ORIE 4741: Learning with Big Messy Data

Generalization

Professor Udell
Operations Research and Information Engineering
Cornell

September 10, 2019
Announcements

- midterm 10/5
- makeup exam 10/2, by arrangement with instructors (by email)
- homework due next Tuesday
- substitute next week: Sumanta Basu on decision trees!
- I won’t hold OH next week
Generalization and Overfitting

- goal of model is **not** to predict well on $D$
- 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:

<table>
<thead>
<tr>
<th>low training set error</th>
<th>low test set error</th>
<th>generalizes</th>
<th>overfits</th>
</tr>
</thead>
<tbody>
<tr>
<td>high training set error</td>
<td>high test set error</td>
<td>?!?!</td>
<td>underfits</td>
</tr>
</tbody>
</table>
Simplest case: generalizing from a mean

exit polling

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

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

▶ 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 Clinton’s expected vote share

---

1and suppose no one votes by mail or early or . . . , so these are iid samples from the US electorate
Simplest case: generalizing from a mean

exit polling

- sample $n$ voters leaving polling places\(^1\)
- for each voter $i$, define the Boolean random variable
  
  $$z_i = \begin{cases} 
  1 & \text{if voter } i \text{ voted for Clinton} \\
  0 & \text{otherwise}
  \end{cases}$$

- 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 Clinton’s expected vote share

what does the sample mean $\nu$ tell us about the true mean $\mu$?

\(^1\) and suppose no one votes by mail or early or . . . , so these are iid samples from the US electorate
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 $\epsilon > 0$,

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

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 $Ez_i = \mu$. Define the sample mean $\nu = \frac{1}{n} \sum_{i=1}^{n} z_i$. Then for any $\epsilon > 0$,

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

an example of a concentration inequality

▶ $\mu$ can’t be much higher than $\nu$
▶ $\mu$ can’t be much lower than $\nu$
▶ more samples $n$ improve estimate exponentially quickly
Back to the learning problem

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

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

$$= \mathbb{1}(y_i = h(x_i))$$
Back to the learning problem

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

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

**example.** build a model of voting behavior:

- $y_i = f(x_i)$ is 1 if voter $i$ voted for Clinton, 0 otherwise
- $h(x_i)$ is our guess of how the voter will vote, using hypothesis $h$
- $z_i$ 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 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 $x$:

\[
P(y | x) = \begin{cases} 1 & y = f(x) \\ 0 & \text{otherwise} \end{cases}
\]
Adding in probability

make our model probabilistic:

- fix a probability distribution $P(x, y)$
- sample $(x_i, y_i)$ iid 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 $x$:

$$P(y|x) = \begin{cases} 1 & y = f(x) \\ 0 & \text{otherwise} \end{cases}$$
fix a hypothesis $h : \mathcal{X} \rightarrow \mathcal{Y}$.

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 $z_i = 1(y_i = h(x_i))$

$z_i$ are iid (since $(x_i, y_i)$ are iid, and $h$ is fixed)

$Ez = E_{(x,y) \sim P(x,y)} 1(y = h(x))$

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

$$\mathbb{P}\left[ \left| \frac{1}{n} \sum_{i=1}^{n} z_i - Ez \right| > \epsilon \right] \leq 2 \exp \left( -2\epsilon^2 n \right)$$
Hoeffding for the noisy learning problem

- fix a hypothesis $h : \mathcal{X} \rightarrow \mathcal{Y}$.
- 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 $z_i = 1(y_i = h(x_i))$
- $z_i$ are iid (since $(x_i, y_i)$ are iid, and $h$ is fixed)
- $\mathbb{E} z = \mathbb{E}_{(x,y) \sim P(x,y)} 1(y = h(x))$

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

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

Q: what is the probability over?
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 $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$
- take $z_i = 1(y_i = h(x_i))$
- $z_i$ are iid (since $(x_i, y_i)$ are iid, and $h$ is fixed)
- $\mathbb{E}z = \mathbb{E}_{(x,y)\sim P(x,y)}1(y = h(x))$

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

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

Q: what is the probability over?
A: other data sets $D' = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ 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 \( z_i = 1(y_i = h(x_i)) \).
- \( z_i \) are iid (since \((x_i, y_i)\) are iid, and \( h \) is fixed).
- \( \mathbb{E}z = \mathbb{E}_{(x,y)\sim P(x,y)} 1(y = h(x)) \)

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

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

**Q:** what is the probability over?
**A:** other data sets \( \mathcal{D}' = \{(x_1, y_1), \ldots, (x_n, y_n)\} \) 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 $z_i = 1(y_i = h(x_i))$
- $z_i$ are iid (since $(x_i, y_i)$ are iid, and $h$ is fixed)
- $\mathbb{E}z = \mathbb{E}_{(x,y) \sim P(x,y)} 1(y = h(x))$

so we can apply Hoeffding! for any $\epsilon > 0$,\[
\mathbb{P}\left[ \left| \frac{1}{n} \sum_{i=1}^{n} z_i - \mathbb{E}z \right| > \epsilon \right] \leq 2 \exp \left( -2\epsilon^2 n \right)
\]

**Q:** what is the probability over?
**A:** other data sets $\mathcal{D}' = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ 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:** test set error, since $h$ is independent of $\mathcal{D}$
some new terminology:

▶ **in-sample error.**

\[
E_{\text{in}}(h) = \text{fraction of } D \text{ where } y_i \text{ and } h(x_i) \text{ disagree} \\
= \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(y_i \neq h(x_i))
\]

▶ **out-of-sample error.**

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

notice

\[
E_{\text{out}}(h) = \mathbb{E}[E_{\text{in}}(h)]
\]
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 $z_i = \mathbb{1}(y_i = h(x_i))$
- $z_i$ are iid (since $(x_i, y_i)$ are iid, and $h$ is fixed)
- $\mathbb{E}z = \mathbb{E}_{(x,y) \sim P(x,y)} \mathbb{1}(y = h(x))$

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

$$\mathbb{P}[|E_{in}(h) - E_{out}(h)| > \varepsilon] \leq 2 \exp\left(-2\varepsilon^2 n\right)$$
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 Clinton. Then sample data $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, and set $z_i = 1(y_i = h(x_i))$.

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 Clinton (e.g., using perceptron). Set $z_i' = 1(y_i = g(x_i))$.

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 $h: \mathcal{X} \rightarrow \mathcal{Y}$ to predict who will vote for Clinton. Then sample data $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, and set $z_i = 1(y_i = h(x_i))$.

2. Sample the data $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, and use it to develop a model $g: \mathcal{X} \rightarrow \mathcal{Y}$ to predict who will vote for Clinton (e.g., using perceptron). Set $z'_i = 1(y_i = g(x_i))$.

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 $z_i$s iid? What about the $z'_i$s?
Does Hoeffding work for our learned model?

two scenarios:

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

2. Sample the data \( \mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\} \), and use it to develop a model \( g : \mathcal{X} \rightarrow \mathcal{Y} \) to predict who will vote for Clinton (e.g., using perceptron). Set \( z'_i = 1(y_i = g(x_i)) \).

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 \( \mathbb{E}z' \)?

\( \textbf{Q:} \) Are the \( z_i \)s iid? What about the \( z'_i \)s?

\( \textbf{A:} \) \( z_i \)s are iid. \( z'_i \)s are not independent: they depend on \( g \), which depends on the whole data set \( \mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\} \).
Does Hoeffding work for our learned model?

two scenarios:

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

2. Sample the data \( \mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\} \), and use it to develop a model \( g : \mathcal{X} \rightarrow \mathcal{Y} \) to predict who will vote for Clinton (e.g., using perceptron). Set \( z'_i = 1(y_i = g(x_i)) \).

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 \( z \)'s iid? What about the \( z'_i \)'s?

**A:** \( z \)'s 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?
Does Hoeffding work for our learned model?

two scenarios:

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

2. Sample the data \( \mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\} \), and use it to develop a model \( g : \mathcal{X} \rightarrow \mathcal{Y} \) to predict who will vote for Clinton (e.g., using perceptron). Set \( z'_i = 1(y_i = g(x_i)) \).

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 \( z_i \)'s iid? What about the \( z'_i \)'s?

**A:** \( z_i \)'s 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 $D_{\text{train}}$ and test set $D_{\text{test}}$
▶ pick $m$ different interesting model classes
e.g., different $\phi$s: $\phi_1$, $\phi_2$, $\ldots$, $\phi_m$
▶ fit (“train”) models on training set $D_{\text{train}}$
get one model $h : \mathcal{X} \rightarrow \mathcal{Y}$ for each $\phi$s, and set

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

▶ compute error of each model on test set $D_{\text{test}}$ and choose lowest:

$$g = \arg\min_{h \in \mathcal{H}} E_{D_{\text{test}}} (h)$$

$g$ was trained on $D_{\text{test}}$!

Hoeffding does not directly apply:

$E_{D_{\text{test}}}(g)$ may not accurately estimate $E_{\text{out}}(g)$.
Recall validation procedure

how to decide which model to use?

▶ split data into training set $\mathcal{D}_{\text{train}}$ and test set $\mathcal{D}_{\text{test}}$
▶ pick $m$ different interesting model classes
e.g., different $\phi$s: $\phi_1$, $\phi_2$, $\ldots$, $\phi_m$
▶ fit (“train”) models on training set $\mathcal{D}_{\text{train}}$
get one model $h : \mathcal{X} \to \mathcal{Y}$ for each $\phi$s, and set

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

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

$$g = \arg\min_{h \in \mathcal{H}} E_{\mathcal{D}_{\text{test}}}(h)$$

**Q:** Are $\{z_i = 1(y_i = g(x_i) : (x_i, y_i) \in \mathcal{D}_{\text{test}})\}$ independent?
Recall validation procedure

how to decide which model to use?

- split data into training set $\mathcal{D}_{\text{train}}$ and test set $\mathcal{D}_{\text{test}}$
- pick $m$ different interesting model classes
e.g., different $\phi$s: $\phi_1, \phi_2, \ldots, \phi_m$
- fit (“train”) models on training set $\mathcal{D}_{\text{train}}$
  get one model $h: \mathcal{X} \rightarrow \mathcal{Y}$ for each $\phi$s, and set
  \[ \mathcal{H} = \{ h_1, h_2, \ldots, h_m \} \]
- compute error of each model on test set $\mathcal{D}_{\text{test}}$ and choose lowest:
  \[ g = \arg\min_{h \in \mathcal{H}} E_{\mathcal{D}_{\text{test}}}(h) \]

Q: Are $\{ z_i = 1(y_i = g(x_i) : (x_i, y_i) \in \mathcal{D}_{\text{test}}) \}$ independent?
A: No; $g$ was trained on $\mathcal{D}_{\text{test}}$!

**Hoeffding does not directly apply:**
$E_{\mathcal{D}_{\text{test}}}(g)$ may not accurately estimate $E_{\text{out}}(g)$
The union bound

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

$$\Pr(A \cup B) \leq \Pr(A) + \Pr(B)$$
The union bound

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

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

**Q:** When is this bound tight? *i.e.*, when is

$$
\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)?
$$
The union bound

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

\[
P(A \cup B) \leq P(A) + P(B)
\]

Q: When is this bound tight? i.e., when is
\[
P(A \cup B) = P(A) + P(B)
\]
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 $h$s 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 $h$s for different data sets)

So

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

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

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

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

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

**Q:** do you think this bound is tight?

**A:** no, it can overcount badly. for random events $A$ and $B$, if $\Pr(A \cap B)$ is large, then $\Pr(A \cup B) \ll \Pr(A) + \Pr(B)$

**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

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

to apply Hoeffding to real-valued outputs:

- pick some small \( \varepsilon > 0 \)
- \( 1((y_i - h(x_i))^2 \leq \varepsilon) \) is 0 if hypothesis \( h \) predicts well, 1 if hypothesis \( h \) predicts poorly
- define error of hypothesis \( h \) on data set \( D \) as

\[
E_D(h) = \frac{1}{|D|} \sum_{(x,y)\in D} 1((y - h(x))^2 \leq \varepsilon)
\]
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
The in-sample error and training error do not obey the Hoeffding inequality (without using, e.g., a union bound).

If test set is only used for testing (not for model selection), the test error does obey the Hoeffding inequality:

$$\Pr[|E_{\text{test}}(g) - E_{\text{out}}(g)| > \varepsilon] \leq 2 \exp\left(-2\varepsilon^2|D_{\text{test}}|\right).$$

So we can use the test error to predict generalization.
Overfitting to the test set

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

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

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

▶ Concentration bounds for infinite model classes: see introduction to VC dimension in “Learning from Data” by Abu-Mostafa et al.


▶ Concentration bounds for time series: see papers by Cosma Shalizi