Mathematical Foundations of ML (CS 4785/5783)
Lecture 2

Statistical Learning and Uniform Convergence

http://www.cs.cornell.edu/Courses/cs4783/2022sp/notes02.pdf

Scenario II

On each round t:

Email x_t is composed, possibly by spammer!

System classifies email as $\hat{y}_t$

True label $y_t = f_{i^*}(x_t)$ revealed

We get feedback every round. But spammer can pick next email.

Goal: Make as few mistakes as possible.

Scenario II

How about using the same algorithm from scenario 1 for each t (re-run)?

How many mistakes would it make?

Ans: N-1

Scenario II

Algorithm:

$$\text{Pick } \mathcal{F}_t = \{f_i : i \in [N], \forall s < t, f_i(x_s) = y_s\}$$
$$\text{Set } \hat{y}_t = \text{Majority}(\{f(x_t) : f \in \mathcal{F}_t\})$$

Mistake Bound:

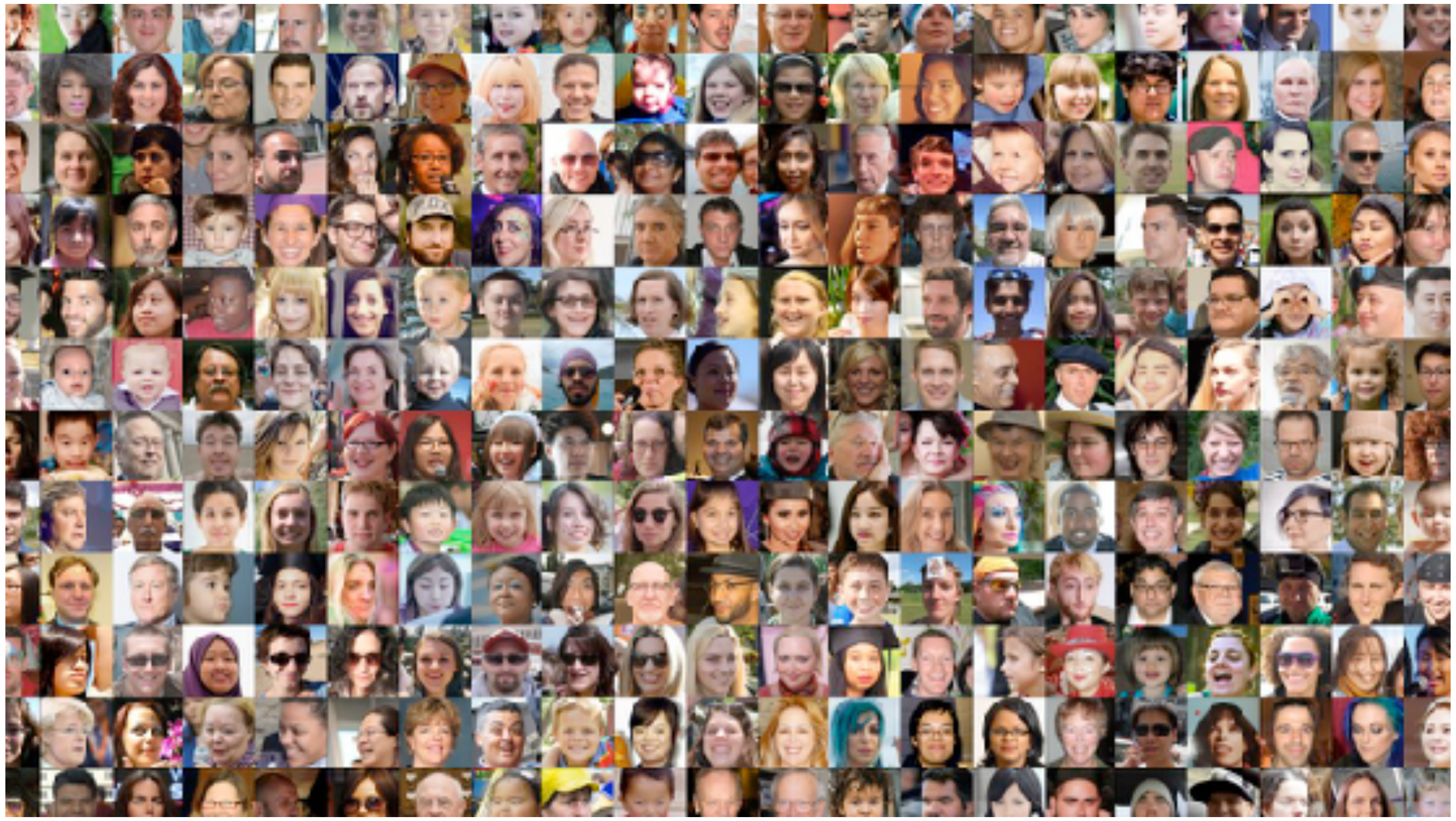
$$\sum_t \mathbf{1}\{\hat{y}_t \neq y_t\} \leq \log_2 N$$

Why?

Statistical Learning Framework

Eg: ML for Face recognition

$$\mathcal{X}: \text{ set of all images}$$



U

We don't have access to U, we just need the samples

When we deploy the system, do we really sample from U at random?


In summer


In winter

No assumption is right but some are useful!

Statistical Learning Framework

$$\mathbf{D} \text{ is a distribution on } \mathcal{X} \times \mathcal{Y}$$

D captures the idea of this set **U**

$$\text{Training sample } S = \{(x_1, y_1), \ldots, (x_n, y_n)\} \quad \text{Each } (x_t, y_t) \sim \mathbf{D}$$

$$\text{Risk of a model } g \text{ defined as } L_{\mathbf{D}}(g) = \mathbb{E}_{(x,y)\sim\mathbf{D}}\left[\ell(g(x), y)\right]$$

(Future instances drawn from **D**)

Excess risk of model g w.r.t. model class $\mathcal{F}$ defined as

$$L_{\mathbf{D}}(g) - \min_{f \in \mathcal{F}} L_{\mathbf{D}}(f)$$

Goal: provide an algorithm for which excess risk is small

Training Loss Vs Test Loss

$$\text{Training loss: } \widehat{L}_S(g) = \frac{1}{|S|} \sum_{(x,y) \in S} \ell(g(x), y)$$

Test loss: Draw fresh samples (not used by algorithm) and compute average error on that

Test loss is a good proxy for risk
(provided we never use it in any sense for training/parameter tuning etc.)

The Common Fallacy

$$\forall f \in \mathcal{F}, P\left(\left|L_{\mathbf{D}}(f) - \widehat{L}_S(f)\right| \text{ is large}\right) \text{ is small}$$

Algorithm picks $\hat{f}_S \in \mathcal{F}$ and so

$$P\left(\left|L_{\mathbf{D}}(\hat{f}_S) - \widehat{L}_S(\hat{f}_S)\right| \text{ is large}\right) \text{ is small}$$

THIS IS FALSE IN GENERAL!

Breakout room 3 mins

The Common Fallacy

- The issue with benchmark dataset like CIFAR and Imagenet
- Double edged sword

Empirical Risk Minimization

Pick a model in class that minimizes training error

$$\hat{f}_{\mathrm{ERM}} \in \arg\min_{f \in \mathcal{F}} \widehat{L}_S(f)$$

- When does this succeed?
  - When model class is too complex, we already saw this can fail
  - When model class is say just one function, it succeeds due to law of large numbers (concentration)
  - In general how well does this algorithm do?

ERM and Uniform Convergence

$$
\begin{aligned}
P\left(L_{\mathbf{D}}(\hat{f}_{\text{ERM}}) - \min_{f\in\mathcal{F}} L_{\mathbf{D}}(f) > 2\epsilon\right) &= P\left(L_{\mathbf{D}}(\hat{f}_{\text{ERM}}) - \widehat{L}_S(\hat{f}_{\text{ERM}}) + \widehat{L}_S(\hat{f}_{\text{ERM}}) - \min_{f\in\mathcal{F}} L_{\mathbf{D}}(f) > 2\epsilon\right) \\
&= P\left(L_{\mathbf{D}}(\hat{f}_{\text{ERM}}) - \widehat{L}_S(\hat{f}_{\text{ERM}}) + \max_{f\in\mathcal{F}}\left(\widehat{L}_S(\hat{f}_{\text{ERM}}) - L_{\mathbf{D}}(f)\right) > 2\epsilon\right) \\
&\leq P\left(L_{\mathbf{D}}(\hat{f}_{\text{ERM}}) - \widehat{L}_S(\hat{f}_{\text{ERM}}) + \max_{f\in\mathcal{F}}\left(\widehat{L}_S(f) - L_{\mathbf{D}}(f)\right) > 2\epsilon\right) \\
&\leq P\left(\max_{f\in\mathcal{F}}\left|\widehat{L}_S(f) - L_{\mathbf{D}}(f)\right| > \epsilon\right) \qquad (1)
\end{aligned}
$$

ERM Over Finite Class

If losses are bounded by 1 (in absolute) and $|\mathcal{F}| < \infty$, then, for any $\delta > 0$ with probability at least $1 - \delta$,

$$L_{\mathbf{D}}(\hat{f}_{\mathrm{ERM}}) - \min_{f \in \mathcal{F}} L_{\mathbf{D}}(f) \leq \sqrt{\frac{8 \log\left(2|\mathcal{F}|/\delta\right)}{n}}$$

ERM Over Finite Class

Hoeffding Inequality: Let $Z_1, \dots, Z_n$ be a sequence of n random variables bounded by 1, drawn iid from a fixed distribution. Then:

$$P\left(\left|\frac{1}{n}\sum_{t=1}^{n} Z_t - \mathbb{E}Z\right| > \epsilon\right) \le 2\exp\left(-\frac{n\epsilon^2}{2}\right)$$

ERM OVER FINITE CLASS

Hoeffding Inequality: Let $Z_1, \ldots, Z_n$ be a sequence of n random variables bounded by 1, drawn iid from a fixed distribution. Then:

$$P\left(\left|\frac{1}{n}\sum_{t=1}^{n} Z_t - \mathbb{E}Z\right| > \epsilon\right) \leq 2\exp\left(-\frac{n\epsilon^2}{2}\right)$$

Proof idea:

$$\text{For each } f \in \mathcal{F} \text{ define } Z_t^f = \ell(f(x_t), y_t)$$

Apply Hoeffding for each f individually

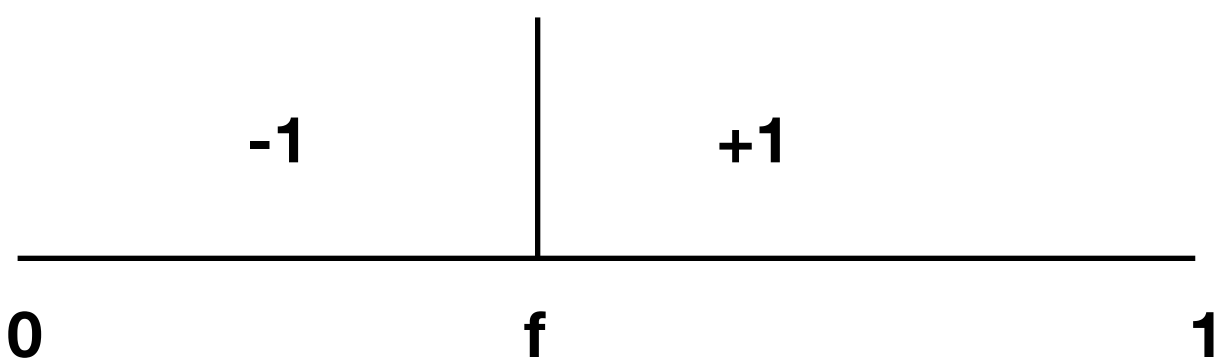
Use union bound to move to uniform deviation

Beyond Finite Model Class

- Idea 1: Find a finite set $\mathcal{F}'$ such that for any $f \in \mathcal{F}$ there exists an. $f' \in \mathcal{F}'$ s.t.

$$\forall x, y, \quad |\ell(f'(x), y) - \ell(f(x), y)| < \Delta$$

- But this may not always work, consider the example of learning thresholds:



$$\mathcal{X} = [0, 1] \qquad f(x) = \text{sign}(x - f)$$

$$\mathcal{F} \text{ indexed by set } [0, 1]$$

For any $\Delta < 1/2$, this class cannot be approximated by a finite set.