Where does the error come from?
A more complex model does not always lead to better performance on testing data.
\[ \hat{y} = \hat{f}(x) \]

Only Niantic knows \( \hat{f} \)

From training data, we find \( f^* \)

\( f^* \) is an estimator of \( \hat{f} \)
Bias and Variance of Estimator

• Estimate the mean of a variable $x$
  • assume the mean of $x$ is $\mu$
  • assume the variance of $x$ is $\sigma^2$

• Estimator of mean $\mu$
  • Sample $N$ points: $\{x^1, x^2, ..., x^N\}$

$$m = \frac{1}{N} \sum_n x^n \neq \mu$$

$$E[m] = E \left[ \frac{1}{N} \sum_n x^n \right] = \frac{1}{N} \sum_n E[x^n] = \mu$$
Bias and Variance of Estimator

• Estimate the mean of a variable $x$
  • assume the mean of $x$ is $\mu$
  • assume the variance of $x$ is $\sigma^2$
• Estimator of mean $\mu$
  • Sample $N$ points: $\{x^1, x^2, \ldots, x^N\}$

$$m = \frac{1}{N} \sum_{n} x^n \neq \mu$$

$$Var[m] = \frac{\sigma^2}{N}$$

Variance depends on the number of samples
Bias and Variance of Estimator

- Estimate the mean of a variable $x$
  - assume the mean of $x$ is $\mu$
  - assume the variance of $x$ is $\sigma^2$
- Estimator of variance $\sigma^2$
  - Sample $N$ points: $\{x^1, x^2, \ldots, x^N\}$

$$m = \frac{1}{N} \sum_n x^n \quad s^2 = \frac{1}{N} \sum_n (x^n - m)^2$$

Biased estimator

$$E[s^2] = \frac{N - 1}{N} \sigma^2 \neq \sigma^2$$
$$E[f^*] = \bar{f}$$

- **Low Bias**
  - Low Variance
  - Low Bias

- **High Bias**
  - High Variance
  - High Bias

- **Variance**
- **Bias**

- **$f^*$**
- **$\hat{f}$**
Parallel Universes

• In all the universes, we are collecting (catching) 10 Pokémon as training data to find $f^*$
Parallel Universes

• In different universes, we use the same model, but obtain different $f^*$

$$y = b + w \cdot x_{cp}$$

Universe 123

Universe 345
\( f^* \) in 100 Universes

\[ y = b + w \cdot x_{cp} \]

\[ y = b + w_1 \cdot x_{cp} + w_2 \cdot (x_{cp})^2 + w_3 \cdot (x_{cp})^3 + w_4 \cdot (x_{cp})^4 + w_5 \cdot (x_{cp})^5 \]
Variance

\[ y = b + w \cdot x_{cp} \]

\[ y = b + w_1 \cdot x_{cp} + w_2 \cdot (x_{cp})^2 + w_3 \cdot (x_{cp})^3 + w_4 \cdot (x_{cp})^4 + w_5 \cdot (x_{cp})^5 \]

Small Variance

Large Variance

Simpler model is less influenced by the sampled data

Consider the extreme case \( f(x) = c \)
Bias

\[ E[f^*] = \bar{f} \]

- Bias: If we average all the \( f^* \), is it close to \( \hat{f} \)

![Diagram showing large and small bias with a graph assuming this is \( \hat{f} \)]
Black curve: the true function $\hat{f}$
Red curves: 5000 $f^*$
Blue curve: the average of 5000 $f^*$

$$= \bar{f}$$
Bias

\[ y = b + w \cdot x_{cp} \]

\[ y = b + w_1 \cdot x_{cp} + w_2 \cdot (x_{cp})^2 + w_3 \cdot (x_{cp})^3 + w_4 \cdot (x_{cp})^4 + w_5 \cdot (x_{cp})^5 \]
Bias v.s. Variance

- Large Bias
- Small Bias
- Large Variance
- Small Variance

Error from bias
Error from variance
Error observed

Underfitting
Overfitting

Large Bias → Small Bias
Small Variance → Large Variance
What to do with large bias?

• Diagnosis:
  • If your model cannot even fit the training examples, then you have large bias (Underfitting).
  • If you can fit the training data, but large error on testing data, then you probably have large variance (Overfitting).

• For bias, redesign your model:
  • Add more features as input
  • A more complex model
What to do with large variance?

• More data
  Very effective, but not always practical

• Regularization
Model Selection

• There is usually a trade-off between bias and variance.
• Select a model that balances two kinds of error to minimize total error
• What you should NOT do:

```
Training Set
Model 1  Err = 0.9
Model 2  Err = 0.7
Model 3  Err = 0.5
```

```
Testing Set (real)
Err > 0.5
```

```
Testing Set
Err > 0.5
```
Homework

Training Set

Model 1 \[\Rightarrow \text{Err} = 0.9\]
Model 2 \[\Rightarrow \text{Err} = 0.7\]
Model 3 \[\Rightarrow \text{Err} = 0.5\]

Testing Set

public

private

Err > 0.5

I beat baseline!

No, you don’t

What will happen next Friday?

http://www.chioka.in/how-to-select-your-final-models-in-a-kaggle-competition/
Cross Validation

Using the results of public testing data to tune your model
You are making public set better than private set.

Not recommend
N-fold Cross Validation

Training Set

Train
Train
Val

Train
Val
Train

Val
Train
Train

Model 1
Err = 0.2
Err = 0.4
Err = 0.3

Model 2
Err = 0.4
Err = 0.5
Err = 0.5

Model 3
Err = 0.4
Err = 0.6
Err = 0.3

Avg Err
= 0.3
Avg Err
= 0.5
Avg Err
= 0.4

Testing Set

public

Testing Set

private
Reference

- Bishop: Chapter 3.2