

Common Theme, Toolbox and Research workflow in Data Science

Apply different algorithms to solve different problems based on the same
<Theme> and <Research Workflow>

Algorithms

- SVM
- KNN
- Naïve Bayes
- Neural Network
- Logistics Regression
- NLP

**All these algorithms can
be used to solve classification
problem**

Theme
Research
Workflow
Toolbox

Problems

- Regression
- Classification
- Recommendation System
- Clustering
- Association

Common
Theme in
Machine
Learning

confusion matrix bias
bias vs variance train test split
train test split precision vs recall
features engineering
regularization **overfitting** cross validation
encoding categorical var traintestsplit
features engineering cost function
data normalization r-square
model performance
type ii error type i error

Common Data Scientists Toolbox

Every professional
requires mastery of
their tools box

Every profession
has a standard
protocol



You need to know why, when and how to
different tools properly

This requires understanding of the each of
concepts behind the tools

Common Terminology

Underfit vs Overfit

Precision vs Accuracy

Bias vs Variance

Precision vs Recall

Train/Test Split

In-sample vs Out-of-sample data

Removing Bias

Cross Validation

Normalized Data

Encoding Categorical Variables

Data Normalization

Normalization or scaling refers to bringing all the columns into similar range.

Two common ways: Min-Max normalization and Z-score normalization

$$z = \frac{x - \min(x)}{\max(x) - \min(x)}$$

$$z = \frac{x - \mu}{\sigma}$$

μ = Mean

σ = Standard Deviation

One-Hot encoding or Using dummies variables

Method in bringing in categorical variables as additional features

A categorical variable that can have N possible values will turn into a $(N-1)$ additional dummy variables (or $N-1$ features) where each one of the dummy variable can take only either 0 or 1 and for one of the possible values in the categorical variable. Reason for only $N - 1$ is because when all other dummy variables are 0, it implicitly mean it is for the last N value

Example: Sex has only two possible values: Male and Female. We can create 1 dummy variable, say, Male. And if it is male, Male will be 1, otherwise it is 0. We do NOT need two dummy variable

Example: Education has “Primary”, “High-School”, “College”, “Master” and “PhD”. Then one can create 4 dummy variable, called edu1, edu2, edu3, edu4.

For Primary, the 4 dummy variables values will be (1, 0, 0, 0)

For College, the 4 dummy variables values will be (0, 0, 1, 0)

For PhD, the 4 dummy variables values will be (0, 0, 0, 0)

K-Fold Cross Validation

To make sure your model is not too sensitive to the training samples

Say $K = 10$

Split your dataset into 10 parts, label them as Part1, Part2, Part3, ... , Part10

Run through the following algorithms:

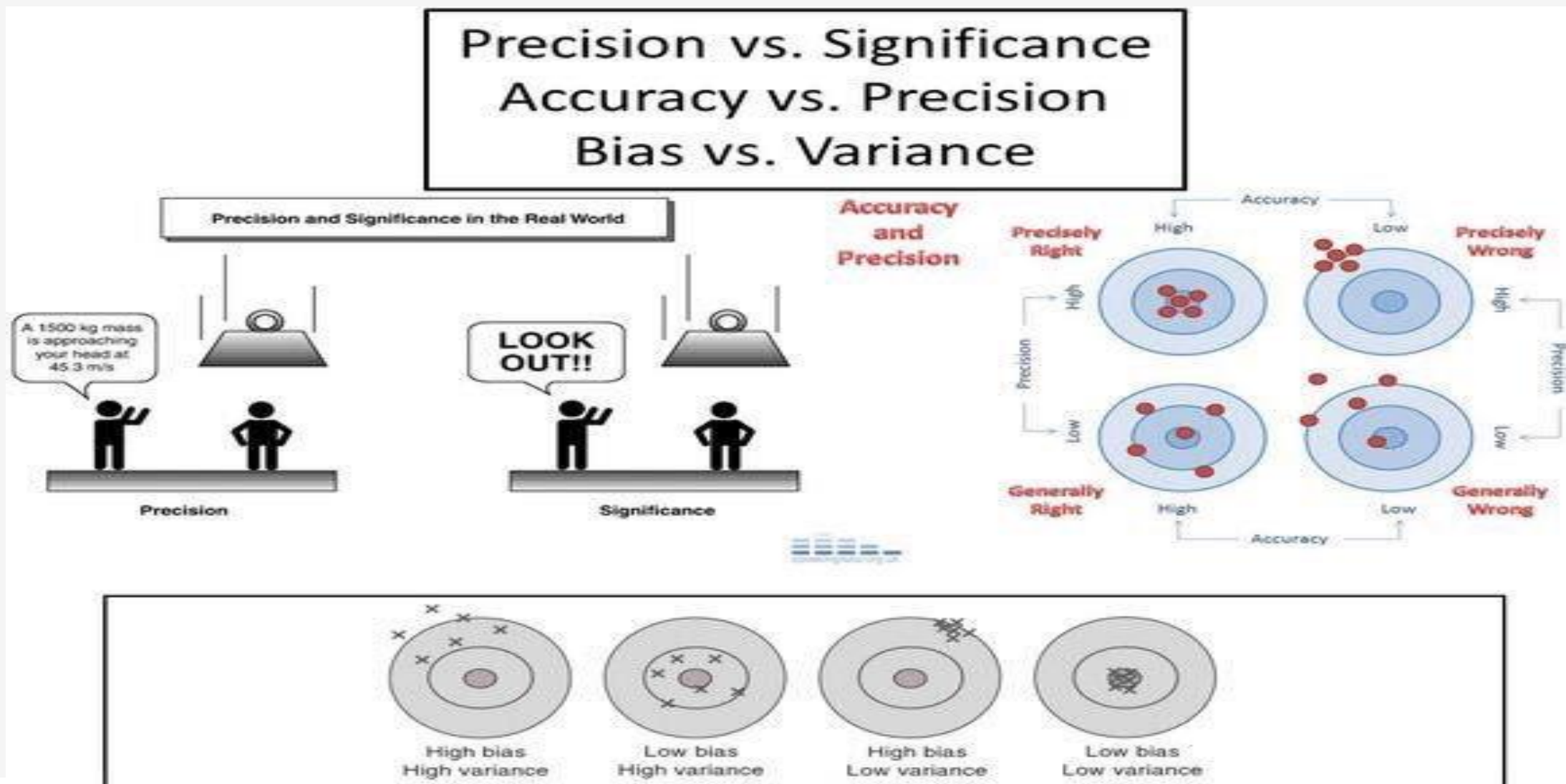
For i from 1 to 10:

Use Part_ i as your testing dataset, all the others Part1, ... Part_ $(i-1)$, Part_ $(i+1)$, ... Part_10 as the training set.

Build a different model, called, Model_ i

Compare all the 10 models, Model_1, Model_2, ..., Model_10 to make sure their performance are similar

Bias vs Variance trade-off



Regularization

- **Dilemma**: Too simple model will underfit (high bias, low variance) while more complicated model will overfit (low bias, high variance)
- **Is there any tools to help finding the right balance?**
 - Idea is to penalize complicated model automatically (adjusted R-square instead of R-square)

Regularize your model so that when you drop data points or model parameters, your model prediction is still pretty much the same

1000 model parameters, drop 100 of them.

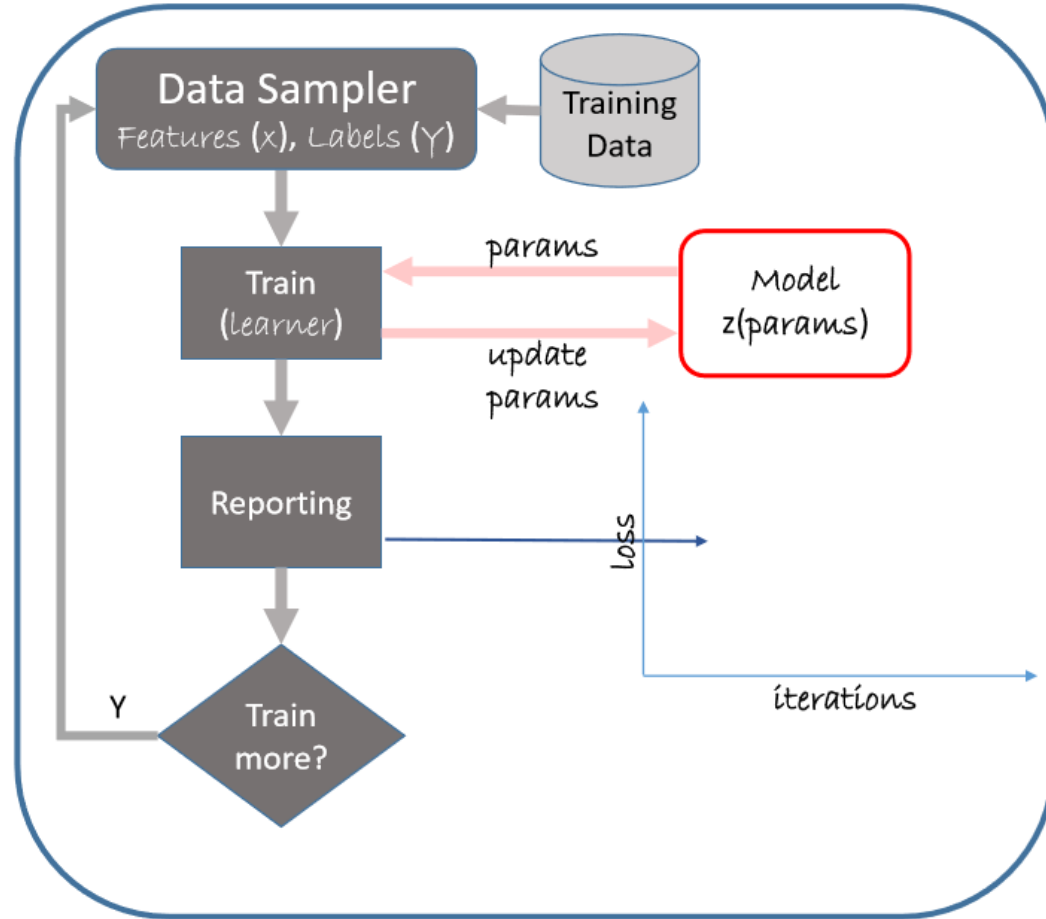
Model depends on the training data set also model assumptions. We want to make sure the model is not sensitive.

Important Concepts

Every professions have a “Best Practice” workflow

Research Workflow (Train-Validate-Test Cycle)

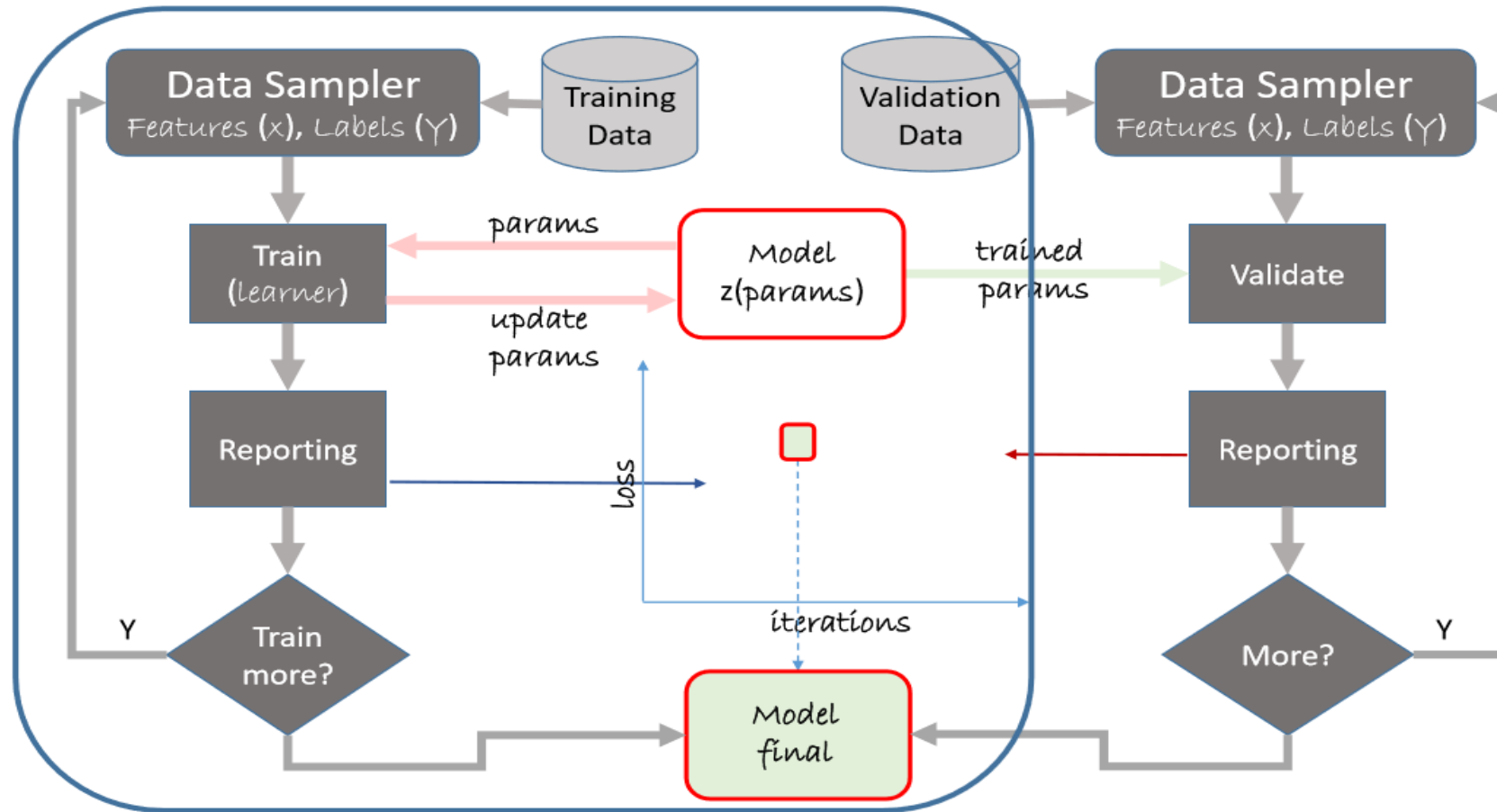
Train Workflow



Reference: Microsoft Professional Program in Artificial Intelligence

Research Workflow (Train-Validate-Test Cycle)

Validation Workflow



Research Workflow (Train-Validate-Test Cycle)

Test Workflow

