7: SAMPLING, CROSS VALIDATION, BOOTSTRAPPING, BIAS VARIANCE, OVER/UNDER FITTING

LEARNING OBJECTIVES

- Bias - Variance decomposition
- Cross validation
- Sampling, Boostrapping
- Strategies to deal with Overfitting and Underfitting

REVIEW OF LESSON 6

LAST LESSON REVIEW

- Scikit: model, fit(), predict()
- Linear regression
- Loss function
- Ridge, Lasso, Elastic
- Polynomial regression

LAST SESSION

ANY QUESTIONS FROM LAST CLASS?

- What's a loss function?
- What is Ridge?
  - What's the impact of α?
- What is Polynomial regression?

AND

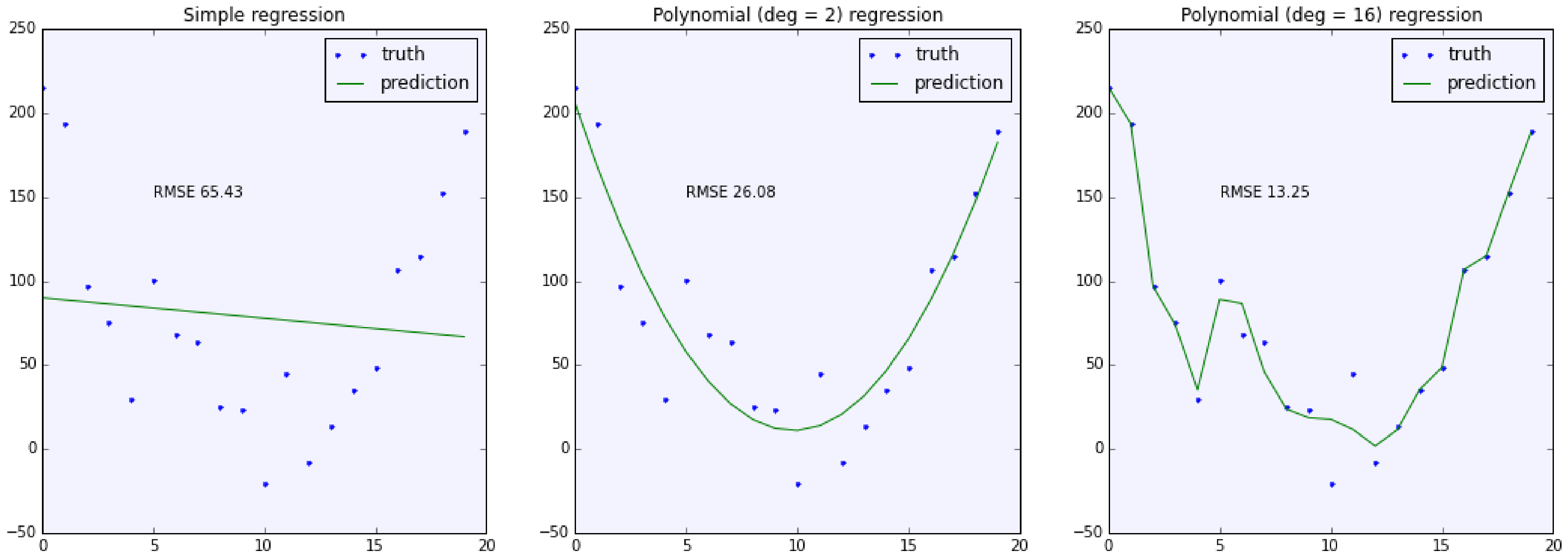
- Matlab vs Scikit / Matplotlib
- Time spent on viz vs analysis
- In depth understanding of the models needed? Where is the info?
- Is there a prefered library for a given model such as OLS or do you change libraries?

TODAY

MAKING THE MOST OUT OF YOUR DATA AND MODEL SELECTION

POLYNOMIAL REGRESSION

if you remember we had the following plots for polynomial regression:



The higher the degree of the polynomial, the better the fit

How does it work with new data?

OVER FITTING / UNDER FITTING

- Exercise: on housing dataset, with linear and polynomial (N = 4, 16)
- Split the dataset into 2 chunks. train set of 100 samples. the rest for the test set
- Train 3 models on the train set with X= df.NOX and y = df.NIX
- Estimate the prediction error on the test set with these 3 models
- What's hapenning here?

OVER FITTING / UNDER FITTING

Notebook 1: Overfitting or Underfitting

- Housing dataset on df.NOX ~ df.DIS
- with seaborn, sns.jointplot(df.NOX, df.DIS)
- 3 different polynomial regression models: degree = 1 (linear), 4 and 16

```
poly = PolynomialFeatures(4)
X4 = poly.fit_transform(X)
```

- split the datset in 2: train (first 100 rows) and test (the rest)

```
df = df[:100]
df_test = df[100:]
```

- Plot the prediction (NOX) vs the predictors (DIS, DIS^4 and DIS ^16)
- Calculate the MSE on the training set and test set for the different models
- Although the MSE on training improves, the MSE on test gets worse

BIAS VARIANCE DECOMPOSITION

The prediction error of your model can be decomposed in 2 terms:

- The bias
- The Variance

Total Error = Bias Error + Variance Error

BIAS VARIANCE DECOMPOSITION

Error due to Bias: Error due to bias is taken as the difference between the average prediction of our model and the correct value which we are trying to predict.

Bias measures how far off in general your models' predictions are from the correct value.

Error due to Variance: The error due to variance is taken as the variability of a model prediction for a given data point.

The variance is how much the predictions for a given point vary between different realizations of the model.

BIAS VARIANCE DECOMPOSITION

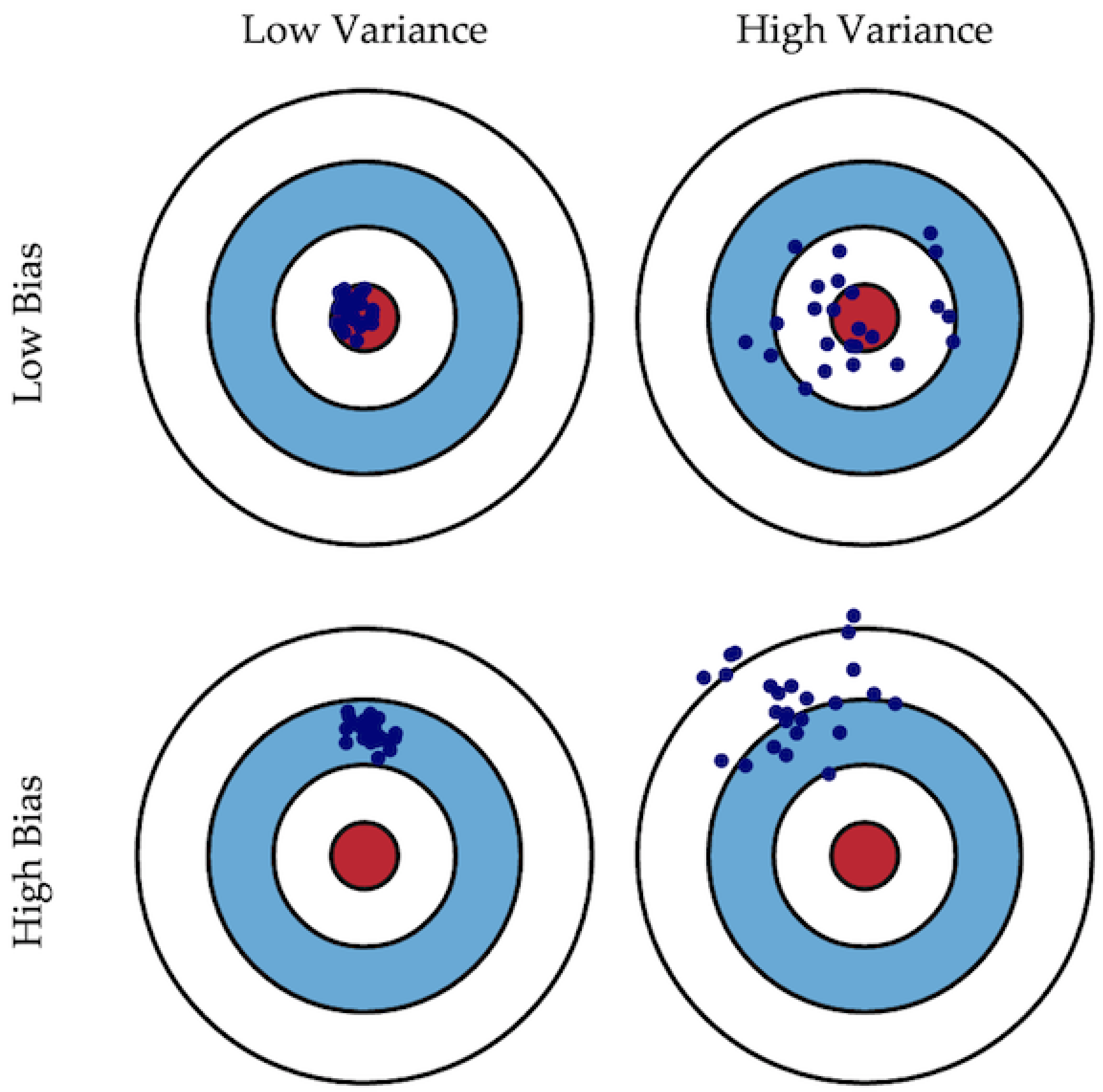


Fig. 1 Graphical illustration of bias and variance.

BIAS VARIANCE DECOMPOSITION

$$\text{MSE}(\hat{Y}) = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$$

Which can be rewritten as

$$\text{MSE}(\hat{Y}) = \mathbb{E}\left[(\hat{Y} - Y)^2\right] = \mathbb{E}\left[\left(\hat{Y} - \mathbb{E}(\hat{Y})\right)^2\right] + \left(\mathbb{E}(\hat{Y}) - Y\right)^2$$

$$\text{MSE}(\hat{Y}) = \text{Var}(\hat{Y}) + \text{Bias}(\hat{Y}, Y)^2$$

with

- $$\text{Var}(\hat{Y}) = \mathbb{E}\left[\left(\hat{Y} - \mathbb{E}(\hat{Y})\right)^2\right]$$
- $\text{Bias}(\hat{Y}, Y) = \mathbb{E}(\hat{Y}) - Y$

$$
\begin{aligned}
\mathbb{E}((\hat{Y}-Y)^2) &= \mathbb{E}\left[\left(\hat{Y}-\mathbb{E}(\hat{Y})+\mathbb{E}(\hat{Y})-Y\right)^2\right] \\
&= \mathbb{E}\left[\left(\hat{Y}-\mathbb{E}(\hat{Y})\right)^2 + 2\left((\hat{Y}-\mathbb{E}(\hat{Y}))(\mathbb{E}(\hat{Y})-Y)\right) + \left(\mathbb{E}(\hat{Y})-Y\right)^2\right] \\
&= \mathbb{E}\left[\left(\hat{Y}-\mathbb{E}(\hat{Y})\right)^2\right] + 2\mathbb{E}\left[(\hat{Y}-\mathbb{E}(\hat{Y}))\overbrace{(\mathbb{E}(\hat{Y})-Y)}^{\text{This is a constant, so it can be pulled out.}}\right] + \mathbb{E}\Big[\overbrace{\Big(\mathbb{E}(\hat{Y})-Y}^{\text{This is a constant, so i expected valu is itself.}} \\
&= \mathbb{E}\left[\left(\hat{Y}-\mathbb{E}(\hat{Y})\right)^2\right] + 2\underbrace{\mathbb{E}(\hat{Y}-\mathbb{E}(\hat{Y}))}_{=\mathbb{E}(\hat{Y})-\mathbb{E}(\hat{Y})=0}(\mathbb{E}(\hat{Y})-Y) + \left(\mathbb{E}(\hat{Y})-Y\right)^2 \\
&= \mathbb{E}\left[\left(\hat{Y}-\mathbb{E}(\hat{Y})\right)^2\right] + \left(\mathbb{E}(\hat{Y})-Y\right)^2 \\
&= \mathrm{Var}(\hat{Y}) + \mathrm{Bias}(\hat{Y},Y)^2
\end{aligned}
$$

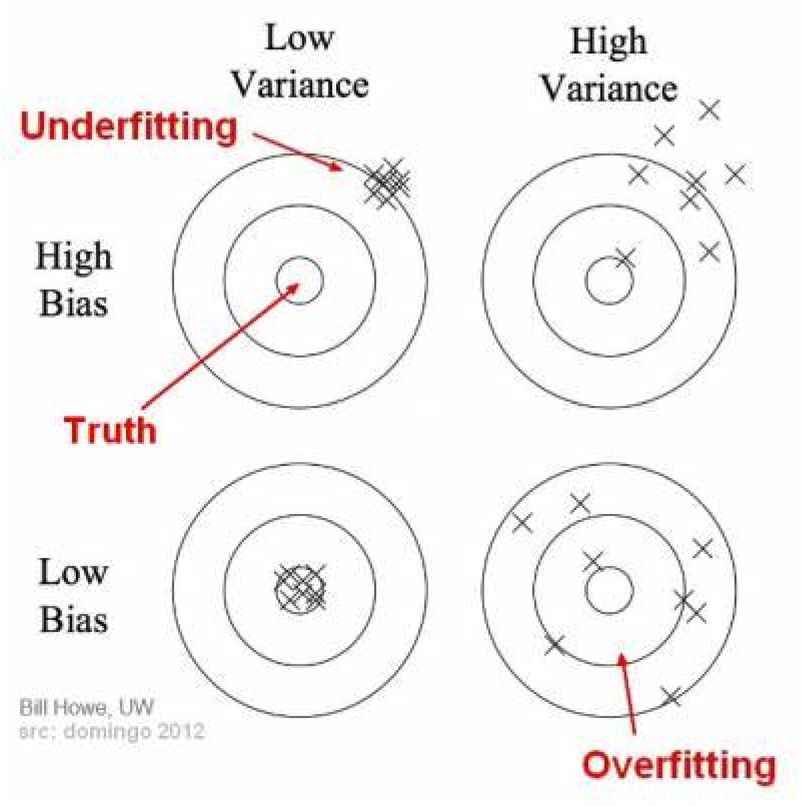
RECAP

In order to minimize the Mean Squared Error, we have to reduce both the **Bias** and the **Variance**

- High Bias <-> Underfitting
  - Not a good estimator
- High Variance <-> Overfitting
  - too **sensitive** to training data
  - no predictive power

RECAP

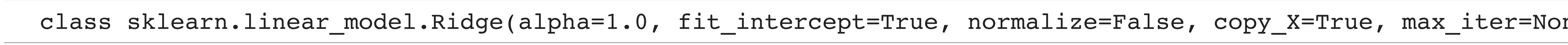
On data that has not been seen before by the model:



IS YOUR MODEL GOOD?
IS IT PREDICTIVE?

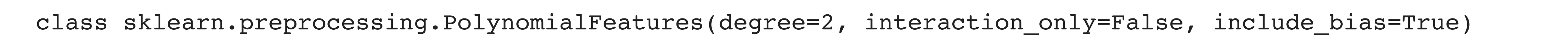
META PARAMETERS - TUNING YOUR MODEL

- α in Ridge



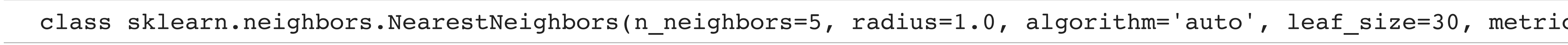
```
class sklearn.linear_model.Ridge(alpha=1.0, fit_intercept=True, normalize=False, copy_X=True, max_iter=Nor
```

- Degree in polynomial regression



```
class sklearn.preprocessing.PolynomialFeatures(degree=2, interaction_only=False, include_bias=True)
```

- K in K-NN



```
class sklearn.neighbors.NearestNeighbors(n_neighbors=5, radius=1.0, algorithm='auto', leaf_size=30, metric
```

Meta parameters allow you to train several different models on a given set of data

TRAIN, VALIDATION, TEST

Split your data in 3 subsets: for instance: 60%, 20%, 20%

1. **Training set** : This *training* dataset is used to *train* different prediction models with different hyper parameters.
2. **Validation** set: Since this data has not yet been seen by the model we can use that validation set to assess the prediction power of our models.

Model selection: select the model with the best performance.

1. **Test set**: Reality check. how is our *best* model performing on totally new data? You obtain an idea of the real predictive power of your chosen model

Do not use the results obtained on your test data to further optimize your model.

TRAIN, VALIDATION, TEST

EXERCISE: MODEL SELECTION ON HOUSING DATA

- Split your data in 3 subsets: train validation and test
- Train 10 Ridge models with 10 different α
- Calculate MSE on the validation set for these 10 models
- Pick the best one
- Run the best one on the test set and calculate the resulting MSE
- **Do not touch your model anymore**

L7 N2 - train, validation test split and model selection

K-FOLD CROSS VALIDATION

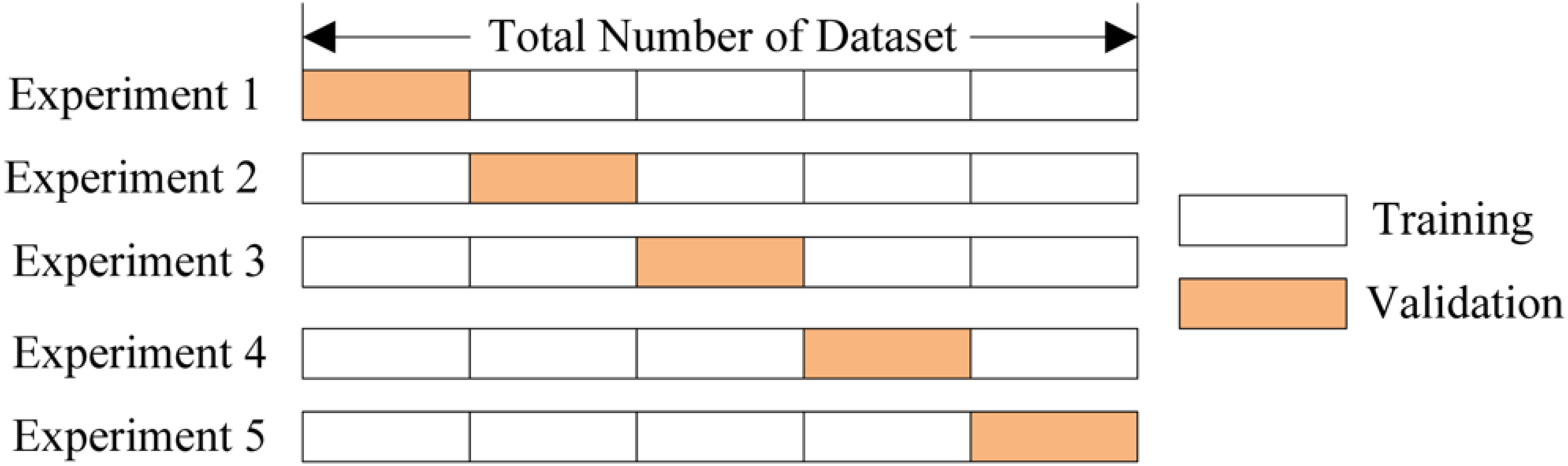
Train, valid, test: You're wasting a lot of data

K-FOLD CROSS VALIDATION

1. Split your data into train (80%) / test (20%), leave the test alone
2. Then further split your training set on K subsets (K = 4)
    - train on 1,2,3, validate on 4
    - train on 1,2,4, validate on 3
    - train on 1,3,4, validate on 2
    - train on 2,3,4, validate on 1

The average of the errors obtained on the validations subsets is a better estimation on the performance of your model than if you had just one validation set.

K-FOLD CROSS VALIDATION



CROSS VALIDATION

Many other ways to do cross validation in Scikit learn.

- Stratified K-Fold: preserving the percentage of samples for each class.
- Leave one out: Each sample is used once as a test set (singleton) while the remaining samples form the training set.
- Shuffle Split: Random permutation cross-validation iterator. Yields indices.
- cross_val_score, see this example

K-FOLD CROSS VALIDATION

Exercise

On the diabetes dataset, find the optimal regularization parameter alpha.

Bonus: How much can you trust the selection of alpha?

http://scikit-learn.org/stable/modules/generated/sklearn.cross_validation.KFold.html#sklearn.cross_v

http://scikit-learn.org/stable/auto_examples/exercises/plot_cv_diabetes.html#example-exercises-plot-cv-diabetes-py

DON'T LET SMALL DATASETS BRING YOU DOWN!

BOOTSTRAPING

NOW YOU CAN GET EVEN MORE DATA OUT OF YOUR DATA WITH …

BOOTSTRAPING

BOOTSTRAPPING

BOOTSTRAPPING

BOOTSTRAPPING

BOOTSTRAPPING

BOOTSTRAPPING

You have a small number of samples (N samples). You want a good understanding of some statictics / measure of the overall population: for instance the mean, or the Standard Deviation

1. You calculate the mean on your whole sample

Gives you one number but not much in terms of the reliability of this number.

1. You run N experiments
   - Each time you select half of your samples with replacement Which means some samples are counted selevarl times
   - You calculate the mean for each experiment You end up with a pretty solid estimation of the distribution of the mean of your population.

Bootstrapping animated

BOOTSTRAPPING

Notebook - Boostrapping for the mean

BOOTSTRAPPING FOR MODEL SELECTION

Using Scikit's Bootstrap or resample

Exercise: estimate the distribution of your linear regression coeffficients

IS YOUR MODEL OVERFITTING OR UNDER FITTING

LEARNING CURVE

- set aside a test set
- train your model on increasing number of training samples
- for each model, calculate the training error and the test error

LEARNING CURVE IN SCIKIT

scikit example 1

scikit example 2

LEARNING CURVE

How you know if your model is over fitting or under fitting?

The learning curve traces the error on training and the error on the validation set as the sample sizes increases.

For small sample sizes, the model does not have enough data to learn and both the training error and the testing error are high.

As the sample size increases the learning rate decreases. The model is learning the data. And the testing error is also decreasing. The model becomes a better predictor.

As you increase the sample size, there comes a point where the training error keeps on decreasing but the testing error stops decreasing and may instead starts increasing.

LEARNING CURVE - ANDREW NG

https://www.coursera.org/learn/machine-learning/lecture/Kont7/learning-curves

LEARNING CURVE - LINKS

Good illustrations of learning curves

- http://www.ultravioletanalytics.com/2014/12/12/kaggle-titanic-competition-part-ix-bias-variance-and-learning-curves/
- http://www.astroml.org/sklearn_tutorial/practical.html

OVER FITTING OR UNDER FITTING?

Exercise: polynomial with n = 3 on housing dataset

Plot the learning curve

```
estimator = ...
X_train, y_train, X_test, y_test = split(X, y)
n_samples = X_train.shape[0]
train_scores, test_scores = [], []
for n in range(10, 10, n_samples):
    estimator.fit(X_train[:n], y_train[n])
    train_scores.append(estimator.score(X_train[:n], y_train[n]))
    test_scores.append(estimator.score(X_test, y_test))
plot(range(10, 10, n_samples), train_scores)
plot(range(10, 10, n_samples), test_scores)
```

STRATEGIES

HIGH BIAS - UNDER FITTING

http://www.astroml.org/sklearn_tutorial/practical.html

- Add more features. In our example of predicting home prices, it may be helpful to make use of information such as the neighborhood the house is in, the year the house was built, the size of the lot, etc. Adding these features to the training and test sets can improve a high-bias estimator

- Use a more sophisticated model. Adding complexity to the model can help improve on bias. For a polynomial fit, this can be accomplished by increasing the degree d. Each learning technique has its own methods of adding complexity.

- Decrease regularization. Regularization is a technique used to impose simplicity in some machine learning models, by adding a penalty term that depends on the characteristics of the parameters. If a model has high bias, decreasing the effect of regularization can lead to better results.

HIGH VARIANCE - OVER FITTING

- Use fewer features. Using a feature selection technique may be useful, and decrease the over-fitting of the estimator.
- Use more training samples. Adding training samples can reduce the effect of over-fitting, and lead to improvements in a high variance estimator.
- Increase Regularization. Regularization is designed to prevent over-fitting. In a high-variance model, increasing regularization can lead to better results.
- Bagging to reduce over fitting (average several overfitting models)
  https://class.coursera.org/datasci-001/lecture/157
  https://en.wikipedia.org/wiki/Bootstrap_aggregating

LESSON REVIEW

LESSON REVIEW

- Bias - Variance decomposition
- Cross validation
- Sampling, Boostrapping
- Strategies to deal with Overfitting and Underfitting

BEFORE NEXT CLASS

5 QUESTIONS ABOUT TODAY

EXIT TICKET