Model Behavior
How to evaluate machine learning models
Recap: Model Fitting

**Fitting:** Observe $x_i$, $y_i$, infer $\theta$

Regression Problems $\rightarrow$ Continuous $Y$

Classification Problems $\rightarrow$ Discrete $Y$
Recap: Model Fitting - Machine Learning

**Fitting:** Observe $x_i$, $y_i$, **infer** $\theta$

Regression Problems -> Continuous $Y$

Classification Problems -> Discrete $Y$
Convolutional Neural Network

Cascade of filters that can be trained (fit) as one unit

“Layers” that consist of convolutions, classifications, regressions
Natural Language Understanding

Word2vec

king

man

woman
All Models: Training and Testing

Original Data → Randomly Selected Rows → Training Data → Testing Data

Use this data for modeling
Use this data for evaluation
Why?

Unbiased sample of unseen data for evaluation
Recap: Back to the beginning

Error in an estimate is normally distributed

$$\varepsilon \approx N(0, \frac{\sigma^2_{pop}}{K})$$

Variance of the population

Size of the sample
Recap: Back to the beginning

Hypothetical Infinite Dataset

Test Set

“Sample Error”

“Expected Population Error”
Recap: Back to the beginning

Hypothetical Infinite Dataset

Test Set

“Empirical Risk”

“Risk”
When Can A Test Set Be Misleading?

Test performance deviates from actual performance
Recap: Response Bias

<table>
<thead>
<tr>
<th>Sampling Probability</th>
<th>Response Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRS 1</td>
<td>1/7<em>1 1/7</em>1 1/7*1</td>
</tr>
<tr>
<td>SRS 2</td>
<td>1/7<em>0.2 1/7</em>0.2 1/7*0.2</td>
</tr>
</tbody>
</table>

Note they don’t have the same sampling rate!
Covariate (Dataset) Shift

Sample Error Rate: 2/15

Overall Error Rate: 10/33
Why?
Why?

- **sudden**
  - Blue circles: gradual change
  - Red triangles: sudden change

- **gradual**
  - Blue circles: gradual change
  - Red triangles: gradual change
Systematically Excluded Training Data

First Self-Driving Car Called ALVINN
Systematically Excluded Training Data

Problem: Only trained on example of “good” driving

Could never recover from errors
Why?
Gradual Shift in Dataset

Facebook happened “before” smart phones
When Can A Test Set Be Misleading?

Test performance deviates from actual performance
Systematically Corrupted Data
Systematically Corrupted Data

Societal Biases in Training Datasets

REPORT
Reducing bias in AI-based financial services
Aaron Klein - Friday, July 10, 2020
Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016
When Can A Test Set Be Misleading?

Test performance deviates from actual performance
Why and How Baidu Cheated an Artificial Intelligence Test

Machine learning gets its first cheating scandal.
All Models: Training and Testing

Original Data → Randomly Selected Rows → Training Data

Use this data for modeling

Testing Data

Use this data for evaluation
All Models: Training and Testing

Each new model needs to draw a new test set!

Model 1 or Model 2

Use this data for evaluation
All Models: Training and Testing

Each new model needs to draw a new test set!

Original Data \rightarrow Randomly Selected Rows \rightarrow Training Data

Training Data \rightarrow Testing Data

Model 1 or Model 2

Use this data for evaluation

Information leaked
Possible Solution

Keep untouched test set and split training data
**Possible Solution**

What if the training data is too small?

Average over different “validation” sets