ORIE 4741: Learning with Big Messy Data

Generalization

Professor Udell

Operations Research and Information Engineering
Cornell

October 16, 2021

Announcements 9/30/21

- hw2 due Thursday morning 9:15am
- hw3 is out; do it early to enjoy Fall break
  - save slip days for emergencies
- project proposals due Sunday 11:59pm
  - final project must use at least 3 techniques from class
- section next week: generalization and validation

Generalization and Overfitting

- goal of model is **not** to predict well on $\mathcal{D}$
- goal of model is to predict well **on new data**

if the model has ____ training set error and ____ test set error, we say the model:

	low test set error	high test set error
low training set error	generalizes	overfits
high training set error	?!?!	underfits

Simplest case: generalizing from a mean

exit polling

- sample n voters leaving polling places
- for each voter i, define the Boolean random variable

$$z_i = \begin{cases} 1 & \text{if voter } i \text{ voted for Biden} \\ 0 & \text{otherwise} \end{cases}$$

- sample mean: $\nu = \frac{1}{n} \sum_{i=1}^{n} z_i$
- true mean: $\mu = \mathbb{E}_{i \sim \text{US electorate}} z_i$ is Biden's expected vote share

Simplest case: generalizing from a mean

exit polling

- sample n voters leaving polling places
- for each voter i, define the Boolean random variable

$$z_i = \begin{cases} 1 & \text{if voter } i \text{ voted for Biden} \\ 0 & \text{otherwise} \end{cases}$$

- sample mean: $\nu = \frac{1}{n}\sum_{i=1}^{n} z_i$
- true mean: $\mu = \mathbb{E}_{i \sim \text{US electorate}} z_i$ is Biden's expected vote share

Q: When does sample mean ν estimate true mean μ well?

Simplest case: generalizing from a mean

exit polling

- sample n voters leaving polling places
- for each voter i, define the Boolean random variable

$$z_i = \begin{cases} 1 & \text{if voter } i \text{ voted for Biden} \\ 0 & \text{otherwise} \end{cases}$$

- sample mean: $\nu = \frac{1}{n} \sum_{i=1}^{n} z_i$
- true mean: $\mu = \mathbb{E}_{i \sim \text{US electorate}} z_i$ is Biden's expected vote share

Q: When does sample mean ν estimate true mean μ well?
A: (1) sample voters uniformly from all voters (2) n large!

Simplest case: generalizing from a mean

exit polling

- sample n voters leaving polling places
- for each voter i, define the Boolean random variable

$$z_i = \begin{cases} 1 & \text{if voter } i \text{ voted for Biden} \\ 0 & \text{otherwise} \end{cases}$$

- sample mean: $\nu = \frac{1}{n}\sum_{i=1}^{n} z_i$
- true mean: $\mu = \mathbb{E}_{i \sim \text{US electorate}} z_i$ is Biden's expected vote share

Q: When does sample mean ν estimate true mean μ well?
A: (1) sample voters uniformly from all voters (2) n large!
Q: Why might these conditions fail to hold?

Simplest case: generalizing from a mean

exit polling

- sample n voters leaving polling places
- for each voter i, define the Boolean random variable

$$z_i = \begin{cases} 1 & \text{if voter } i \text{ voted for Biden} \\ 0 & \text{otherwise} \end{cases}$$

- sample mean: $\nu = \frac{1}{n} \sum_{i=1}^{n} z_i$
- true mean: $\mu = \mathbb{E}_{i \sim \text{US electorate}} z_i$ is Biden's expected vote share

Q: When does sample mean ν estimate true mean μ well?
A: (1) sample voters uniformly from all voters (2) n large!
Q: Why might these conditions fail to hold?
A: Absentee votes; failure to sample small or remote polling places; voters who refuse to answer; limited polling resources

Poll: true mean and sample mean

Suppose voters in our polling sample are uniformly sampled from the set of all voters, and give truthful answers. The strong law of large numbers states that sample mean converges to the true mean

A. false

B. true

C. as the number of samples $n \to \infty$

D. as the number of voters in the US $\to \infty$

E. so long as the poll is conducted by a respectable nonpartisan organization

Hoeffding inequality

Theorem (Hoeffding Inequality)

Let $z_i \in \{0,1\}$, $i = 1, \ldots, n$, be independent Boolean random variables with mean $\mathbb{E}z_i = \mu$. Define the sample mean $\nu = \frac{1}{n}\sum_{i=1}^{n} z_i$. Then for any $\epsilon > 0$,

$$\mathbb{P}[|\nu - \mu| > \varepsilon] \leq 2\exp\left(-2\varepsilon^2 n\right).$$

an example of a **concentration inequality**

Hoeffding inequality

Theorem (Hoeffding Inequality)

Let $z_i \in \{0, 1\}$, $i = 1, \ldots, n$, be independent Boolean random variables with mean $\mathbb{E} z_i = \mu$. Define the sample mean $\nu = \frac{1}{n} \sum_{i=1}^{n} z_i$. Then for any $\epsilon > 0$,

$$\mathbb{P}[|\nu - \mu| > \varepsilon] \leq 2 \exp\left(-2\varepsilon^2 n\right).$$

an example of a **concentration inequality**

- μ can't be much higher than ν
- μ can't be much lower than ν
- more samples n improve estimate **exponentially** quickly

Compare with law of large numbers

Theorem (Strong Law of Large Numbers)

Let $z_i \in \mathbf{R}$, $i = 1, \dots, n$, be independent random variables with mean $\mathbb{E}z_i = \mu$. Define the sample mean $\nu = \frac{1}{n}\sum_{i=1}^{n} z_i$. Then

$$\nu \to \mu \qquad \text{as} \qquad n \to \infty$$

compare with the Hoeffding bound:

- the Hoeffding bound provides **quantitative** predictions on how fast the sample mean ν **concentrates** near μ.
- the Hoeffding bound only holds for Boolean random variables
- similar **concentration inequalities** (named, *e.g.*, Azuma, McDiarmid, Bennet, Bernstein, Chernoff, ...) hold for other kinds of random variables

Back to the learning problem

fix a hypothesis $h : \mathcal{X} \to \mathcal{Y}$. take

$$
\begin{aligned}
z_i &= \begin{cases} 1 & y_i = h(x_i) \\ 0 & \text{otherwise} \end{cases} \\
&= \mathbb{1}(y_i = h(x_i))
\end{aligned}
$$

Back to the learning problem

fix a hypothesis $h : \mathcal{X} \to \mathcal{Y}$. take

$$
\begin{aligned}
z_i &= \begin{cases} 1 & y_i = h(x_i) \\ 0 & \text{otherwise} \end{cases} \\
&= \mathbb{1}(y_i = h(x_i))
\end{aligned}
$$

example. build a model of voting behavior:

- y_i is 1 if voter i voted for Biden, 0 otherwise
- $h(x_i)$ is our guess of how the voter will vote, using hypothesis h

Back to the learning problem

fix a hypothesis $h : \mathcal{X} \to \mathcal{Y}$. take

$$\begin{aligned} z_i &= \begin{cases} 1 & y_i = h(x_i) \\ 0 & \text{otherwise} \end{cases} \\ &= \mathbb{1}(y_i = h(x_i)) \end{aligned}$$

example. build a model of voting behavior:

- y_i is 1 if voter i voted for Biden, 0 otherwise
- $h(x_i)$ is our guess of how the voter will vote, using hypothesis h
- $z_i = \mathbb{1}(y_i = h(x_i))$ is 1 if we guess correctly for voter i, 0 otherwise

Back to the learning problem

fix a hypothesis $h : \mathcal{X} \to \mathcal{Y}$. take

$$
\begin{aligned}
z_i &= \begin{cases} 1 & y_i = h(x_i) \\ 0 & \text{otherwise} \end{cases} \\
&= \mathbb{1}(y_i = h(x_i))
\end{aligned}
$$

example. build a model of voting behavior:

- y_i is 1 if voter i voted for Biden, 0 otherwise
- $h(x_i)$ is our guess of how the voter will vote, using hypothesis h
- $z_i = \mathbb{1}(y_i = h(x_i))$ is 1 if we guess correctly for voter i, 0 otherwise
- z_i depends on x_i, y_i, and h

Adding in probability

make our model probabilistic:

- fix a probability distribution $P(x, y)$
- sample (x_i, y_i) iid[1] from $P(x, y)$
- form data set $\mathcal{D}$ by sampling:
  - for $i = 1, \ldots, n$
    - sample $(x_i, y_i) \sim P(x, y)$
  - set $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$

[1] iid: independent and identically distributed

Adding in probability

make our model probabilistic:

- fix a probability distribution $P(x, y)$
- sample (x_i, y_i) iid[1] from $P(x, y)$
- form data set $\mathcal{D}$ by sampling:
  - for $i = 1, \ldots, n$
    - sample $(x_i, y_i) \sim P(x, y)$
  - set $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$

special case. $y = f(x)$ is deterministic conditioned on x:

$$
\begin{aligned}
P(y|x) &= \begin{cases} 1 & y = f(x) \\ 0 & \text{otherwise} \end{cases} \\
P(x, y) &= P(x)P(y|x) = \begin{cases} P(x) & y = f(x) \\ 0 & \text{otherwise} \end{cases}
\end{aligned}
$$

[1] iid: independent and identically distributed

Hoeffding for the noisy learning problem

- fix a hypothesis $h : \mathcal{X} \to \mathcal{Y}$.
- draw samples (x_i, y_i) iid from $P(x, y)$ to form $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$

Hoeffding for the noisy learning problem

- fix a hypothesis $h : \mathcal{X} \to \mathcal{Y}$.
- draw samples (x_i, y_i) iid from $P(x, y)$ to form $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$
- take $z_i = \mathbb{1}(y_i = h(x_i))$

Hoeffding for the noisy learning problem

- fix a hypothesis $h : \mathcal{X} \to \mathcal{Y}$.
- draw samples (x_i, y_i) iid from $P(x, y)$ to form $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$
- take $z_i = \mathbb{1}(y_i = h(x_i))$
- z_i are iid (since (x_i, y_i) are iid, and h is fixed)

Hoeffding for the noisy learning problem

- fix a hypothesis $h : \mathcal{X} \to \mathcal{Y}$.
- draw samples (x_i, y_i) iid from $P(x, y)$ to form $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$
- take $z_i = \mathbb{1}(y_i = h(x_i))$
- z_i are iid (since (x_i, y_i) are iid, and h is fixed)
- $\mathbb{E}z = \mathbb{E}_{(x,y)\sim P(x,y)} \mathbb{1}(y = h(x))$

Hoeffding for the noisy learning problem

- fix a hypothesis $h : \mathcal{X} \to \mathcal{Y}$.
- draw samples (x_i, y_i) iid from $P(x, y)$ to form $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$
- take $z_i = \mathbb{1}(y_i = h(x_i))$
- z_i are iid (since (x_i, y_i) are iid, and h is fixed)
- $\mathbb{E}z = \mathbb{E}_{(x,y)\sim P(x,y)} \mathbb{1}(y = h(x))$

so we can apply Hoeffding! for any $\epsilon > 0$,

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} z_i - \mathbb{E}z\right| > \varepsilon\right] \leq 2\exp\left(-2\varepsilon^2 n\right)$$

Hoeffding for the noisy learning problem

- fix a hypothesis $h : \mathcal{X} \to \mathcal{Y}$.
- draw samples (x_i, y_i) iid from $P(x, y)$ to form $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$
- take $z_i = \mathbb{1}(y_i = h(x_i))$
- z_i are iid (since (x_i, y_i) are iid, and h is fixed)
- $\mathbb{E}z = \mathbb{E}_{(x,y)\sim P(x,y)} \mathbb{1}(y = h(x))$

so we can apply Hoeffding! for any $\epsilon > 0$,

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} z_i - \mathbb{E}z\right| > \varepsilon\right] \leq 2\exp\left(-2\varepsilon^2 n\right)$$

Q: Probability? Where's the randomness here?

Hoeffding for the noisy learning problem

- fix a hypothesis $h : \mathcal{X} \to \mathcal{Y}$.
- draw samples (x_i, y_i) iid from $P(x, y)$ to form $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$
- take $z_i = \mathbb{1}(y_i = h(x_i))$
- z_i are iid (since (x_i, y_i) are iid, and h is fixed)
- $\mathbb{E}z = \mathbb{E}_{(x,y)\sim P(x,y)} \mathbb{1}(y = h(x))$

so we can apply Hoeffding! for any $\epsilon > 0$,

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} z_i - \mathbb{E}z\right| > \varepsilon\right] \leq 2\exp\left(-2\varepsilon^2 n\right)$$

Q: Probability? Where's the randomness here?
A: The dataset $\mathcal{D}$ is random, drawn iid according to $P(x, y)$

Hoeffding for the noisy learning problem

- fix a hypothesis $h : \mathcal{X} \to \mathcal{Y}$.
- draw samples (x_i, y_i) iid from $P(x, y)$ to form $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$
- take $z_i = \mathbb{1}(y_i = h(x_i))$
- z_i are iid (since (x_i, y_i) are iid, and h is fixed)
- $\mathbb{E}z = \mathbb{E}_{(x,y)\sim P(x,y)} \mathbb{1}(y = h(x))$

so we can apply Hoeffding! for any $\epsilon > 0$,

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} z_i - \mathbb{E}z\right| > \varepsilon\right] \leq 2\exp\left(-2\varepsilon^2 n\right)$$

Q: Probability? Where's the randomness here?
A: The dataset $\mathcal{D}$ is random, drawn iid according to $P(x, y)$
Q: Is $\frac{1}{n}\sum_{i=1}^{n} z_i$ more like training set error or test set error?

Hoeffding for the noisy learning problem

- fix a hypothesis $h : \mathcal{X} \to \mathcal{Y}$.
- draw samples (x_i, y_i) iid from $P(x, y)$ to form $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$
- take $z_i = \mathbb{1}(y_i = h(x_i))$
- z_i are iid (since (x_i, y_i) are iid, and h is fixed)
- $\mathbb{E}z = \mathbb{E}_{(x,y)\sim P(x,y)} \mathbb{1}(y = h(x))$

so we can apply Hoeffding! for any $\epsilon > 0$,

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} z_i - \mathbb{E}z\right| > \varepsilon\right] \leq 2\exp\left(-2\varepsilon^2 n\right)$$

Q: Probability? Where's the randomness here?
A: The dataset $\mathcal{D}$ is random, drawn iid according to $P(x, y)$
Q: Is $\frac{1}{n}\sum_{i=1}^{n} z_i$ more like training set error or test set error?
A: It's more like test set error, since h is independent of $\mathcal{D}$

In-sample and out-of-sample error

some new terminology:

- **in-sample error.**

$$\begin{aligned} E_{\text{in}}(h) &= \text{fraction of } \mathcal{D} \text{ where } y_i \text{ and } h(x_i) \text{ disagree} \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(y_i \neq h(x_i)) \end{aligned}$$

- **out-of-sample error.**

$$\begin{aligned} E_{\text{out}}(h) &= \text{probability that } y \text{ and } h(x) \text{ disagree} \\ &= \mathop{\mathbb{P}}_{(x,y)\sim P(x,y)}[y \neq h(x)] \end{aligned}$$

In-sample and out-of-sample error

some new terminology:

- **in-sample error.**

$$\begin{aligned} E_{\text{in}}(h) &= \text{fraction of } \mathcal{D} \text{ where } y_i \text{ and } h(x_i) \text{ disagree} \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(y_i \neq h(x_i)) \end{aligned}$$

- **out-of-sample error.**

$$\begin{aligned} E_{\text{out}}(h) &= \text{probability that } y \text{ and } h(x) \text{ disagree} \\ &= \mathop{\mathbb{P}}_{(x,y)\sim P(x,y)} [y \neq h(x)] \end{aligned}$$

notice

$$E_{\text{out}}(h) = \mathbb{E}\left[E_{\text{in}}(h)\right]$$

Hoeffding for the noisy learning problem

- fix a hypothesis $h : \mathcal{X} \to \mathcal{Y}$.
- consider (x_i, y_i) as samples drawn from $P(x, y)$

Hoeffding for the noisy learning problem

- fix a hypothesis $h : \mathcal{X} \to \mathcal{Y}$.
- consider (x_i, y_i) as samples drawn from $P(x, y)$
- take $z_i = \mathbb{1}(y_i \neq h(x_i))$
- z_i are iid (since (x_i, y_i) are iid, and h is fixed)

Hoeffding for the noisy learning problem

- fix a hypothesis $h : \mathcal{X} \to \mathcal{Y}$.
- consider (x_i, y_i) as samples drawn from $P(x, y)$
- take $z_i = \mathbb{1}(y_i \neq h(x_i))$
- z_i are iid (since (x_i, y_i) are iid, and h is fixed)
- $\mathbb{E}z = \mathbb{E}_{(x,y)\sim P(x,y)}\mathbb{1}(y \neq h(x)) = E_{\text{out}}(h)$

Hoeffding for the noisy learning problem

- fix a hypothesis $h : \mathcal{X} \to \mathcal{Y}$.
- consider (x_i, y_i) as samples drawn from $P(x, y)$
- take $z_i = \mathbb{1}(y_i \neq h(x_i))$
- z_i are iid (since (x_i, y_i) are iid, and h is fixed)
- $\mathbb{E}z = \mathbb{E}_{(x,y)\sim P(x,y)} \mathbb{1}(y \neq h(x)) = E_{\text{out}}(h)$
- $\frac{1}{n} \sum_{i=1}^{n} z_i = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(y_i \neq h(x_i)) = E_{\text{in}}(h)$

Hoeffding for the noisy learning problem

- fix a hypothesis $h : \mathcal{X} \to \mathcal{Y}$.
- consider (x_i, y_i) as samples drawn from $P(x, y)$
- take $z_i = \mathbb{1}(y_i \neq h(x_i))$
- z_i are iid (since (x_i, y_i) are iid, and h is fixed)
- $\mathbb{E}z = \mathbb{E}_{(x,y)\sim P(x,y)} \mathbb{1}(y \neq h(x)) = E_{\text{out}}(h)$
- $\frac{1}{n}\sum_{i=1}^{n} z_i = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(y_i \neq h(x_i)) = E_{\text{in}}(h)$

apply Hoeffding: for any $\varepsilon > 0$,

$$\mathbb{P}[|E_{\text{in}}(h) - E_{\text{out}}(h)| > \varepsilon] \leq 2 \exp\left(-2\varepsilon^2 n\right)$$

Does Hoeffding work for our learned model?

two scenarios:

1. Without looking at any data, pick a model $h : \mathcal{X} \to \mathcal{Y}$ to predict who will vote for Biden. Then sample data $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, and set $z_i = \mathbb{1}(y_i = h(x_i))$.

Does Hoeffding work for our learned model?

two scenarios:

1. Without looking at any data, pick a model $h : \mathcal{X} \to \mathcal{Y}$ to predict who will vote for Biden. Then sample data $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, and set $z_i = \mathbb{1}(y_i = h(x_i))$.
2. Sample the data $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, and use it to develop a model $g : \mathcal{X} \to \mathcal{Y}$ to predict who will vote for Biden (*e.g.*, using perceptron). Set $z_i' = \mathbb{1}(y_i = g(x_i))$.

Is the sample mean $\frac{1}{n}\sum_{i=1}^{n} z_i$ a good estimate for the expected performance $\mathbb{E}z$? Is $\frac{1}{n}\sum_{i=1}^{n} z_i'$ a good estimate for $\mathbb{E}z'$?

Does Hoeffding work for our learned model?

two scenarios:

1. Without looking at any data, pick a model $h : \mathcal{X} \to \mathcal{Y}$ to predict who will vote for Biden. Then sample data $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, and set $z_i = \mathbb{1}(y_i = h(x_i))$.
2. Sample the data $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, and use it to develop a model $g : \mathcal{X} \to \mathcal{Y}$ to predict who will vote for Biden (*e.g.*, using perceptron). Set $z_i' = \mathbb{1}(y_i = g(x_i))$.

Are the z_is iid?

A. yes
B. no

Does Hoeffding work for our learned model?

two scenarios:

1. Without looking at any data, pick a model $h : \mathcal{X} \to \mathcal{Y}$ to predict who will vote for Biden. Then sample data $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, and set $z_i = \mathbb{1}(y_i = h(x_i))$.
2. Sample the data $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, and use it to develop a model $g : \mathcal{X} \to \mathcal{Y}$ to predict who will vote for Biden (*e.g.*, using perceptron). Set $z_i' = \mathbb{1}(y_i = g(x_i))$.

Are the z_is iid?

A. yes
B. no

Are the z_i's iid?

A. yes
B. no

Does Hoeffding work for our learned model?

two scenarios:

1. Without looking at any data, pick a model $h : \mathcal{X} \to \mathcal{Y}$ to predict who will vote for Biden. Then sample data $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, and set $z_i = \mathbb{1}(y_i = h(x_i))$.
2. Sample the data $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, and use it to develop a model $g : \mathcal{X} \to \mathcal{Y}$ to predict who will vote for Biden (*e.g.*, using perceptron). Set $z_i' = \mathbb{1}(y_i = g(x_i))$.

Are the z_is iid?

A. yes
B. no

Are the z_i's iid?

A. yes
B. no

the z_i's depend on g, which depends on the whole data set $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$.

Does Hoeffding work for our learned model?

two scenarios:

1. Without looking at any data, pick a model $h : \mathcal{X} \to \mathcal{Y}$ to predict who will vote for Biden. Then sample data $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, and set $z_i = \mathbb{1}(y_i = h(x_i))$.
2. Sample the data $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, and use it to develop a model $g : \mathcal{X} \to \mathcal{Y}$ to predict who will vote for Biden (*e.g.*, using perceptron). Set $z_i' = \mathbb{1}(y_i = g(x_i))$.

Does the Hoeffding bound apply to the sample mean of the (iid) z_is?

A. yes
B. no

Does Hoeffding work for our learned model?

two scenarios:

1. Without looking at any data, pick a model $h : \mathcal{X} \to \mathcal{Y}$ to predict who will vote for Biden. Then sample data $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, and set $z_i = \mathbb{1}(y_i = h(x_i))$.
2. Sample the data $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, and use it to develop a model $g : \mathcal{X} \to \mathcal{Y}$ to predict who will vote for Biden (*e.g.*, using perceptron). Set $z_i' = \mathbb{1}(y_i = g(x_i))$.

Does the Hoeffding bound apply to the sample mean of the (iid) z_is?

A. yes
B. no

Does the Hoeffding bound apply to the sample mean of the (not iid) z_i's?

A. yes
B. no

Does Hoeffding work for our learned model?

two scenarios:

1. Without looking at any data, pick a model $h : \mathcal{X} \to \mathcal{Y}$ to predict who will vote for Biden. Then sample data $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, and set $z_i = \mathbb{1}(y_i = h(x_i))$.
2. Sample the data $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, and use it to develop a model $g : \mathcal{X} \to \mathcal{Y}$ to predict who will vote for Biden (*e.g.*, using perceptron). Set $z_i' = \mathbb{1}(y_i = g(x_i))$.

Is the sample mean $\frac{1}{n}\sum_{i=1}^{n} z_i$ a good estimate for the expected performance $\mathbb{E}z$? Is $\frac{1}{n}\sum_{i=1}^{n} z_i'$ a good estimate for $\mathbb{E}z'$?

Q: Are the z_is iid? What about the z_i's?
A: z_is are iid. z_i's are not independent: they depend on g, which depends on the whole data set $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$.
Q: Does Hoeffding apply to the first? the second?
A: Hoeffding applies to first, not to second.
Extreme case for second scenario: model memorizes the data.

Recall validation procedure

how to decide which model to use?

- split data into training set $\mathcal{D}_{\text{train}}$ and test set $\mathcal{D}_{\text{valid}}$
- pick m different interesting model classes
  e.g., different ϕs: ϕ_1, ϕ_2, ..., ϕ_m
- fit ("train") models on training set $\mathcal{D}_{\text{train}}$
  get one model $h : \mathcal{X} \to \mathcal{Y}$ for each ϕs, and set

$$\mathcal{H} = \{h_1, h_2, \dots, h_m\}$$

- compute error of each model on test set $\mathcal{D}_{\text{valid}}$ and choose lowest:

$$g = \operatorname*{argmin}_{h \in \mathcal{H}} E_{\mathcal{D}_{\text{valid}}}(h)$$

Recall validation procedure

how to decide which model to use?

- split data into training set $\mathcal{D}_{\text{train}}$ and test set $\mathcal{D}_{\text{valid}}$
- pick m different interesting model classes
  e.g., different ϕs: ϕ_1, ϕ_2, ..., ϕ_m
- fit ("train") models on training set $\mathcal{D}_{\text{train}}$
  get one model $h : \mathcal{X} \to \mathcal{Y}$ for each ϕs, and set

$$\mathcal{H} = \{h_1, h_2, \ldots, h_m\}$$

- compute error of each model on test set $\mathcal{D}_{\text{valid}}$ and choose lowest:

$$g = \operatorname*{argmin}_{h \in \mathcal{H}} E_{\mathcal{D}_{\text{valid}}}(h)$$

Q: Are $\{z_i = \mathbb{1}(y_i = g(x_i) : (x_i, y_i) \in \mathcal{D}_{\text{valid}})\}$ independent?

Recall validation procedure

how to decide which model to use?

- split data into training set $\mathcal{D}_{\text{train}}$ and test set $\mathcal{D}_{\text{valid}}$
- pick m different interesting model classes
  e.g., different ϕs: ϕ_1, ϕ_2, ..., ϕ_m
- fit ("train") models on training set $\mathcal{D}_{\text{train}}$
  get one model $h : \mathcal{X} \to \mathcal{Y}$ for each ϕs, and set

  $$\mathcal{H} = \{h_1, h_2, \ldots, h_m\}$$

- compute error of each model on test set $\mathcal{D}_{\text{valid}}$ and choose lowest:

  $$g = \operatorname*{argmin}_{h \in \mathcal{H}} E_{\mathcal{D}_{\text{valid}}}(h)$$

Q: Are $\{z_i = \mathbb{1}(y_i = g(x_i) : (x_i, y_i) \in \mathcal{D}_{\text{valid}})\}$ independent?
A: No; g was trained on $\mathcal{D}_{\text{valid}}$!
Hoeffding does not directly apply:
$E_{\mathcal{D}_{\text{valid}}}(g)$ may not accurately estimate $E_{\text{out}}(g)$

The union bound

recall the **union bound**: for two random events A and B,

$$\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$$

Poll: the union bound

In Ithaca, the probability of rain on any given day is 30%. The probability of sun on any given day is 50%. What is the probability p that there will be sun or rain on any given day?

A. $\leq$ 30%: $p \leq .30$

B. between 30% and 50%: $p \in (.30, .50]$

C. between 50% and 80%: $p \in (.50, .80]$

D. $>$ 80%: $p > 80$

Poll: when is the union bound tight?

In some other hypothetical city, the probability of rain on any given day is 30%; the probability of sun on any given day is 50%; and the probability of sun or rain on any given day is 80%. What can we say about the probability p that it will be sunny **and** rain on the same day?

A. $p = 0$

B. $p \in (0, .30]$

C. $p \in (.30, .50]$

D. $p \in (.50, .80]$

E. $p > .80$

Poll: when is the union bound tight?

In some other hypothetical city, the probability of rain on any given day is 30%; the probability of sun on any given day is 50%; and the probability of sun or rain on any given day is 80%. What can we say about the probability p that it will be sunny **and** rain on the same day?

A. $p = 0$
B. $p \in (0, .30]$
C. $p \in (.30, .50]$
D. $p \in (.50, .80]$
E. $p > .80$

Q: More generally, when is the union bound tight? *i.e.*, when is $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$?

Poll: when is the union bound tight?

In some other hypothetical city, the probability of rain on any given day is 30%; the probability of sun on any given day is 50%; and the probability of sun or rain on any given day is 80%. What can we say about the probability p that it will be sunny **and** rain on the same day?

A. $p = 0$
B. $p \in (0, .30]$
C. $p \in (.30, .50]$
D. $p \in (.50, .80]$
E. $p > .80$

Q: More generally, when is the union bound tight? *i.e.*, when is $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$?
A: When $A \cap B = \emptyset$

Rescuing Hoeffding: the union bound

- let's suppose $\mathcal{H}$ is finite, with m hypotheses in it
- the hypothesis g is one of those m hypotheses
- so if (given a data set $\mathcal{D}$)

  $$|E_{\text{in}}(g) - E_{\text{out}}(g)| > \varepsilon,$$

  then for some $h \in \mathcal{H}$, $|E_{\text{in}}(h) - E_{\text{out}}(h)| > \varepsilon$
- (g depends on the data set; we might choose different hs for different data sets)

Rescuing Hoeffding: the union bound

- let's suppose $\mathcal{H}$ is finite, with m hypotheses in it
- the hypothesis g is one of those m hypotheses
- so if (given a data set $\mathcal{D}$)

$$|E_{\text{in}}(g) - E_{\text{out}}(g)| > \varepsilon,$$

  then for some $h \in \mathcal{H}$, $|E_{\text{in}}(h) - E_{\text{out}}(h)| > \varepsilon$
- (g depends on the data set; we might choose different hs for different data sets)

so

$$\begin{aligned}
\mathbb{P}[|E_{\text{in}}(g) - E_{\text{out}}(g)| > \varepsilon] &\leq \sum_{h \in \mathcal{H}} \mathbb{P}[|E_{\text{in}}(h) - E_{\text{out}}(h)| > \varepsilon] \\
&\leq \sum_{h \in \mathcal{H}} 2 \exp\left(-2\varepsilon^2 n\right) \\
&= 2m \exp\left(-2\varepsilon^2 n\right)
\end{aligned}$$

Hoeffding for learning

we just proved that our learning algorithm generalizes!

Theorem (Generalization bound for learning)

Let g be a hypothesis chosen from among m different hypotheses. Then

$$\mathbb{P}[|E_{in}(g) - E_{out}(g)| > \varepsilon] \leq 2m \exp\left(-2\varepsilon^2 n\right).$$

Hoeffding for learning

we just proved that our learning algorithm generalizes!

Theorem (Generalization bound for learning)

Let g be a hypothesis chosen from among m different hypotheses. Then

$$\mathbb{P}[|E_{in}(g) - E_{out}(g)| > \varepsilon] \leq 2m \exp\left(-2\varepsilon^2 n\right).$$

Q: do you think this bound is tight?

Hoeffding for learning

we just proved that our learning algorithm generalizes!

Theorem (Generalization bound for learning)

Let g be a hypothesis chosen from among m different hypotheses. Then

$$\mathbb{P}[|E_{in}(g) - E_{out}(g)| > \varepsilon] \leq 2m \exp\left(-2\varepsilon^2 n\right).$$

Q: do you think this bound is tight?
A: no, it can overcount badly. for random events A and B, if $\mathbb{P}(A \cap B)$ is large, then $\mathbb{P}(A \cup B) \ll \mathbb{P}(A) + \mathbb{P}(B)$

Hoeffding for learning

we just proved that our learning algorithm generalizes!

Theorem (Generalization bound for learning)

Let g be a hypothesis chosen from among m different hypotheses. Then

$$\mathbb{P}[|E_{in}(g) - E_{out}(g)| > \varepsilon] \leq 2m \exp\left(-2\varepsilon^2 n\right) .$$

Q: do you think this bound is tight?
A: no, it can overcount badly. for random events A and B, if $\mathbb{P}(A \cap B)$ is large, then $\mathbb{P}(A \cup B) \ll \mathbb{P}(A) + \mathbb{P}(B)$

more information. look up the Vapnik-Chervoninkis (VC) dimension, *e.g.*, in *Learning from Data*, by Abu-Mostafa, Magdon-Ismail, and Lin.

A tradeoff for learning

- we want $\mathcal{H}$ to be **big** to make E_{in} small
- we want $\mathcal{H}$ to be **small** to ensure E_{out} is close to E_{in}

A tradeoff for learning

- we want $\mathcal{H}$ to be **big** to make E_{in} small
- we want $\mathcal{H}$ to be **small** to ensure E_{out} is close to E_{in}

what does this tell us about the difficulty of learning complicated functions f?

Generalization for regression

Theorem (Generalization bound for learning)

Let g be a hypothesis chosen from among m different hypotheses. Then

$$\mathbb{P}[|E_{in}(g) - E_{out}(g)| > \varepsilon] \leq 2m \exp\left(-2\varepsilon^2 n\right).$$

to apply Hoeffding to real-valued outputs:

- pick some small $\epsilon > 0$
- $\mathbb{1}((y_i - h(x_i))^2 \geq \varepsilon))$ is 0 if hypothesis h predicts well, 1 if hypothesis h predicts poorly
- define error of hypothesis h on data set $\mathcal{D}$ as

$$E_{\mathcal{D}}(h) = \frac{1}{|\mathcal{D}|} \sum_{(x,y)\in\mathcal{D}} \mathbb{1}((y - h(x))^2 \leq \epsilon)$$

Recap

- we introduced a probabilistic framework for generating data
- we showed that the in-sample error predicts the out-of-sample error for a single hypothesis
- we showed that the in-sample error predicts the out-of-sample error for a learned hypothesis, when $\mathcal{H}$ is finite
- we stopped there, because the math gets much more complicated — but indeed, generalization is possible!
- **the practical lesson:** (especially for complex models), don't **learn** and **estimate your error** on the same data set

in, out, train, test

- the training error does **not** obey the Hoeffding inequality
- the validation error obeys the Hoeffding inequality, with the union bound: if we choose g as the best of m models on the validation set,

$$\mathbb{P}[|E_{\text{valid}}(g) - E_{\text{out}}(g)| > \varepsilon] \leq 2m \exp\left(-2\varepsilon^2 |\mathcal{D}_{\text{valid}}|\right).$$

- the test error **does** obey the Hoeffding inequality

$$\mathbb{P}[|E_{\text{test}}(g) - E_{\text{out}}(g)| > \varepsilon] \leq 2 \exp\left(-2\varepsilon^2 |\mathcal{D}_{\text{test}}|\right).$$

so we can use the (validation error or) test error to predict generalization

Hoeffding for the validation set: details

if validation set is used for model selection, the validation error obeys
the Hoeffding inequality **with the union bound**

- for each model family, optimal model trained on $\mathcal{D}$ is a hypothesis h
- so finite number of models $m \implies$ finite hypothesis space $\mathcal{H}$
- hypotheses $h \in \mathcal{H}$ are independent of validation set $\mathcal{D}'$
- let $g_{\mathcal{D}'}$ be the hypothesis $h \in H$ with lowest error on validation set $\mathcal{D}'$
- Hoeffding with union bound applies!

$$\mathbb{P}[|E_{\mathcal{D}'}(g_{\mathcal{D}'}) - E_{\text{out}}(g_{\mathcal{D}'})| > \varepsilon] \leq 2m \exp\left(-2\varepsilon^2|\mathcal{D}'|\right).$$

References

- Concentration bounds for infinite model classes: see introduction to VC dimension in “Learning from Data” by Abu-Mostafa et al.
- Concentration bounds for cross validation: `https://arxiv.org/pdf/1706.05801.pdf`
- Concentration bounds for time series: see papers by Cosma Shalizi