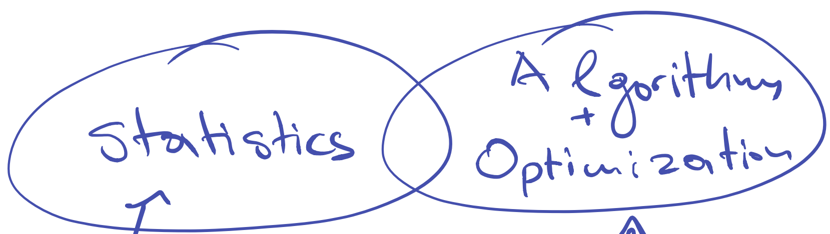
Lecture 2:

• quiz • scribe

Concetration of the emprical risk

ML Research



Explains the "why"s | Explains the "how"s

Today: Why/When does ERM work?

Reminder:

$(\vec{x}_i, \vec{y}_i) \sim D$

hypothesis class $\mathcal{H}$
(aka predictor)

$\begin{bmatrix} \text{linear} \\ \text{SVM} \\ \text{NN} \\ \text{dec. tree} \\ \vdots \end{bmatrix}$

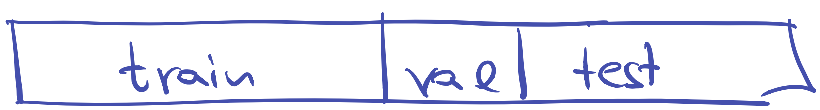
Goal:

"We want to find the best $h \in \mathcal{H}$ for a given D and loss function"

Empirical Risk Minimization (ERM):

$$\underbrace{\min_{h \in \mathcal{H}}}_{\text{model}} \frac{1}{n} \sum_{i=1}^{n} \underbrace{\ell(h(\vec{x}_i); y_i)}_{\text{performance of model on data point } i}$$

Usually data set is split in 3 parts:



find models — eval. and pick the best — report and "forecast"

please "google": • cross validation • hold out set • read intro. to stat. learn.

Main Questions for today:

- When is the <u>empirical risk</u> a good estimator for the <u>true risk</u>?
  [i.e, Does the ERM concentrate?]

- How Does the choice of the model affect the "worst case" concentration of the ERM?

Some Definitions:

There is an unknown distribution $\mathcal{D}$ over labeled examples $\mathcal{X} \times \mathcal{Y}$

feature space   label space

We receive a sample data set of n i.i.d examples

$$S = \{z_1, z_2, \dots, z_n\}, \quad z_i = (x_i, y_i) \sim \mathcal{D}$$

Our goal is to find a hypothesis h_S with small expected/true risk

$$R[h_S] = \mathbb{E}_{z \sim D}\left\{ \ell(h_S(\vec{x}); y) \right\}$$

ℓ: loss of hypothesis h_S on example $\vec{x}$ and its true label y.

The loss measures the disagreement between predictions and reality.

Since we can't directly measure $R[\cdot]$, which is our true objective, we can possibly consider optimizing its sample-average proxy, i.e., the empirical risk:

$$\hat{R}_S[h_S] = \frac{1}{n} \sum_{i=1}^{n} \ell(h_S(\vec{x}_i); y_i)$$

Our hope is that $\hat{R}_S$ is close to R.

The generalization gap:

$$\mathcal{E}_{gen}(h_S) = \left| R[h_S] - \hat{R}_S[h_S] \right|$$

- **Question:** When is it possible to bound $\mathcal{E}_{gen}$ by a small constant?

The answer must depend on:

1) n, the sample size
2) $\mathcal{H}$, the hypothesis class
3) $\mathcal{D}$
[4) The optimization algorithm]

- Assumption: Let the loss be bounded

$$0 \le \ell(w; x) \le 1 \quad \forall w, x$$

can be replaced with a constant $c \in \mathbb{R}^+$

We will use Hoeffding's Inequality to prove that the empirical risk $\hat{R}_S$ concentrates:

<u>Theorem</u>: Let $X_1, \dots, X_n$ be independent RVs on $\mathbb{R}$, such that $0 \le X_i \le 1$ and

$$S = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Then, for all $\varepsilon \geq 0$

$$\Pr\left(|S - \mathbb{E}[S]| \geq \varepsilon \right) \le 2 \cdot e^{-2n\varepsilon^2}$$

- The above is true no matter what the distribution of X_i is!

- <u>Use case</u>: How many samples n do we need to guarantee $S = \mathbb{E}[S] \pm \varepsilon$ with $\Pr\{\cdot\} = \delta$?

$$\delta = 2 e^{-2n\varepsilon^2} \Rightarrow \log\left(\frac{\delta}{2}\right) = -2n\varepsilon^2$$

$$\Rightarrow n = -\log\left(\frac{\delta}{2}\right) / \varepsilon^2 \Rightarrow n = O\left(\frac{\log(1/\delta)}{\varepsilon^2}\right)$$

Careful! Powerful statements like the above tend to be very restrictive!
H.I. is "oblivious" to the distr. of x_i.

Let's try to apply Hoeffding to the empirical risk.

Assume that $h(\cdot;\cdot)$ (i.e., our predictor) is fixed, i.e., it does not depend on the data (!)

Let $R_i[h] = \ell(h(x_i); y_i)$ and

$$\hat{R}_S[h] = \frac{1}{n}\sum_{i=1}^{n} R_i[h]$$

[observe that $R_i[\cdot]$'s are independent]

Then, by the H.I. we have

$$\Pr\left(\left|\hat{R}_S[h] - \mathbb{E}[\hat{R}_S[h]]\right| \geq \varepsilon\right) \leq 2 \cdot e^{-2n\varepsilon^2}$$

What is $\mathbb{E}[\hat{R}_S[h]] = ?$ It is equal to

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[R_i[h]] = \frac{1}{n}\sum_{i=1}^{n} \underbrace{R[h]}$$

the true risk!

Hence, for any given (or fixed) h the empirical risk "converges" to the true with rate $\sim \frac{1}{\sqrt{n}}$

- Question: Is that enough?

No! This result only applies to one h.

What we need: Results for at least a subclass $\mathcal{H}$ of predictors

Simple Example: Say I want $\mathcal{H}$ to be all binary linear classifiers $w \in \{0,1\}^d$. Then $|\mathcal{H}| = 2^d$.
How do we handle this?

- Union Bound: $\Pr(\bigcup_i A_i) \leq \sum_i \Pr(A_i)$

- Use U.B. on the set $\mathcal{H}$, e.g.,

$$\Pr\left(\max_{h\in\mathcal{H}} |\hat{R}_S[h] - R[h]| \geq \varepsilon\right)$$

Observe that

$$\Pr\left(\max_{h\in\mathcal{H}} |\hat{R}_S[h] - R[h]| \geq \varepsilon\right)$$

$$\leq \Pr\left(\sum_{h\in\mathcal{H}} |\cdot - \cdot|\right) \leq 2|\mathcal{H}|e^{-2n\varepsilon^2}$$

$$\leq 2e^{\log|\mathcal{H}|}e^{-2n\varepsilon^2} = 2e^{-2n\varepsilon^2 + \log|\mathcal{H}|}$$

- Hence we need $n = O\left(\frac{\log|\mathcal{H}|/\delta)}{\varepsilon^2}\right)$ samples for ε gen gap with prob. $1-\delta$.

Even this simple bound can give some meaningful results.

<u>Examples:</u>

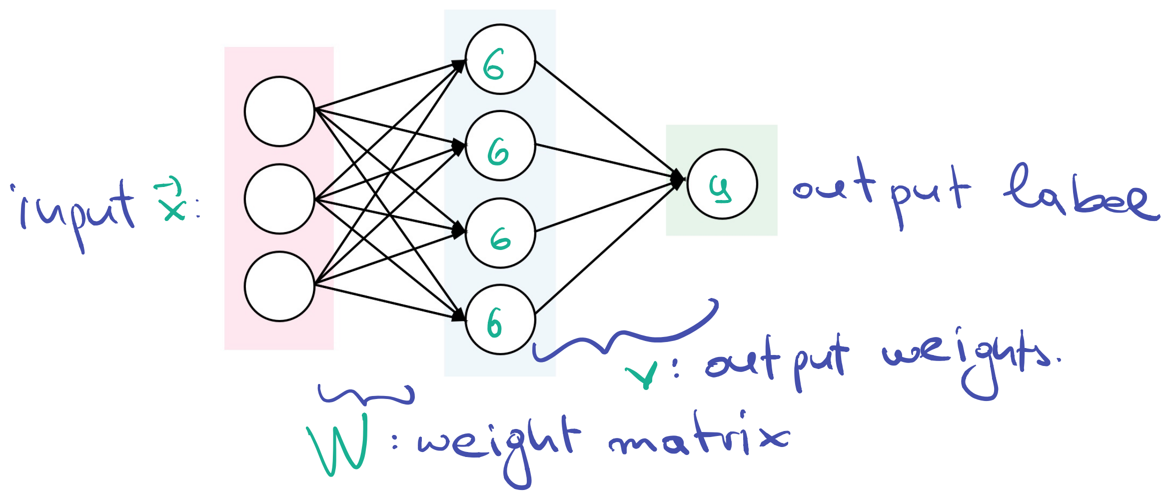
- Binary classification and floating point arithmetic

$$h(w;x) = \text{sign}(w^T x + b)$$

Q: # predictors in class? $|\mathcal{H}| = 2^{16\cdot d}$

So in this case $n = O\left(\frac{d + \log(1/\delta)}{\varepsilon^2}\right)$ "works"

• Neural Nets + floating point arithmetic



$$y = v^T \sigma(Wx)$$

$$|\mathcal{H}| = 2^{16 \cdot \#\text{ parameters/weights}}$$

$$n = O\left(\frac{p + \log(1/\delta)}{\varepsilon^2}\right)$$ samples

suffice for ε-gen. gap.

Warning: The above bounds are very pessimistic because:

- they don't apply to "infinite" classes
- they depend on # parameters of the model
- they are oblivious to the training algo (!)

Main take-aways:

- No matter what the learning problem is if

$$\#\text{samples} = O\left(\frac{\#\text{params}}{\varepsilon^2}\right)$$

then the generalization gap is small

- Smaller # params might be easier to generalize, but not necessary e.g., read:

- Bartlett, Peter L. "For valid generalization the size of the weights is more important than the size of the network." Advances in neural information processing systems. 1997.

- Bartlett, Peter L., Dylan J. Foster, and Matus J. Telgarsky. "Spectrally-normalized margin bounds for neural networks." Advances in Neural Information Processing Systems. 2017.

Next time:
- brief mentions of VC-dim and Rademacher complexity
- Examples of learning problems
- Computational Aspects
- Grad. Descent.