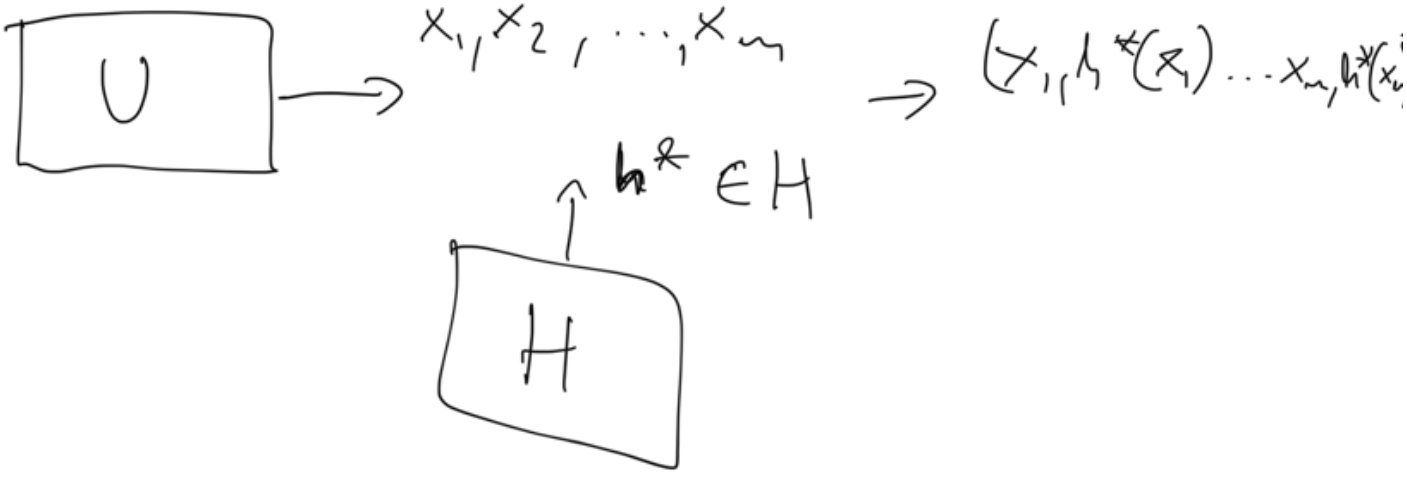
- When $\widehat{err}_S(h)$ is small and $err_D(h)$ is big we call it <u>overfitting</u>.

- <u>Overfitting</u> is the opposite of <u>generalization</u>

<u>Sometimes ERM goes right</u>

- Assume H is finite

- Assume Bayes optimal classifier belongs to H and $err_D(h) = 0$.



Theorem: Suppose $H \subseteq Y^X$ is finite. Let D be a distribution over $X \times Y$ such that for some $h^* \in H$, $err_D(h^*) = 0$. Let S be a labeled sample from D of size $m \geq \frac{\ln|H| + \ln 1/\delta}{\varepsilon}$

then with probality at least $1-\delta$, $err_D(ERM(S)) < \varepsilon$.

Proof:

$$err_{D,h^*}(h) = Pr\left[h^*(x) \neq h(x)\right]$$

Let $H_{BAD} = \{h \in H : err_{D,h^*}(h) \geq \varepsilon\}$

---

1) $err_D(ERM(S)) \geq \varepsilon$

$\Leftrightarrow$

2) $ERM(S) \in H_{BAD}$

$\Rightarrow$

3) $\exists h \in H_{BAD}$ s.t. $\widehat{err}_S(h) = 0$

---

1)' $\forall h \in H_{BAD}$: $err_S(h) > 0$ | Complementary events

2)' $ERM(S) \in H_{BAD}$

$\Leftrightarrow$

3)' $err_D(ERM(S)) < \varepsilon$

- Consider $h \in H_{BAD}$
- Consider $\widehat{err}_S(h)$.
- What is the probility that $\widehat{err}_S(h) = 0$?

$$Pr\left[\widehat{err}_S(h) = 0\right] \leq (1-\varepsilon)^m$$

It suffice to show that

$$Pr\left[\exists h \in H_{BAD} \text{ s.t. } err_D(h) \geq \varepsilon\right] \leq \delta$$

$$\Pr\left[\exists h \in H_{BAD} \text{ s.t. } \widehat{err}_S(h) \geq \varepsilon\right]$$

$$= \Pr\left[\bigcup_{h \in H_{BAD}} \{S : \widehat{err}_S(h) \geq \varepsilon\}\right]$$

$$\leq \sum_{h \in H_{BAD}} \Pr\left[\{S : \widehat{err}_S(h) \geq \varepsilon\}\right] \quad \text{(Union Bound)}$$

$$= \sum_{h \in H_{BAD}} \Pr\left[\widehat{err}_S(h) \geq \varepsilon\right]$$

$$\leq |H_{BAD}| \cdot (1-\varepsilon)^m$$

$$\leq |H| \cdot (1-\varepsilon)^m$$

$$< |H| \cdot e^{-\varepsilon m} \qquad (1+x < e^{\pm x})$$

$$\leq \delta$$