MONASH University

MONASH
INFORMATION
TECHNOLOGY

Introduction to Machine Learning

Prajwol Sangat
Updated by Chee-Ming Ting (15 April 2022)

GROUP OF EIGHT AUSTRALIA

Last week

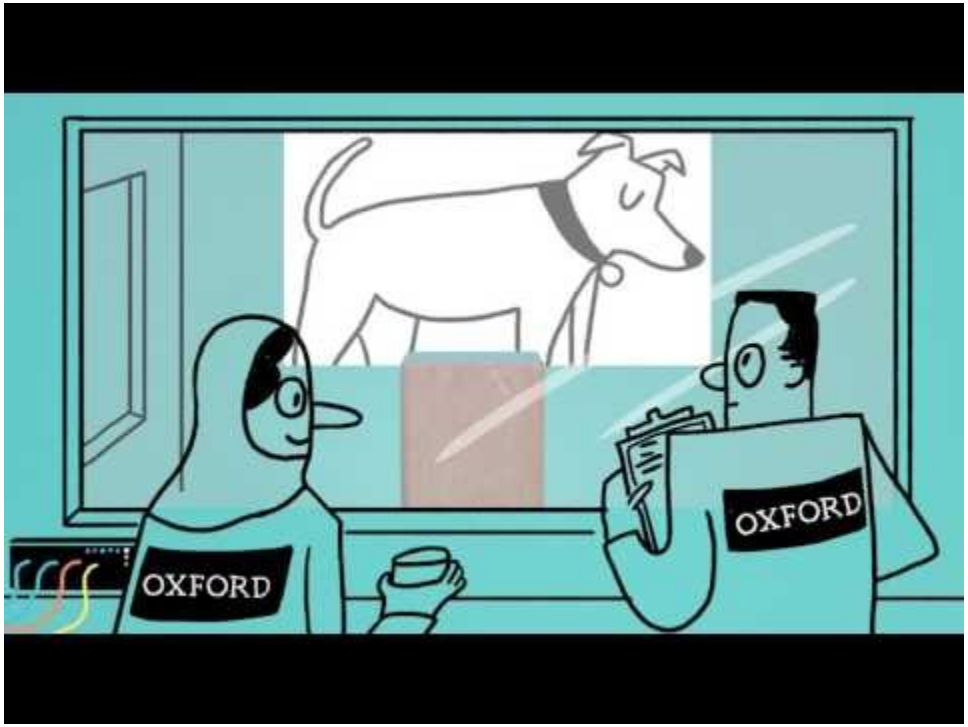
- Parallel Aggregation
  - Parallel Sort
  - Parallel Group-By

This week

- What is Machine Learning?
- Machine Learning Basics
- Types of Machine Learning
- Feature Engineering

According to McKinsey study, 35% of what consumers purchase on Amazon and 75% of what they watch on Netflix is driven by machine learning–based product recommendations.

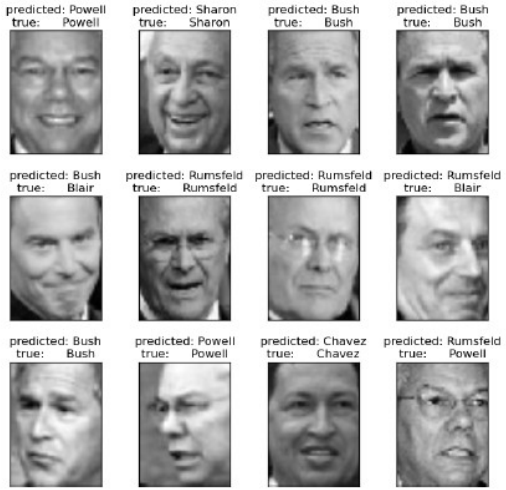
What is Machine Learning?



What is Machine Learning?

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E", (Tom Mitchell, 1997)

Face recognition



Experience E	Task T	Performance P
databases of thousands of known faces	given a new photo, recognise the name of the face	how accurate the recognition is

Examples

Watson, Ken Jennings, and Brad Rutter in their *Jeopardy!* exhibition match.

Experience E	Task T	Performance P
databases of millions of question-answer pairs	given an question, find the best answer	how accurate the answer is

Detecting Spam Emails

Examples of spam emails and not-spam email	To assign a label "spam" or "not-spam" to an email	how accurate spam email can be detected

Detect credit card fraud

Data collected for credit-card transactions deemed as fraud and not-fraud	To assign a label 'fraud' or "not fraud" to a given credit-card transaction	how accurate a credit-card fraud transaction can be detected.

Elements of machine learning

1 Data

feature $x \in \mathbb{R}^d$
label $y \in Y$

Dataset $\mathcal{D} = \{x_i, y_i\}_{i=1}^n$

2 Model

Supervised: $f_\theta : X \to Y$

X is data space
Y is label space
θ: model parameter

3 Assessment

How well is f_θ doing w.r.t data $\mathcal{D}$?

Data processing

feature extraction, feature selection, feature transformation, feature reduction, feature scaling, feature normalization

Predictive Model $Y = f_\theta(X)$

Model Learning (Training)

- Find an optimal model f_θ (by estimating model parameters θ) using training data
- Based on loss function (e.g., minimize error between true and predicted labels)

Model Testing $\hat{y} = f_\theta(x_{test})$

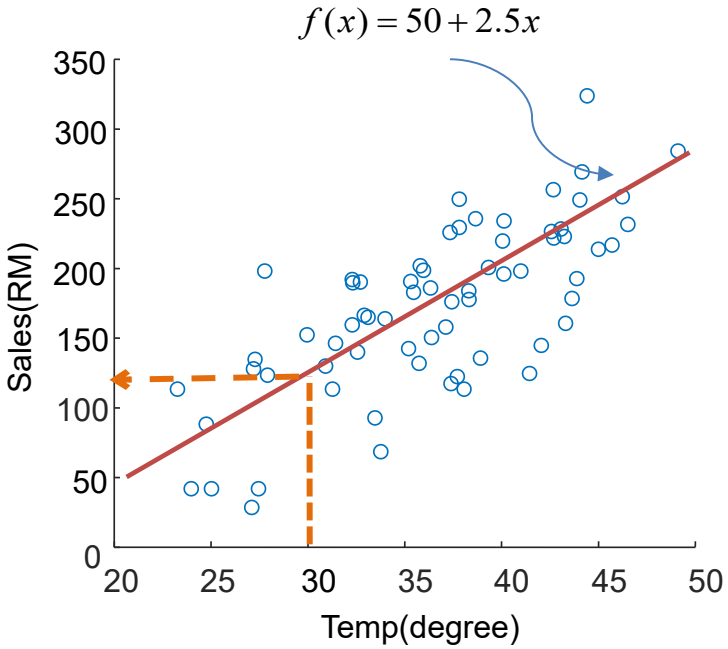
- Test the learned model in predicting unseen test data
- Performance metrics to assess model accuracy

Illustration: Linear Regression model

Problem: Predict ice cream sales given temperature

Data

Day	Temp	Sales
1	36	200
2	31	100
3	24	50
...	...	...
100	38	250



Predictive Model:

- What is good model *f*(.) to maps *x* to *y*?

$$f(x) = \theta_o + \theta_1 x$$

Model Learning/Estimation:

- How to choose parameters θ_o, θ_1?
  - ❑ Define loss function
  - ❑ Estimate using learning algorithm

Estimated parameters: $\theta_o = 50,\ \theta_1 = 2.5$

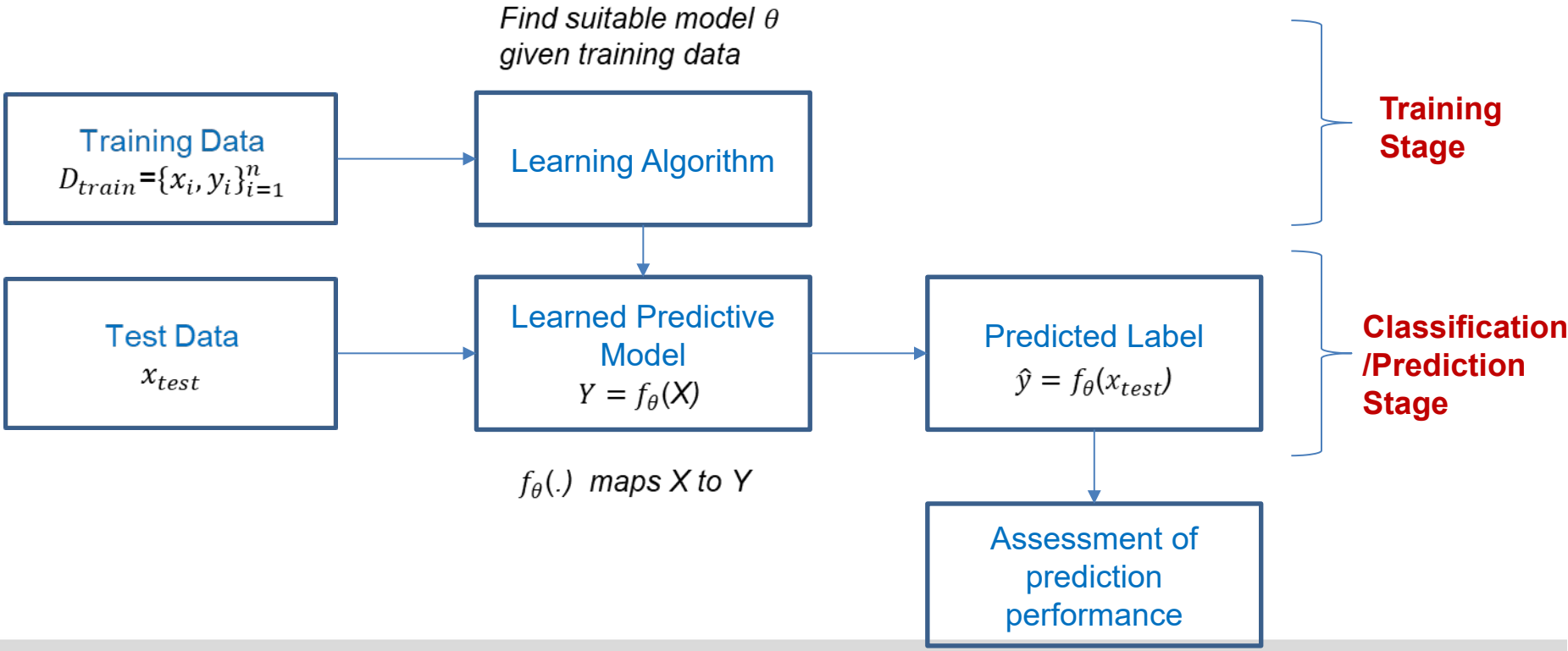
Prediction:

- Given new input, predict *y* with learned model $\hat{y} = f(x_{new}) = \theta_o + \theta_1 x_{new}$

Predicted output $\hat{y} = 50 + 2.5(30) = 125$

Assessment: $\text{error}(\hat{y}, y)$

Overview of machine learning



Data

Features: x_i

- a set of attributes, each is usually in form of a vector or matrix.
- E.g., represent each email (data point) into a bag-of-word vector (feature); or a face photo into a real-valued matrix.

Labels: y_i

- values, categories, classes, assigned to data points.
- E.g., 0 = non-spam, 1 = spam,

Data points (aka instances, samples) $\{x_i\}$ or $\{x_i, y_i\}$

- these are items or instances of data used for training and evaluating ML models.
- E.g., labelled emails in spam detection; transaction data in credit card fraud detection; a photo in face recognition.

data points ... $\{x_i\}$

data points with labels ... $\{x_i, y_i\}$

$$\begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ & & \dots & \\ 0 & 0 & \dots & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 2 & \dots & 0 \\ & & \dots & \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

x_i features

y_i 0 = Jack, 1 = John, etc ...

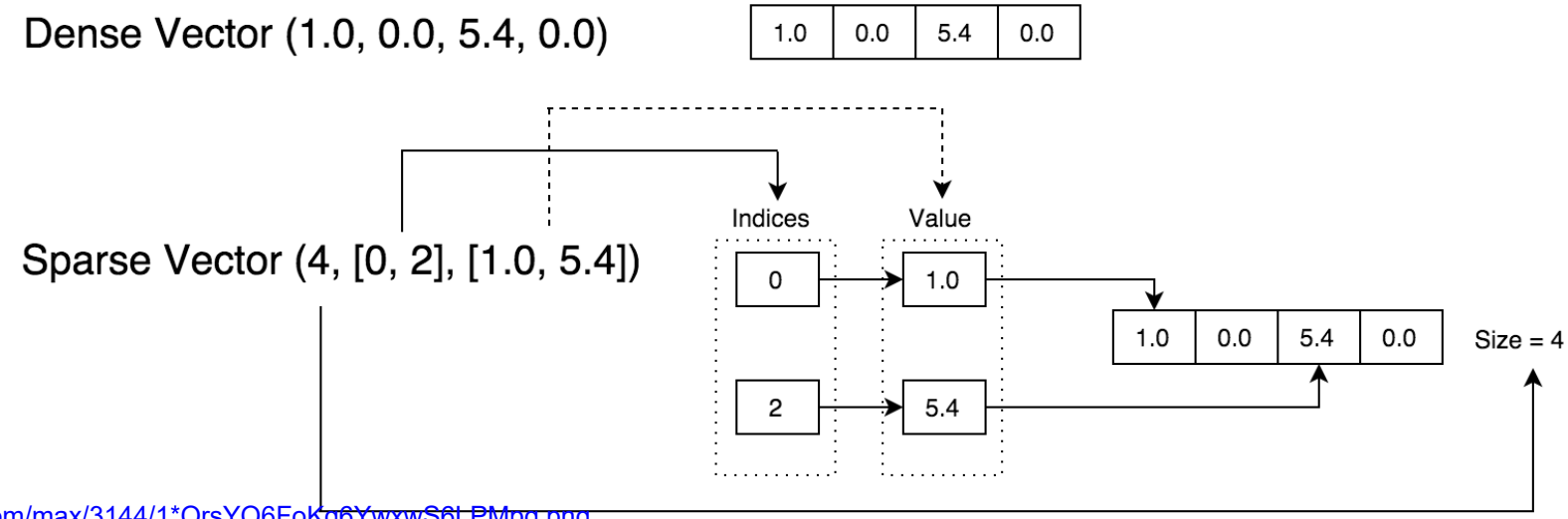
Dataset with n samples: $\mathcal{D} = \{x_i, y_i\}_{i=1}^{n}$

Machine Learning: Data Types

- **Vector**
  - A mathematical vector.
  - *dense vectors*, where every entry is stored, and
  - *sparse vectors*, where only the nonzero entries are stored to save space.
- **Labeled Point**
  - A labeled data point for supervised learning algorithms such as classification and regression.
  - Includes a feature vector and a label (which is a floating point value).

Machine Learning: Data Types

- **Vector**
  - A mathematical vector.
  - *dense vectors*, where every entry is stored, and
  - *sparse vectors*, where only the nonzero entries are stored to save space.



https://miro.medium.com/max/3144/1*OrsYQ6FoKq6YwxwS6LPMpg.png

Features

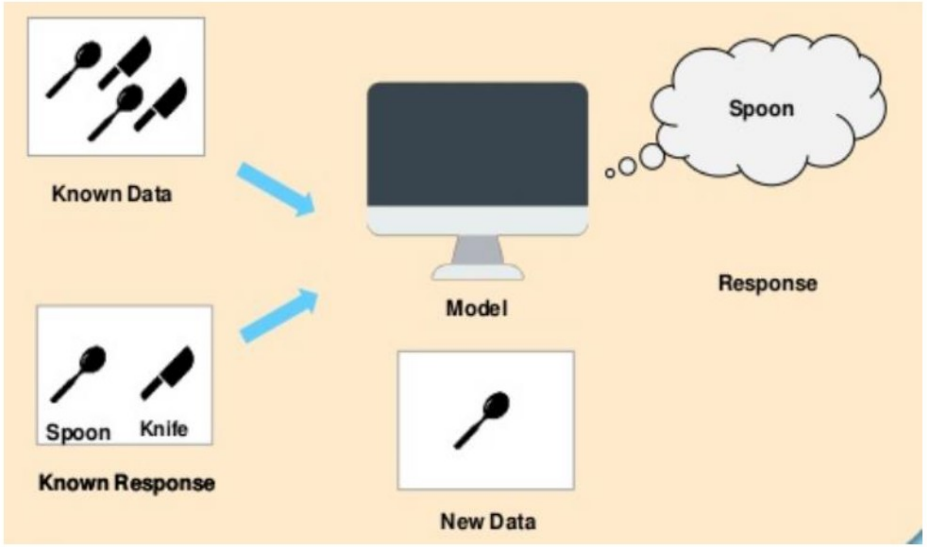
- All learning algorithms require defining a set of *features* for each item, which will be fed into the learning function.
  - *For example, for an email, some features might include the server it comes from, or the number of mentions of the word free, or the color of the text.*
- In many cases, defining the right features is the most challenging
  part of using machine learning.
  - *For example, in a product recommendation task, simply adding another feature (e.g., realizing that which book you should recommend to a user might also depend on which movies she's watched) could give a large improvement in results.*

Machine Learning Fundamentals

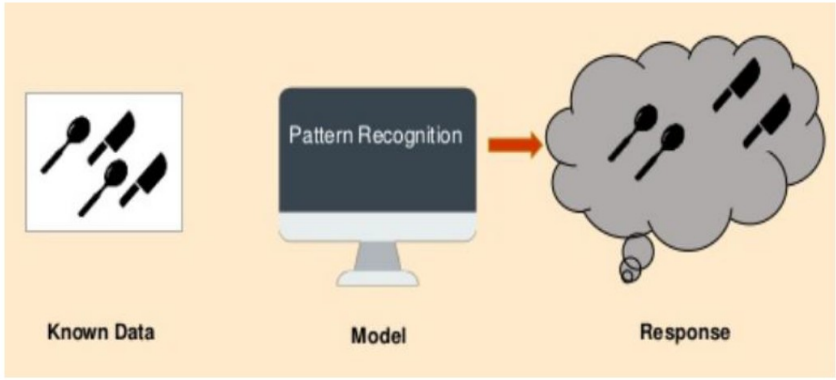
- Supervised and Unsupervised Models
- Bias and Variance
- Underfitting and Overfitting

Model: Types of Model Learning

Supervised



Unsupervised



https://www.simplilearn.com/tutorials/machine-learning-tutorial/supervised-and-unsupervised-learning

Types of Model Learning: Supervised

- Goal: Learn a function from labelled training data to predict the output label(s) given a new unlabeled input.
- Training data consists of input features and output information (labels)
- Two types of supervised learning:
  - ❑ Classification
  - ❑ Regression

Data: $(x_1, y_1), \ldots, (x_n, y_n)$

Function: $f: X \rightarrow Y$

x = feature
y = a discrete label (classification),
y = a continuous value (regression)

Supervised Machine Learning: Classification

Classification problem: To separate inputs into a discrete set of classes or labels

- Binary classification
- Multinomial (Multi-class) classification

Binary classification example: dog or not dog

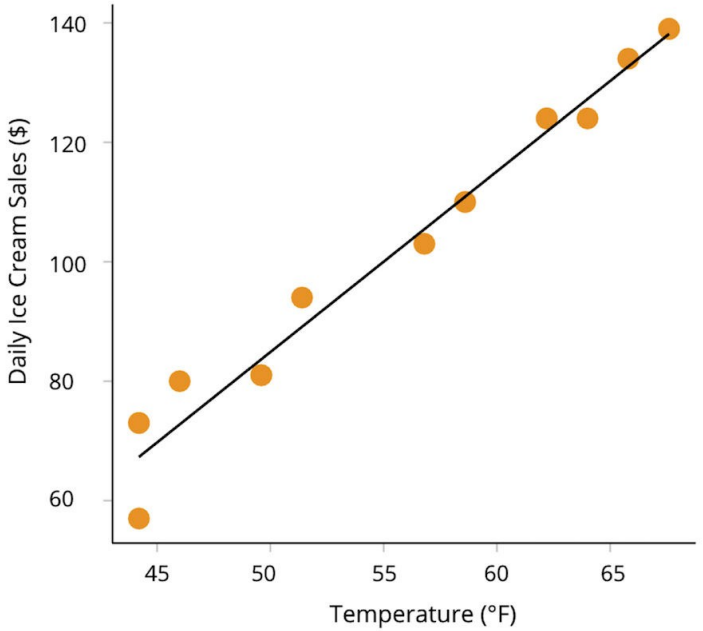
Supervised Machine Learning: Classification



Multinomial classification example: Australian shepherd, golden retriever, or poodle

Supervised Machine Learning: Regression

- A regression problem is when the output variable is a real value, such as “dollars” or “weight”.



Regression example: predicting ice cream sales based on temperature

Supervised Machine Learning in Apache Spark

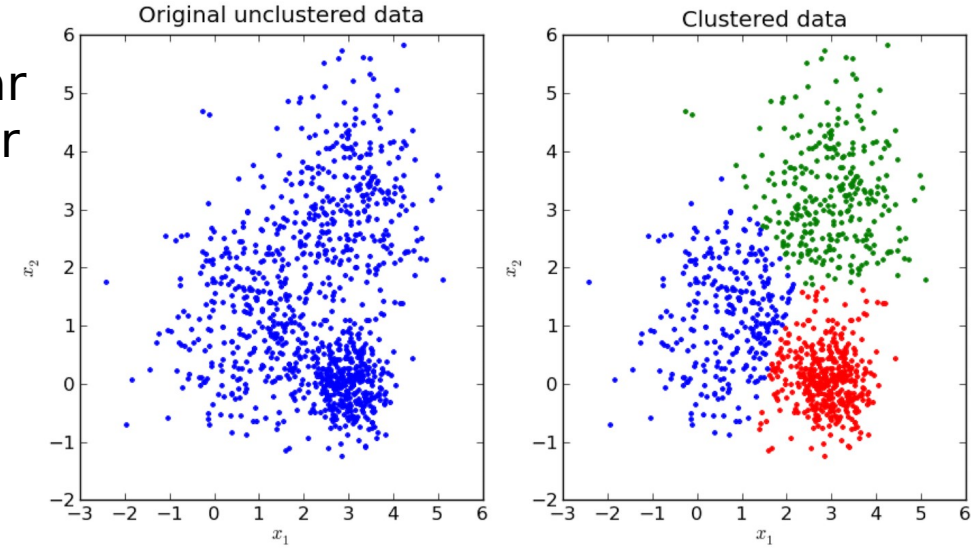
Algorithm	Typical usage
Linear regression	Regression
Logistic regression	Classification (we know, it has regression in the name!)
Decision trees	Both
Gradient boosted trees	Both
Random forests	Both
Naive Bayes	Classification
Support vector machines (SVMs)	Classification

Types of Model Learning: Unsupervised

- Goal: Explore the underlying structure of the data to extract meaningful information. without guidance of known output info.
- Deals with unlabelled data (no output labels)
- Two types of unsupervised learning:
  - Clustering
  - Association

Unsupervised Machine Learning: Clustering

- Clustering problem: Divide data into clusters which are similar between them and are dissimilar to the data belonging to another cluster.
- Where you want to discover the inherent groupings in the data, eg. grouping customers by purchasing behaviour

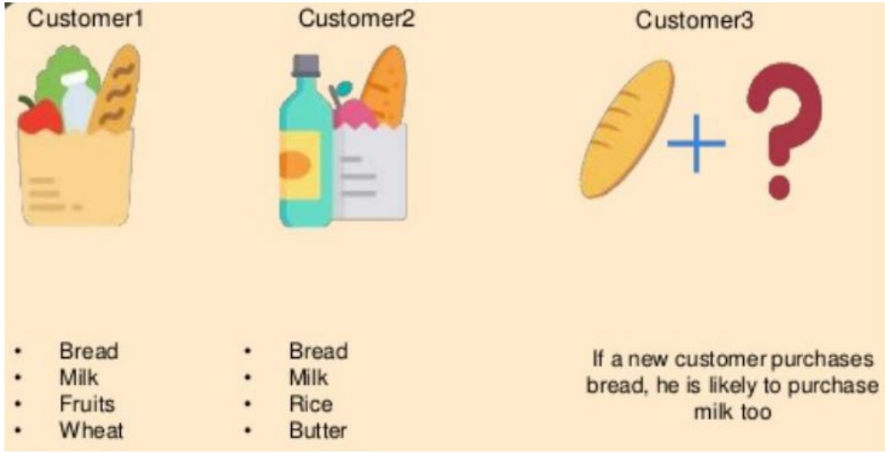


Clustering example

Unsupervised Machine Learning: Association

- Association rule learning problem: Discover the probability of the co-occurrence (association) between items in a large dataset

- Where you want to discover rules that describe large portions of your data, e.g., people who buy X also tend to buy Y.



Unsupervised Machine Learning in Apache Spark

- *k*-means,
- Latent Dirichlet Allocation (LDA), and
- Gaussian mixture models.

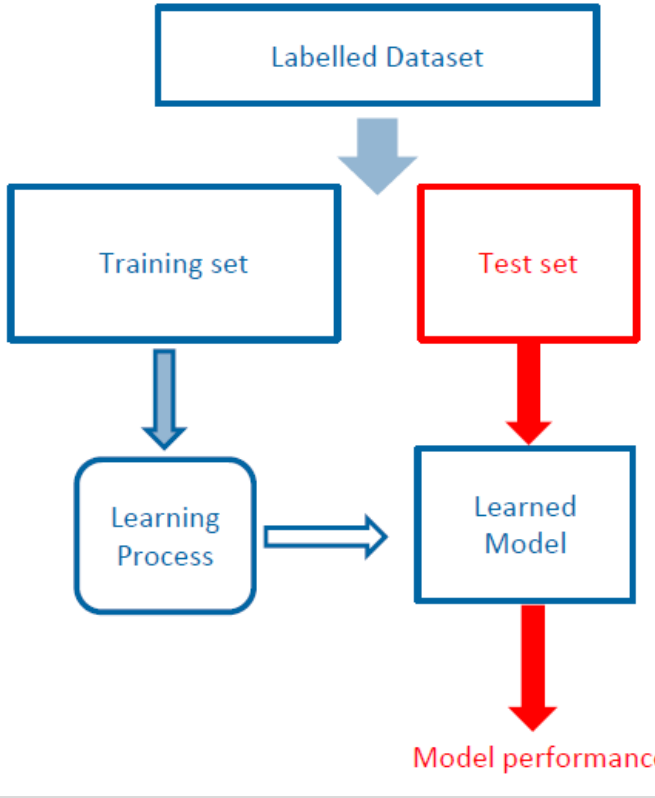
Machine Learning: Assessment

How to prepare the data?

- Train-Test split
- K-fold cross-validation

How to measure performance?

- TP, FP, TN, FP, confusion matrix
- Accuracy, Recall, Precision, F1-score



Day	Temp	Sales
1	36	200
2	31	100
3	24	50
⋮	⋮	⋮
100	38	250

80% - Train set

20% - Test set

Machine Learning: Performance Metrics

Example: Email Spam Detection

In test set: 10 spam, 20 non-spam

Positive = spam

True labels

Predicted labels	SPAM (1)	NON-SPAM (0)
SPAM (1)	7	5
NON-SPAM (0)	3	15

$$\text{Accuracy} = \frac{\#\ correctly\ classified\ samples}{\#\ test\ samples} \text{ x } 100\% = \frac{7+15}{10+20} \text{ x } 100\% = 70\%$$

actual class

predicted class	positive	negative
positive	true positives (TP)	false positives (FP)
negative	false negatives (FN)	true negatives (TN)

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

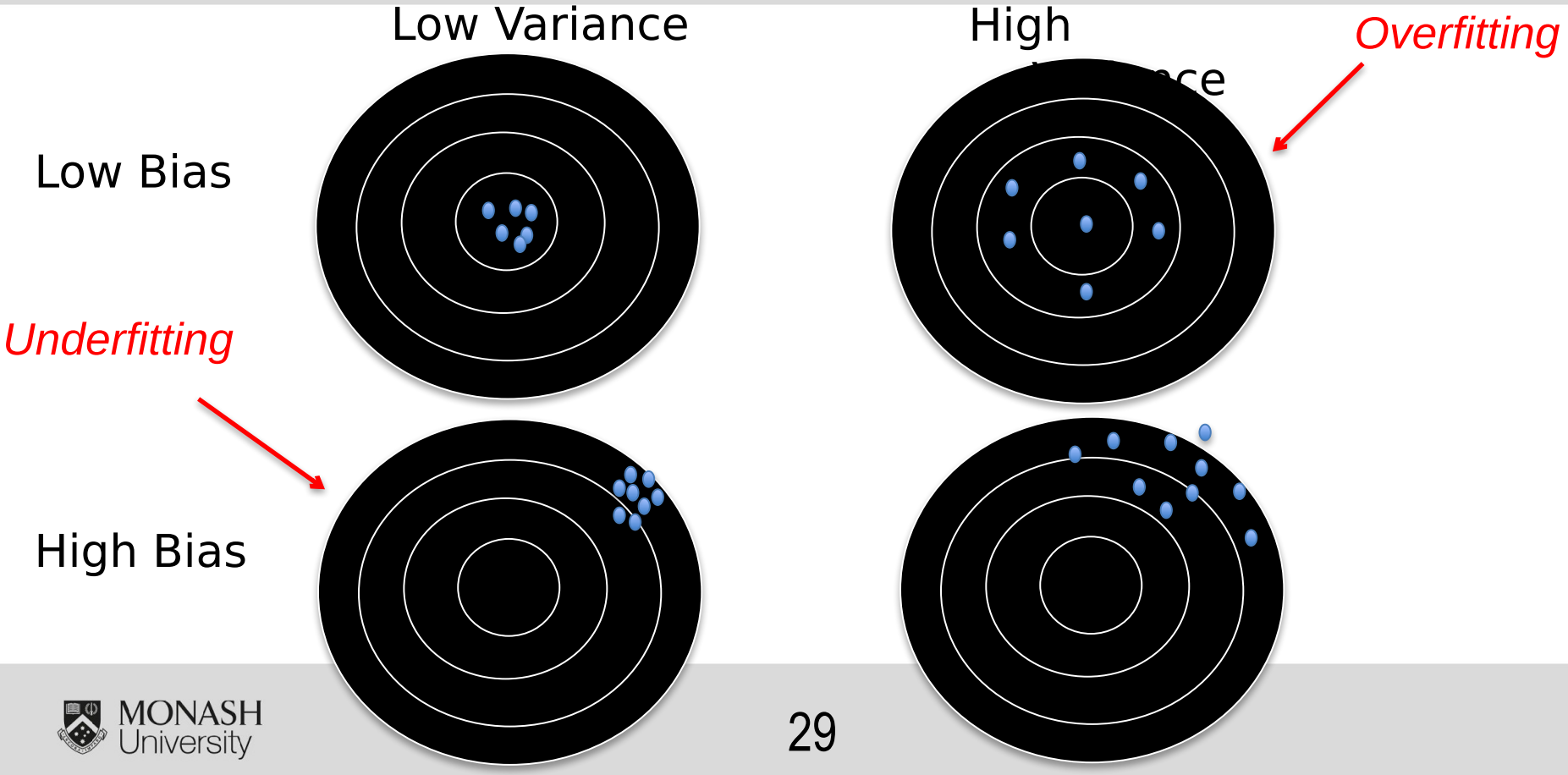
Precision - number of positive class predictions that actually belong to the positive class.
Recall (sensitivity) - number of positive class predictions made out of all positive testing examples.

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Machine Learning: Bias and Variance

- **Bias** is the gap between the averaged predicted value by the model and the actual value of the data.
- **Variance** measures the distance of the predicted values in relation to each other.

Machine Learning: Bias and Variance



Machine Learning: Overfitting and Underfitting

- **Overfitting** (high variance, low bias) is a model that performs well on the training data but generalizes poorly to any new data.

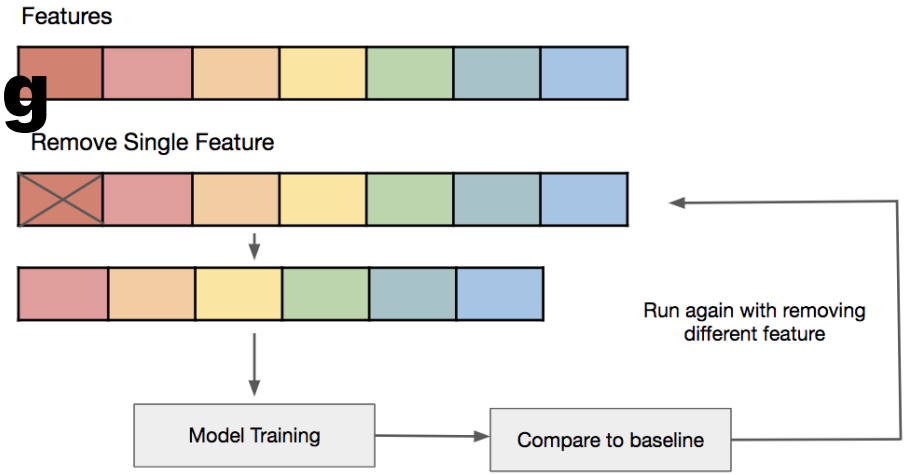
- **Underfitting** (low variance, high bias) is an overly simple model that does not perform well even on the training data.

Machine Learning: Overfitting and Underfitting

- **Preventing Overfitting**
  - **Train with more data**

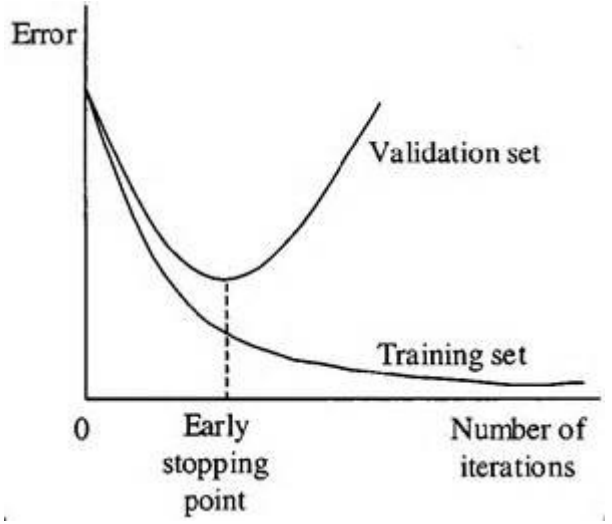
Machine Learning: Overfitting and Underfitting

- **Preventing Overfitting**
  - **Train with more data**
  - **Remove features**



Machine Learning: Overfitting and Underfitting

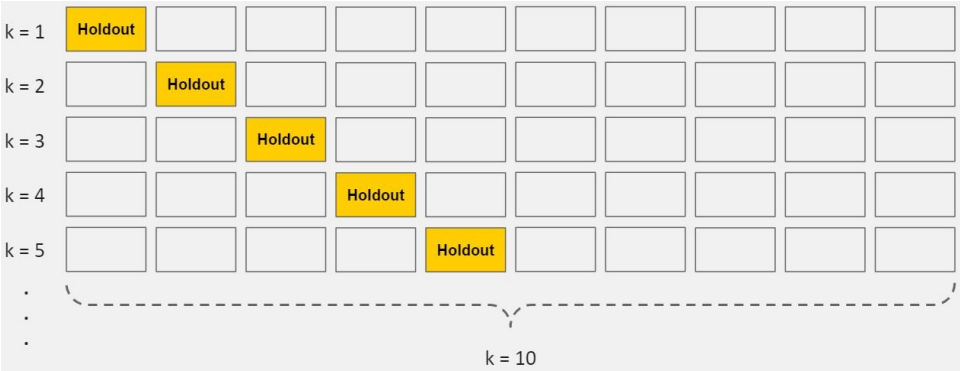
- **Preventing Overfitting**
  - **Train with more data**
  - **Remove features**
  - **Early stopping**



Machine Learning: Overfitting and Underfitting

- **Preventing Overfitting**
  - **Train with more data**
  - **Remove features**
  - **Early stopping**
  - **Cross validation**

K-Fold Cross-Validation



To be continued..

Next topic → Featurization