Week 4: Feasibility of Learning

Table of Contents

Inferring Probability via Sampling

For problems for which patterns cannot be learned, it might be possible to take advantage of probabilities inferred by taking samples from large population

Hoeffding's Inequality

1. **Hoeffding's inequality** is one of the most important inequalities of learning theory, for bounding the probability that sums of bounded random variables differ from their respective expected values.
2. Definition: Let $X_1, \ldots, X_n$ be *independent bounded random variabes* with $X_i \in [a, b]$ for all i, where $-\infty < a \leq b < \infty$. Then for all $t > 0$:

$$\mathbb{P}(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i] \geq t)) \leq exp(-\frac{2nt^2}{(b-a)^2})$$

and

$$\mathbb{P}(\frac{1}{n}\sum_{i=1}^{n}(\mathbb{E}[X_i] - X_i \geq t)) \leq exp(-\frac{2nt^2}{(b-a)^2})$$

Hoeffding's Inequality and Error Probability

1. Combining the two equations above (think of adding up tails at both ends of an error distribution), we get: let

$$S_n = \sum_{i=1}^{n} X_i$$

, then there is:

$$\mathbb{P}(|S_n - \mathbb{E}[S_n]| \geq \epsilon)) \leq 2exp(-\frac{2n\epsilon^2}{(b-a)^2})$$

2. Special case of Hoeffding's Inequality for Bernoulli variables

- For Bernoullli random variables, where possible values are only 0 and 1, equation above further simplifies to:

$$\mathbb{P}(|S_n - \mathbb{E}[S_n]| \geq \epsilon)) \leq 2e^{-2n\epsilon^2}$$

Inferring Unlearnable Population Distribution Via Sampling

1. Assume the unlearnable population distribution μ, given a large *random* sample of size N from the population, and the repective sample distribution ν, probability that ν closely approximates μ can be bounded using Hoeffding's Inequality:

$$\mathbb{P}(|\nu - \mu| \geq \epsilon)) \leq 2e^{-2N\epsilon^2}$$

2. When sample size N is sufficiently large, the statement $\nu = \mu$ is **probably approximately correct (PAC)**
3. The approximation bound determined by Hoeffding's Inequality:
   - Works for all N and ϵ
   - Does **not** depend on value of population distribution μ, does not need to know/learn μ
   - By definition, having larger sample size N, or more lenient error bound ϵ results in higher probabiliy of $\nu \approx \mu$ (assuming the approximated equality is defined by difference within error bound)
   - With a sufficiently large N and restrictive error bound, we can *probably* infer the **unknown** population distribution μ by the **known** sample distribution ν

Guaranteed Bound on Out-of-Sample Error

1. Given Hoeffding's Inequality, when sample size N is sufficiently large, the unknown out-of-sample error rate when a model is applied to the **population**, $[h(x) \neq f(x)]$ can be inferred from the **known** out-of sample error rate when the model is applied to a **sample** (different from training sample, drawn from the same population) drawn from the population, $[h(x_n) \neq y_n]$

- For any fixed hypothesis h, Hoeffinding's Inequality suggests that it's possible to infer unknown $E_{out}(h) = \varepsilon_{x\ P}[h(x) \neq f(x)]$ by known

$$E_{in}(h) = \frac{1}{N} \sum_{n=1} N[h(x_n) \neq y_n]$$

   - Hypothesis h is fixed because the target f is assumed to be fixed

2. Formal guarantee for out-of-sample (OTS) error, given in-sample error:

- For any fixed h, when sample size N is sufficiently large, in-sample error $E_{in}(h)$ is within an error bound *epsilon* to out-of-sample error $E_{out}(h)$

$$\mathbb{P}[|E_{in}(h) - E_{out}(h)| > \epsilon] \leq 2e^{-2N\epsilon^2}$$

- Valid for all N and ϵ
- Does **not** depend on out-of-sample error $E_{out}(h)$, thus holds true even when population is too large that target function f and OTS error $E_{out}(h)$ are practically unknown
- $E_{in}(h) = E_{out}(h)$ is probably approximately correct (PAC)

3. Given formal guarantee above, if a fixed hypothesis h gives small in-sample error, it is likely within good approximation of the target function $f

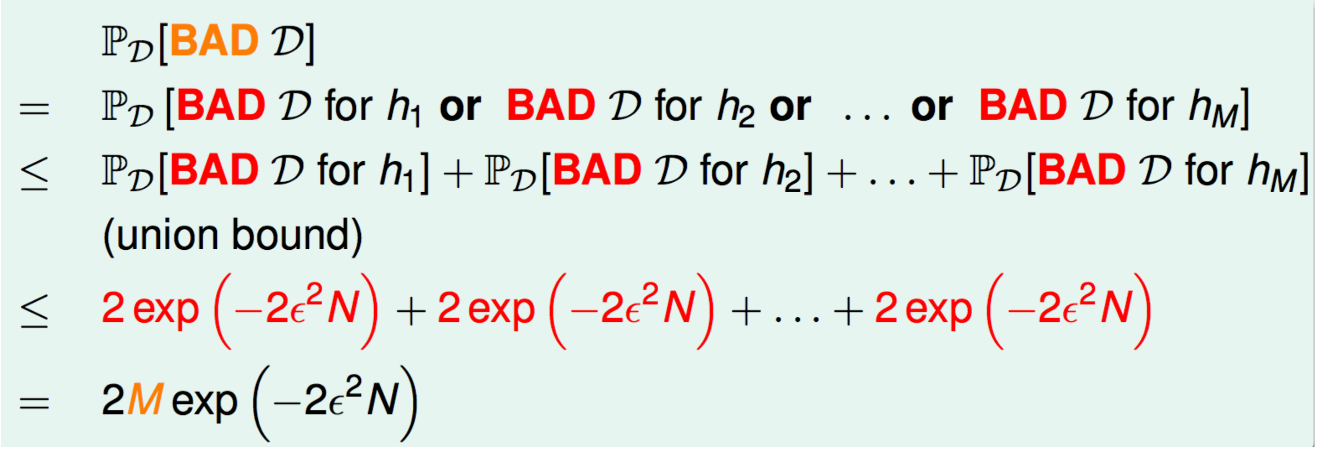
Hoffeding's Ineqaulity When Accounting for Bad Samples

1. It's possible that a biased sample is selected for computing and optimizing $E_{in}(h)$, resulting in high OTS error despite good performance on training data. Even worse, when multiple possible hypotheses are present, bad samples could force the learning algorithm into choosing one that's far from optimal.
2. Hoeffding's Inequality gurantees that, if were to exhaust all possible samples from the given population, the probability of landing on a biased sample is small.

$$\mathbb{P}_{\mathcal{D}}[\text{BAD D}] = \sum_{\text{all possible D}} \mathbb{P}(\mathcal{D}) \cdot [\text{BAD D}] \rightarrow \text{Small}$$

3. When multiple hypotheses (e.g. M candidate hypotheses $h_1, h_2, \ldots, h_M$) are available, finite-bin Hoeffding still guarantees a bound on the possiblility of encountering bad samples, when all possible samples are exhausted.

- A sample is considered "bad" in a multi-hypotheses scenario if there exists some h for which $E_{in}(h)$ and $E_{out}(h)$ are very different.



$$\begin{aligned}
& \mathbb{P}_{\mathcal{D}}[\text{BAD } \mathcal{D}] \\
= \;& \mathbb{P}_{\mathcal{D}}[\text{BAD } \mathcal{D} \text{ for } h_1 \textbf{ or } \text{BAD } \mathcal{D} \text{ for } h_2 \textbf{ or } \ldots \textbf{ or } \text{BAD } \mathcal{D} \text{ for } h_M] \\
\leq \;& \mathbb{P}_{\mathcal{D}}[\text{BAD } \mathcal{D} \text{ for } h_1] + \mathbb{P}_{\mathcal{D}}[\text{BAD } \mathcal{D} \text{ for } h_2] + \ldots + \mathbb{P}_{\mathcal{D}}[\text{BAD } \mathcal{D} \text{ for } h_M] \\
& \text{(union bound)} \\
\leq \;& 2\exp\left(-2\epsilon^2 N\right) + 2\exp\left(-2\epsilon^2 N\right) + \ldots + 2\exp\left(-2\epsilon^2 N\right) \\
= \;& 2M\exp\left(-2\epsilon^2 N\right)
\end{aligned}$$

4. Finite-bin Hoeffding's Inequality:

- Does not depend on any $E_{out}(h_m)$. Target f and the real probability of getting bad sample can stay unknown
- For the optimal hypothesis g, $E_{in}(g) = E_{out}(g)$ is PAC, **regardless of learning algorithm** used to pick the hypothesis

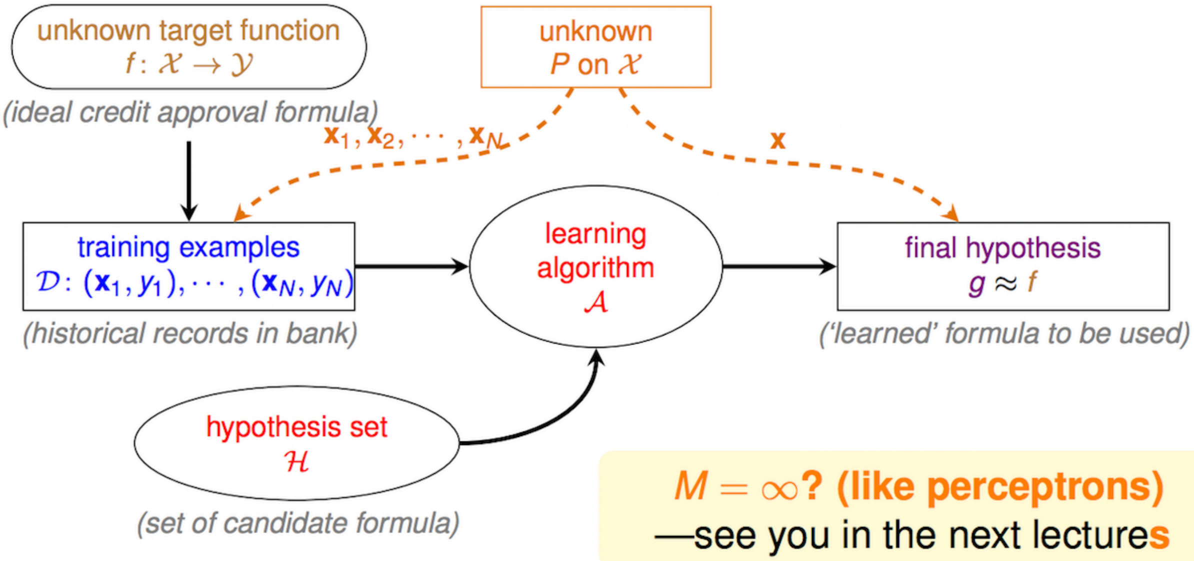
The ‘Statistical’ Learning Flow

if $|\mathcal{H}| = M$ finite, N large enough,

for whatever g picked by $\mathcal{A}$, $E_{\text{out}}(g) \approx E_{\text{in}}(g)$

if $\mathcal{A}$ finds one g with $E_{\text{in}}(g) \approx 0$,

PAC guarantee for $E_{\text{out}}(g) \approx 0 \Longrightarrow$ **learning possible :-)**



unknown target function
$f\colon \mathcal{X} \to \mathcal{Y}$
(ideal credit approval formula)
unknown
P on $\mathcal{X}$
$\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N$
$\mathbf{x}$
training examples
$\mathcal{D}\colon (\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_N, y_N)$
(historical records in bank)
learning
algorithm
$\mathcal{A}$
final hypothesis
$g \approx f$
(‘learned’ formula to be used)
hypothesis set
$\mathcal{H}$
(set of candidate formula)

$M = \infty$? (like perceptrons)

—see you in the next lectures