NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

SC1015 : Review Lecture

Developing Data Science Projects

Dr Sourav SEN GUPTA

SC1015

Admin Announcements

35% of Continuous Assessments Completed

- 5% : LAMS Quizzes for DS. None for AI.
- 15% : Graded Lab Exercises 3, 4, 5 for DS.
- 15% : DS Theory Quiz during Recess Week.

65% of Continuous Assessments Pending

- 10% : Graded Lab Exercises 7, 8 for AI.
- 25% : AI Theory Quiz on 18 April, Monday.
- 30% : Mini-Project (DS). Deadline 24 April.

Study material on Artificial Intelligence released.

SC1015

Mini-Project Grading Rubrics

30% for “Technical” Components

- 10% : Problem definition based on a dataset.
- 20% : Use Machine Learning tools to solve it.

40% for “Analytical” Components

- 20% : Exploratory Analysis and Visualization.
- 20% : Data-driven insights from your solution.

20% for “Supporting” Components

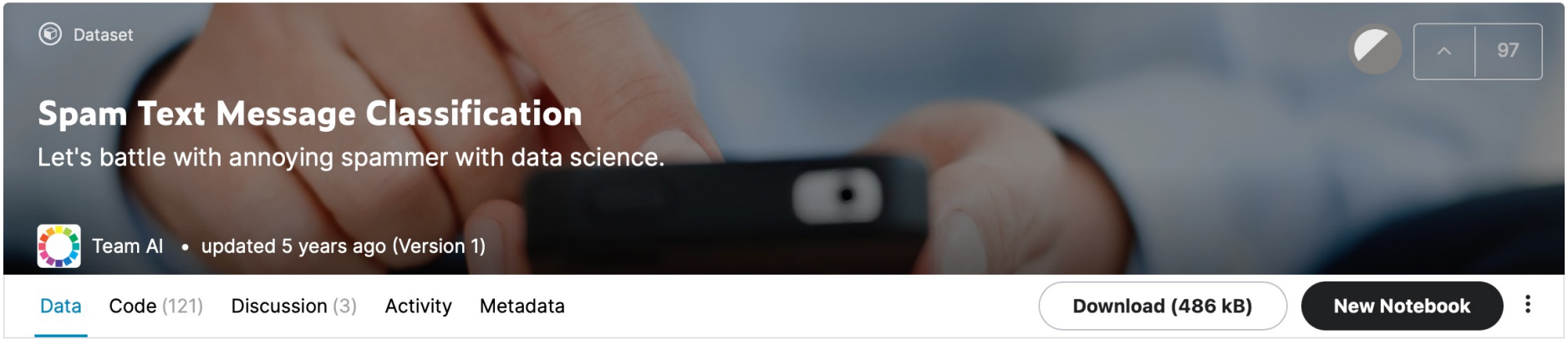
- 10% : Data extraction, preparation, cleaning.
- 10% : Presentation and overall impression.

10% for “Learning” something new (fits anywhere).

Here are a few pointers on the …

“TECHNICAL” COMPONENTS

Primarily inspired by Xavier Bresson’s guide to developing data science projects : https://github.com/xbresson/CE9010_2018 (check Lecture 7 for details)



Step 1 : Choose a dataset according to your interests.

Example : Spam Text Message Classification dataset from Kaggle
https://www.kaggle.com/team-ai/spam-text-message-classification

Your main decision at this level would be …

- Finding a dataset with an associated problem (like classification in this case).
- Finding a dataset without defined problem so that you may define your own.
- Finding a dataset with defined problem, but with scope for you to define too.

Step 2 : Define a problem on the dataset or go with the default.

Example : Spam Text Message Classification dataset from Kaggle
https://www.kaggle.com/team-ai/spam-text-message-classification

Going with the problem defined for the dataset by default.

- Classification of spam messages from a “labelled” set of text data samples.
- Therefore, it falls under the genre of “Supervised Learning” – Classification.
- Your job will be to use the given labelled dataset to build a Binary Classifier.

Defining a new problem based on the same dataset.

- Given the set of good text messages, spams may be considered “Anomalies”.
- Therefore, we can try ideas of “Unsupervised Learning” – Anomaly Detection.
- Your job will be to use the entire dataset to check if spams are truly anomalies.

Extension : Can you build a text-bot that avoids being spammed automatically?

Step 3 : Figure out the proper data representation and features.

Example : Spam Text Message Classification dataset from Kaggle
https://www.kaggle.com/team-ai/spam-text-message-classification

ham	:	Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got a...
spam	:	WINNER!! As a valued network customer you have been selected to receivea £900 prize reward! To claim...

The "default" Textual Data Representation

Strategy 1 : Represent the data in a numeric format.

- Create a dictionary of words and create "word-vectors" out of each message.
- Extract hand-crafted features using domain knowledge – CAPS, numbers etc.

Strategy 2 : Use a tailor-made "model family" in ML.

- Read up Natural Language Processing (NLP) and try to use Neural Networks.

Step 4 : Perform exploratory data analysis on raw/processed data.

Example : Spam Text Message Classification dataset from Kaggle
https://www.kaggle.com/team-ai/spam-text-message-classification

The "default" Textual Data Representation

ham : Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got a...

spam : WINNER!! As a valued network customer you have been selected to receivea £900 prize reward! To claim...

Processed "Numeric" Data

Data representation : Text converted to Numeric

- Create a dictionary of words and create "word-vectors" out of each message.
- Extract hand-crafted features using domain knowledge – CAPS, numbers etc.

Ask questions on the data and extracted features to "illustrate" your problem.

Which features are common in ham and spam? Which ones differentiate them?

Step 5 : Study your chosen model thoroughly before you use it.

Example : Spam Text Message Classification dataset from Kaggle
https://www.kaggle.com/team-ai/spam-text-message-classification

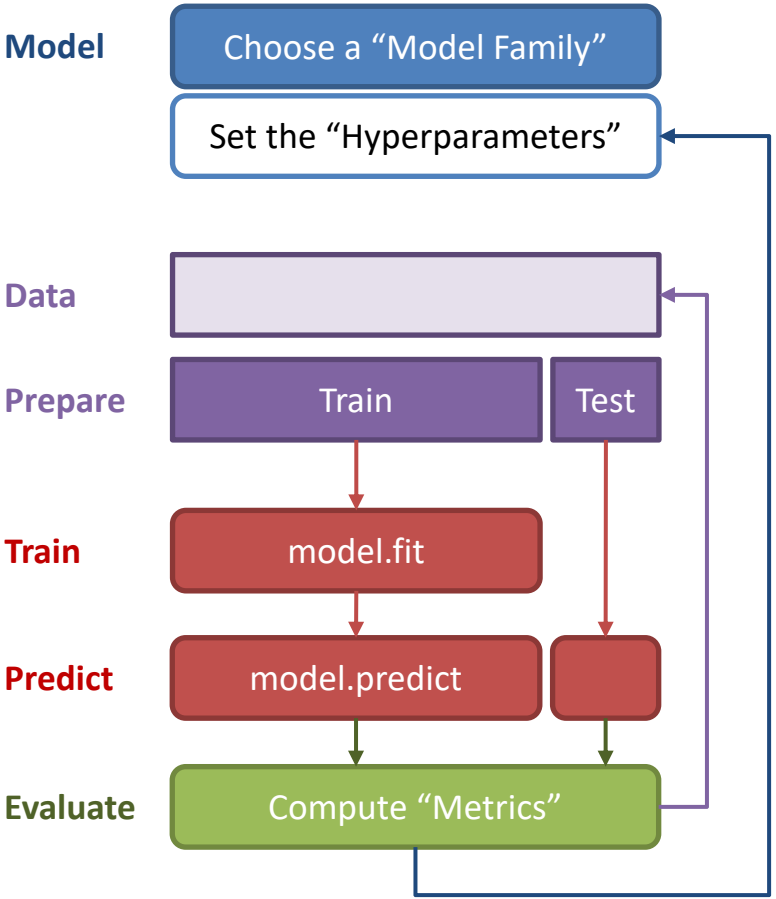
Let's try Random Forest Classifier

Data representation : Text converted to Numeric

- Create a dictionary of words and create "word-vectors" out of each message.
- Extract hand-crafted features using domain knowledge – CAPS, numbers etc.

Random Forest Classifier

- Built using a bunch of Decision Trees put together in an "ensemble model".
- Hyperparameters : The number of trees, maximum depth of each tree, etc.
- Training (fit) : Building a collection of Decision Trees on the training dataset.
- Testing (predict) : Using the Decision Trees together to predict "ham/spam".
- Evaluation Metric : Binary classifier standards – Accuracy, TPR, FPR, F1, AUC.



Step 6 : Set your Model Building process.

Prepare	Train, Validation and Test Data
Train	“Fit” the Model on the Train Set
Test	“Predict” using the trained Model
Evaluate	Performance “metrics” for Model

Once you are happy with your final model(s), you can **derive insights** from the model and the entire exploratory data analysis process.

- Which features/properties differentiate?
- Which features/properties are common?

The 7 steps of Machine Learning : https://youtu.be/nKW8Ndu7Mjw

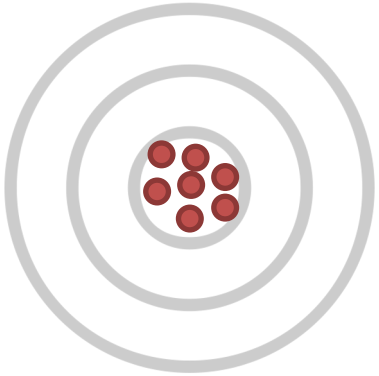
Something you should start thinking about …

ACCURACY OF THE MODEL

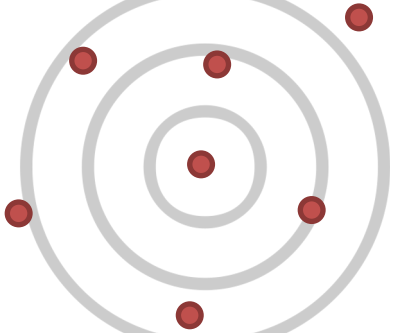

6
5
100
70
2
1

The Dream Model

Low Bias
Low Variance

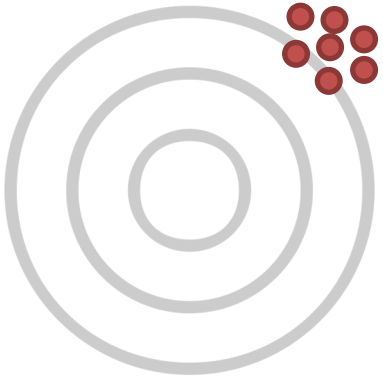


In Practice



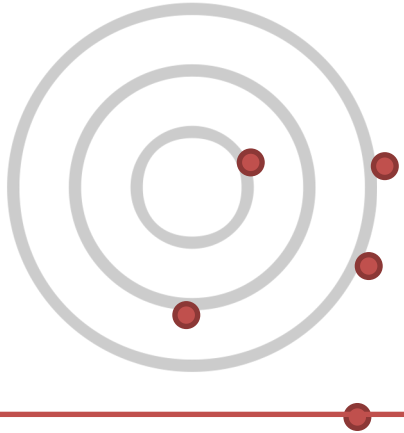
Low Bias
High Variance

High Bias
Low Variance

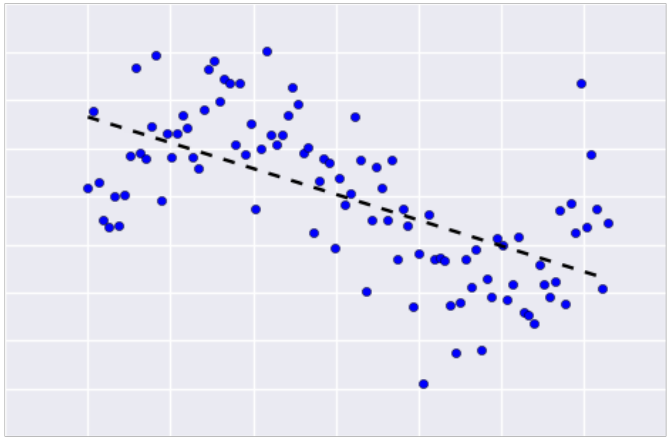


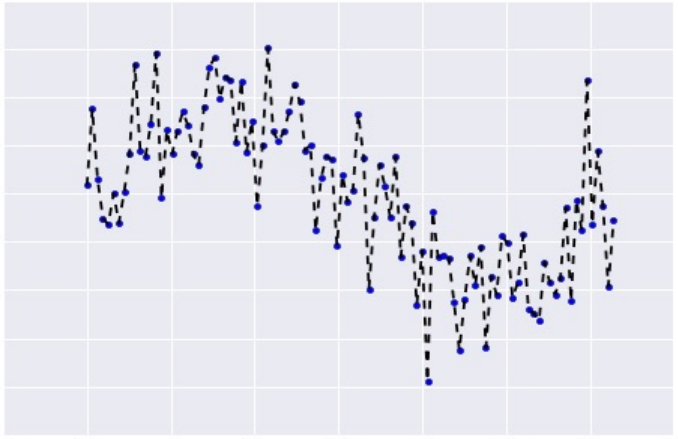
In Practice

High Bias
High Variance

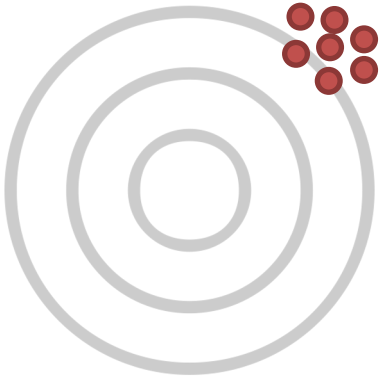


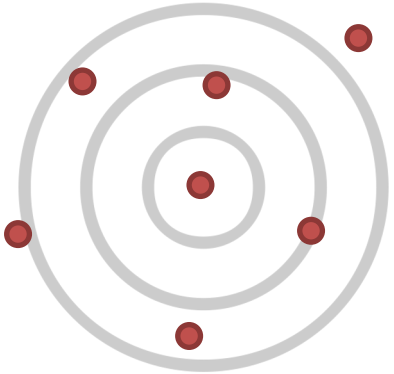
Garbage Model



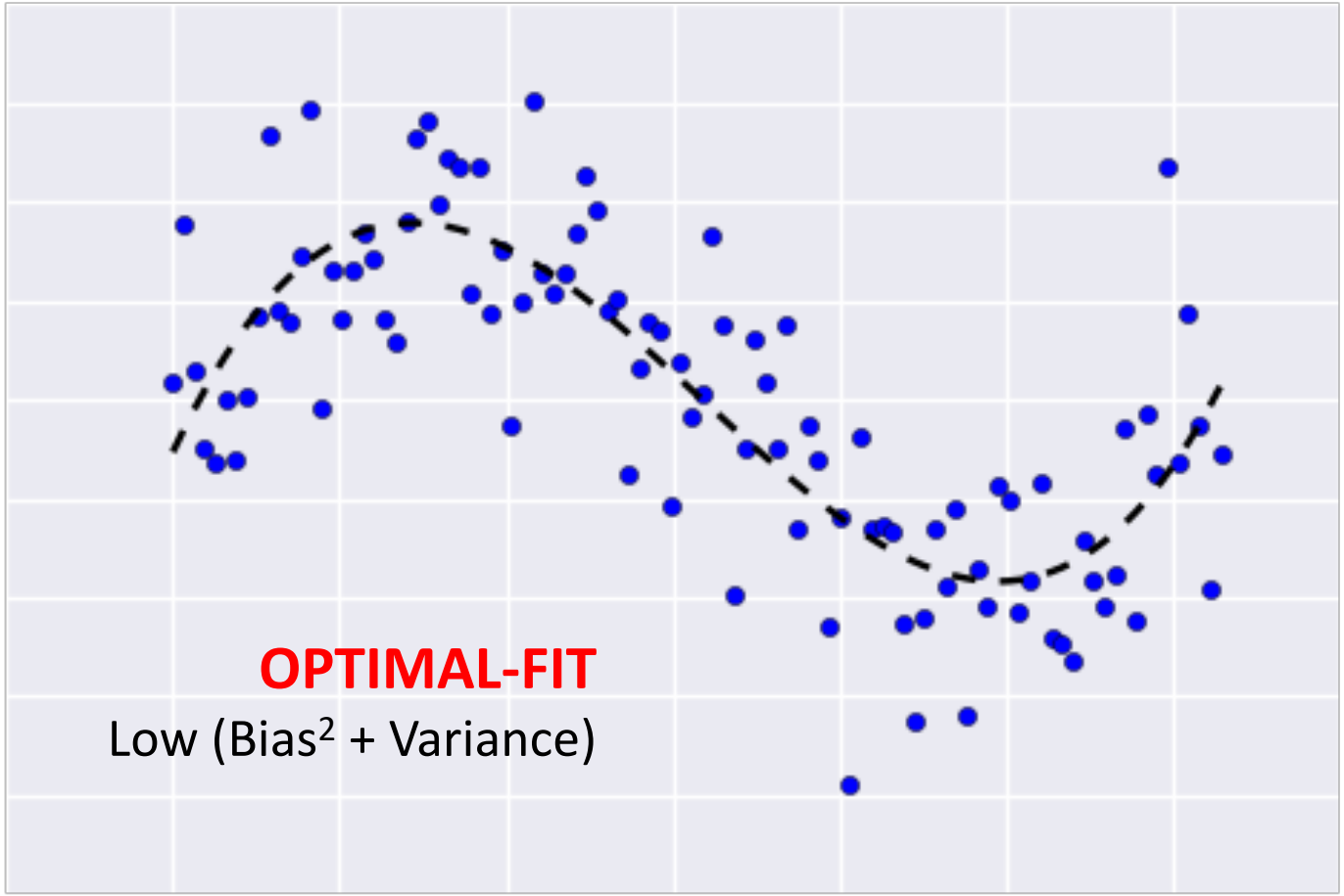


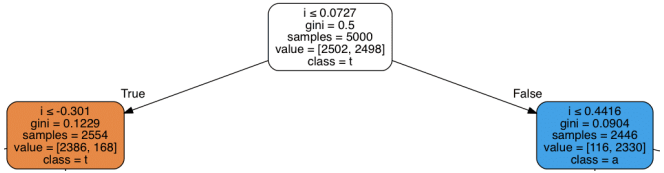
UNDER-FIT
High Bias
Low Variance

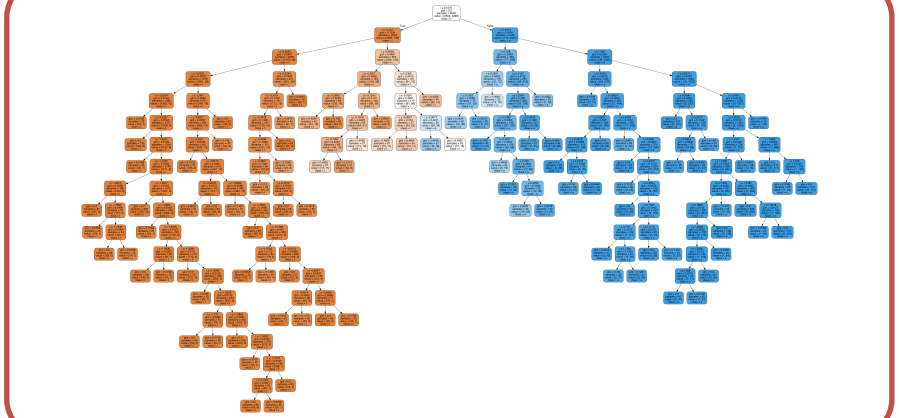




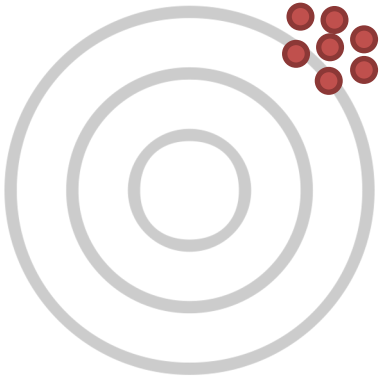
OVER-FIT
Low Bias
High Variance

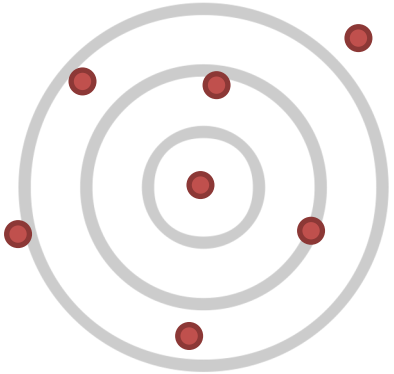

OPTIMAL-FIT
Low (Bias^2 + Variance)



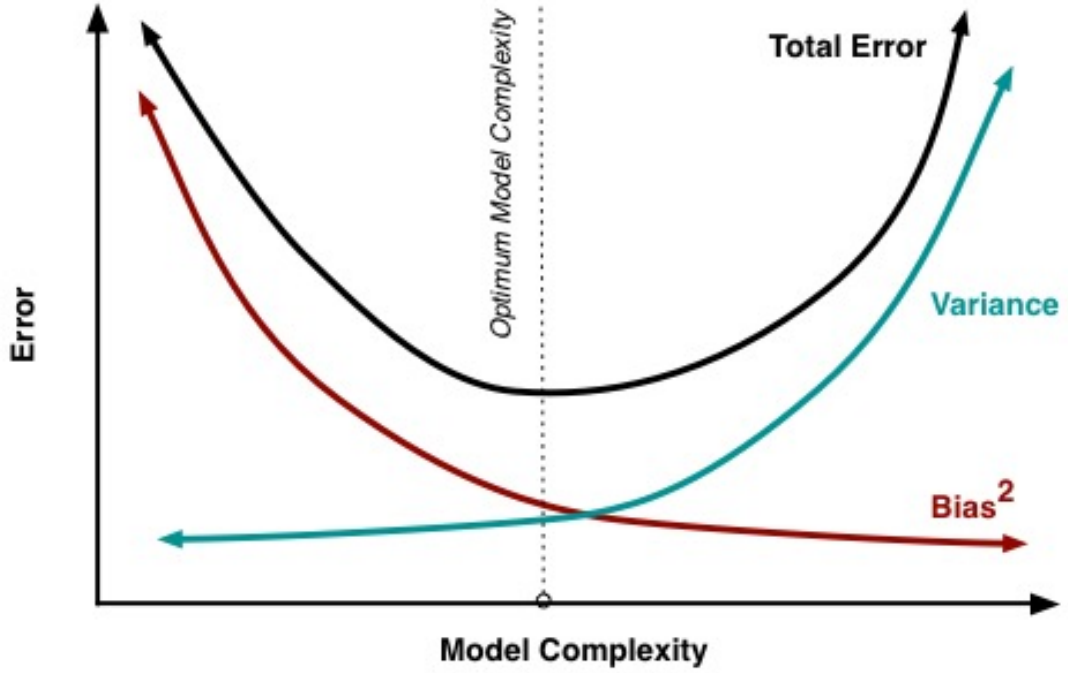


UNDER-FIT
High Bias
Low Variance





OVER-FIT
Low Bias
High Variance

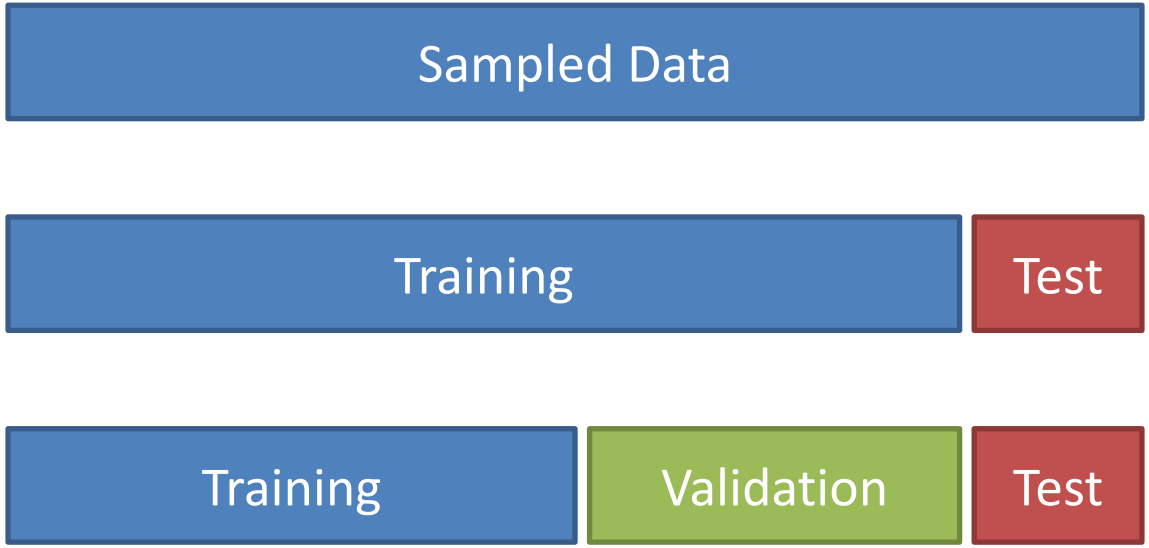


Model Complexity is the Hyper-Parameter to Tune

Hyperparameter Tuning (CV) : https://towardsdatascience.com/cross-validation-and-hyperparameter-tuning-how-to-optimise-your-machine-learning-model-13f005af9d7d

One way to build a generalizable model is …

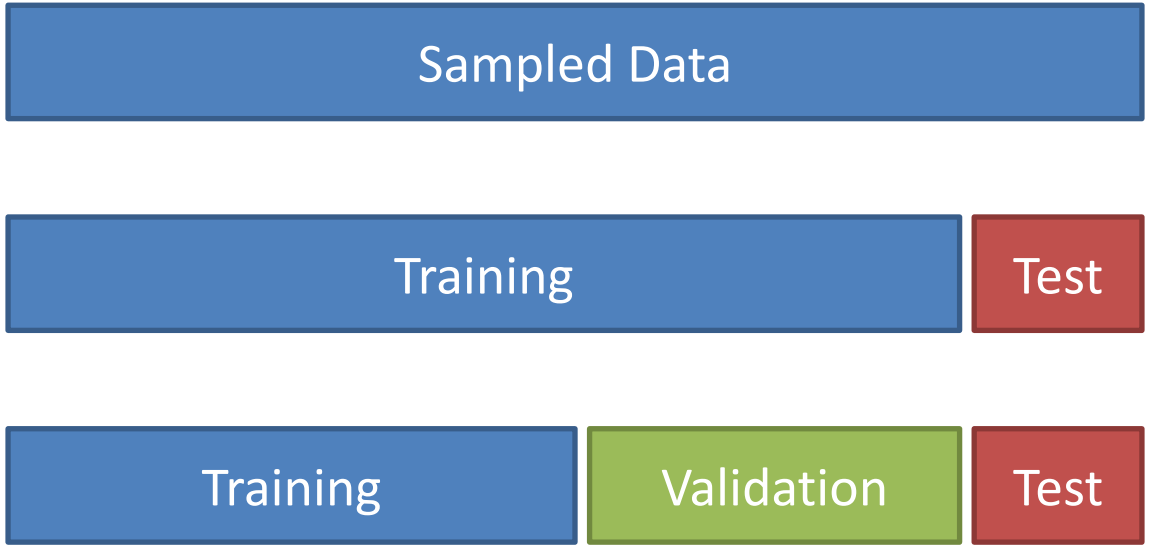
CROSS-VALIDATION



There is a very high chance of Overfitting the labeled Sample Data.

Test Set should never be used for Model re-training or re-learning.

Validation Set offers us scope for Model “testing” and re-learning.



Larger Training Set implies “more Information” in Model Building.
Larger Validation Set implies “more Diversity” in Model Testing.

K-Fold Cross-Validation

Split the dataset into *k* folds/parts

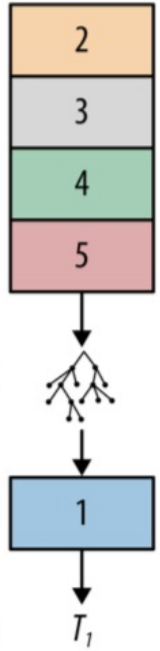
K-Fold Cross-Validation

Use (*k*-1) folds of data for Training

Build the Model

Use the remaining fold for Validation

Note Performance



K-Fold Cross-Validation

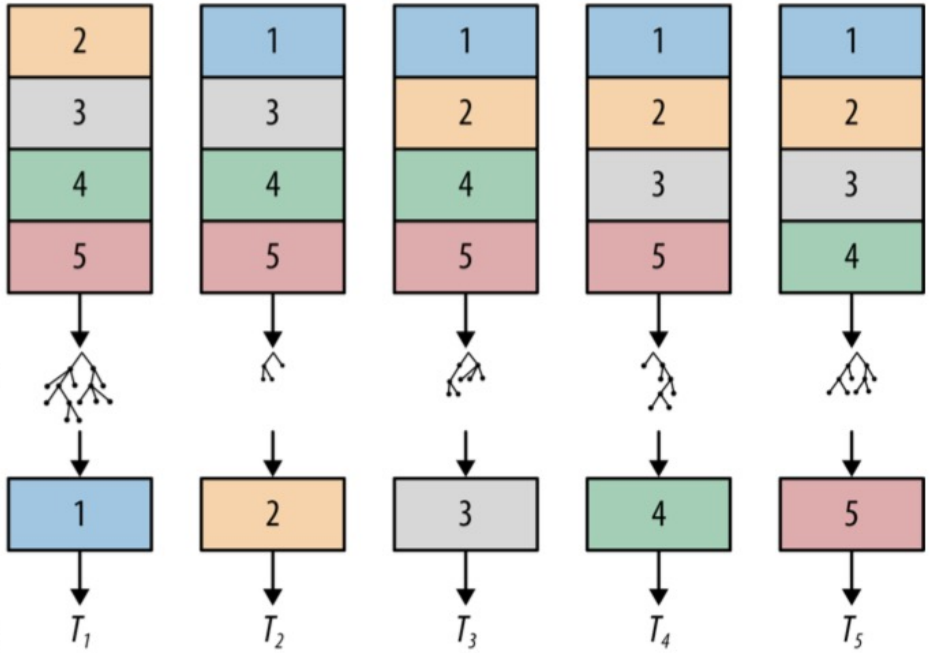
Repeat the process *k* times

Use (*k*-1) folds of data for Training

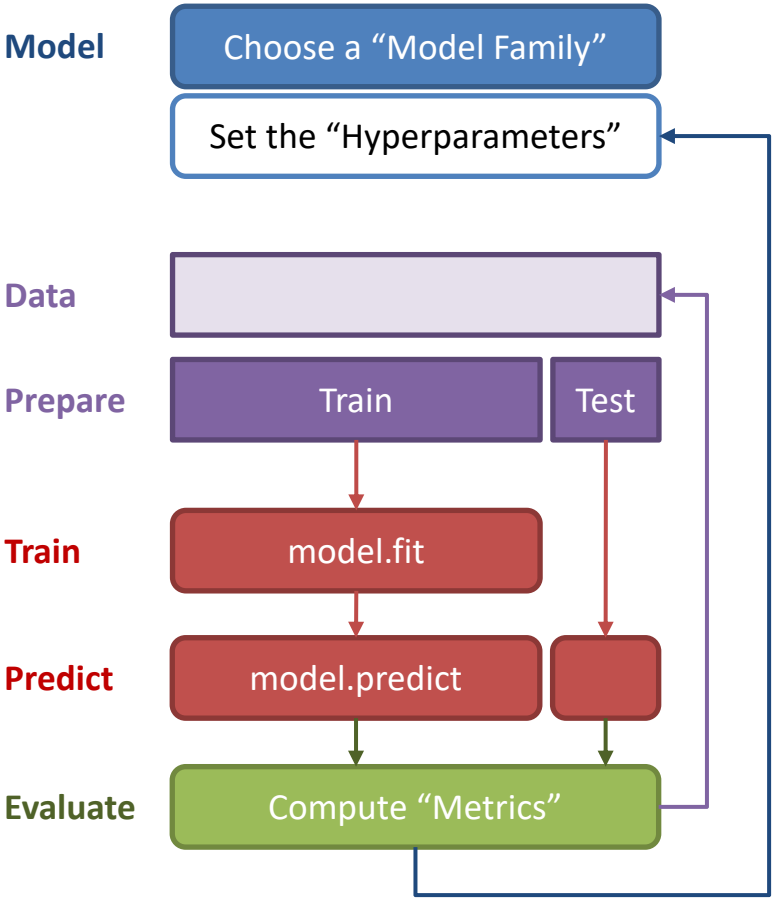
Build the Model

Use the remaining fold for Validation

Note Performance



Note the Mean Performance and its Variance



Do not underestimate the following …

Data Preparation

- Proper representation for your model with feature engineering and feature selection.
- Data pre-processing to fix issues – class imbalance, skewness, scale of variables.

Evaluation Metric

- Different for each problem definition and model family. Also depends on the data.

Presentation of logical insights backed by data analysis. Negative results count too!

Thank You for participating!

Keep learning on you own.