IFN509

Data Exploration and Mining

Week 8

Predictive Data Mining: Introduction

Professor Richi Nayak
r.nayak@qut.edu.au

School of Computer Science
Centre for Data Science
Faculty of Science
https://research.qut.edu.au/adm

Weeks 8 - 11 Learning

- Lectures: Predictive Mining
  - Predictive mining process
  - Decision tree classification
  - Linear and logistic regression
  - Neural Networks
  - K-nearest neighbour (a brief introduction)
- Computer Tutorials (Weeks 10, 11 & 12)
  - Part 1 - Reflective Pen-and-Paper exercises
    - Decision trees, regression and neural networks
  - Part 2 - Practical Exercises
    - Building, evaluating and comparing decision tree models
    - Building, evaluating and comparing logistic regression models
    - Building, evaluating and comparing neural network models

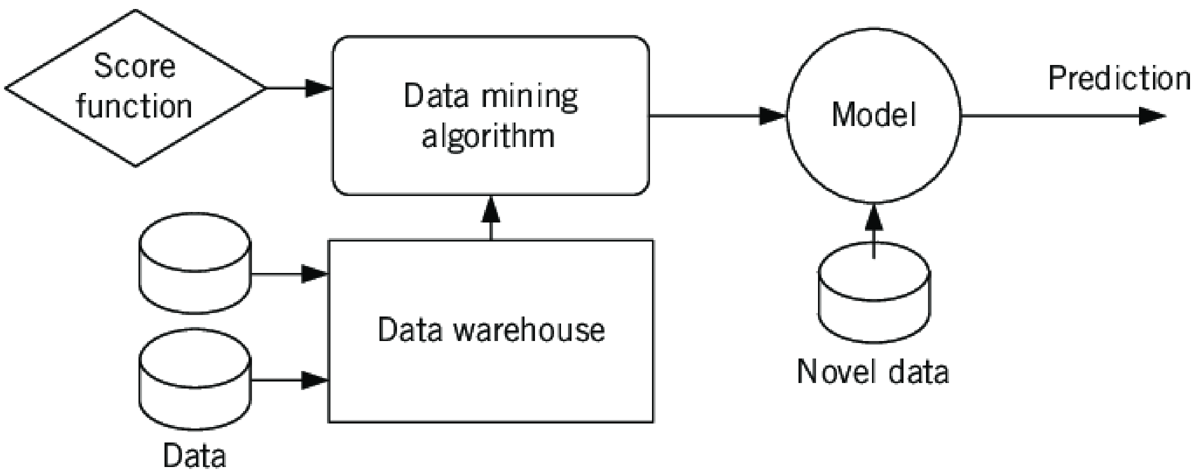
Learning Objectives: Week 8

- What is predictive data mining (classification)?
- Predictive Data Mining
  - Basic Concepts
    - Supervised Learning Process
    - Overfitting or Poor Generalisation
    - Class prediction (classification) vs Value prediction (regression)
  - A quick overview
    - Decision Tree; Neural Networks; Nearest Neighbours; Logistic Regression
  - Evaluation Measures

What Should You Do in Week 8?

- Review the lecture slides and reading materials.
- Attempt the exercise questions on Association Mining in the tutorial
- Complete the Python tasks concerning Association Mining
- Consult the Lecturer/tutor if you have any questions related to the subject.
- Assessment Item 2
  - Read through the specifications
  - Register your team on Canvas
  - Association mining: Should be attempted

Classification: Introduction



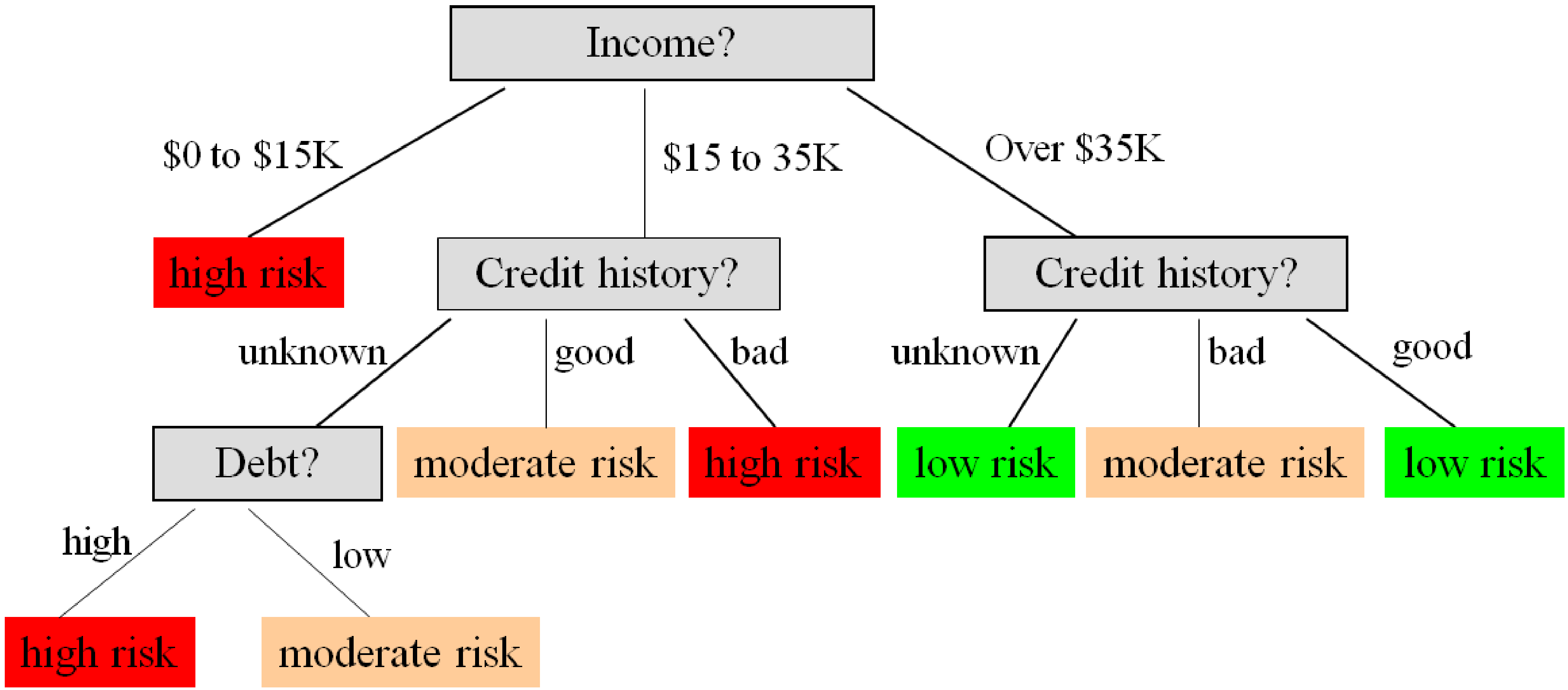
Predictive Modeling: Classification

- Classifying observations/instances into different “given” classes or predicting a value
- Use historical data to make predictions (learn from it)
- Example (Infectious Disease Survival Rate Prediction))
  - Build a model of users based on their history (disease symptoms, physiological data, demographic data, clinical data etc.), output whether the patient made the recovery
  - Exploit factors that lead to survival
  - Predict the chance of recovery for the patient so the treatment strategies can be built accordingly
- Other Examples
  - Classifying credit applicants as low, medium, or high risk; Attrition prediction; Using climate conditions to predict play/not play for a particular event.

Classification Dataset: An Example

No.	Risk	Credit History	Debt	Collateral	Income
1	high	bad	high	none	$0 to $15k
2	high	unknown	high	none	$15 to $35k
3	moderate	unknown	low	none	$15 to $35k
4	high	unknown	low	none	$0 to $15k
5	low	unknown	low	none	over $35k
6	low	unknown	low	adequate	over $35k
7	high	bad	low	none	$0 to $15k
8	moderate	bad	low	adequate	over $35k
9	low	good	low	none	over $35k
10	low	good	high	adequate	over $35k
11	high	good	high	none	$0 to $15k
12	moderate	good	high	none	$15 to $35k
13	low	good	high	none	over $35k
14	high	bad	high	none	$15 to $35k

Classification Model: Decision Tree An Example (cont.)



Classification

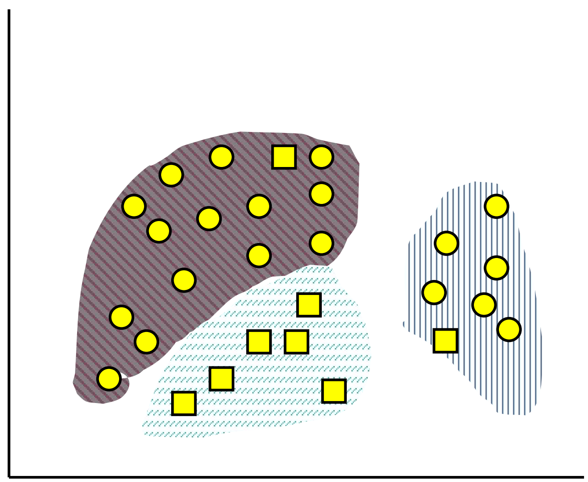
- Like clustering, classification is the organization of data into classes
  - however, class labels are known and it is up to the classification algorithm to use the label information to distinguish the data by learning general features of each class.
  - called supervised classification, because the classification is dictated by given class labels

	Sepal length	Sepal width	Petal length	Petal width	Type
1	5.1	3.5	1.4	0.2	Iris setosa
2	4.9	3.0	1.4	0.2	Iris setosa
…					
51	7.0	3.2	4.7	1.4	Iris versicolor
52	6.4	3.2	4.5	1.5	Iris versicolor
…					
101	6.3	3.3	6.0	2.5	Iris virginica
102	5.8	2.7	5.1	1.9	Iris virginica
…					

Clustering vs. Classification

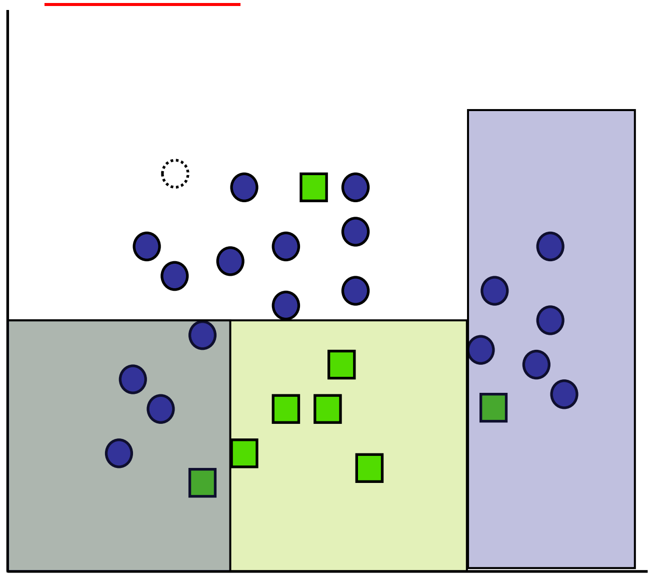
Clustering: Unsupervised learning

Finds "natural" grouping of instances given un-labeled data



Classification: Supervised learning

Learns a model for predicting the instance class from pre-labeled instances



- Both aim to partition data (high-dimensional) into groups / classes/ clusters
- Data items within a group are as similar to each other as possible, but are dissimilar to data items in other groups

Association Mining vs Classification

- Association mining can be applied if **no class is specified** and **any** kind of **structure** is considered "**interesting**"
- Association mining
  - Data is sparse.
  - Can predict any variable's value, not just the class,
    more than one variable's value at a time.
    - Any number of items in the rule body and head.
  - Hence: far **more association rules in numbers** than classification rules

TID	Products
1	A, B, E
2	B, D
3	B, C
4	A, B, D
5	A, C
6	B, C
7	A, C
8	A, B, C, E
9	A, B, C

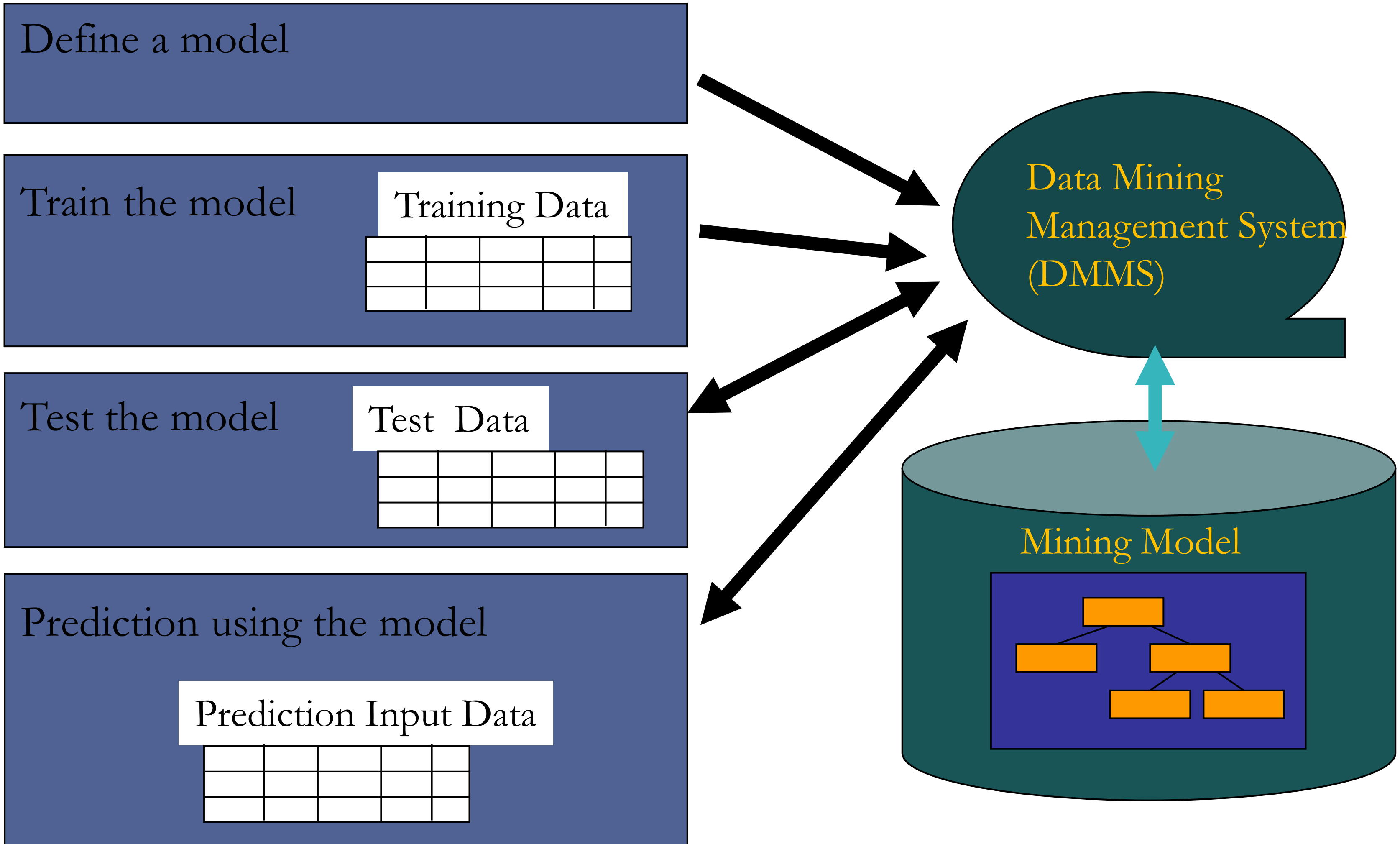
TID	A	B	C	D	E
1	1	1	0	0	1
2	0	1	0	1	0
3	0	1	1	0	0
4	1	1	0	1	0
5	1	0	1	0	0
6	0	1	1	0	0
7	1	0	1	0	0
8	1	1	1	0	1
9	1	1	1	0	0

Supervised learning: A 3-step process

Avoid Overfitting or Poor Generalisation

Bias and Variance

A Typical Predictive DM Process



Classification: 3-Step Process

1. Model Construction /Training/Learning:

- **Each instance is assumed to belong to a predefined class, called the target or class label**
- **Training set: A set of all records used for the construction of the model**
- **The model is usually represented in the form of classification rules, (IF-THEN statements) or decision trees or neural networks**

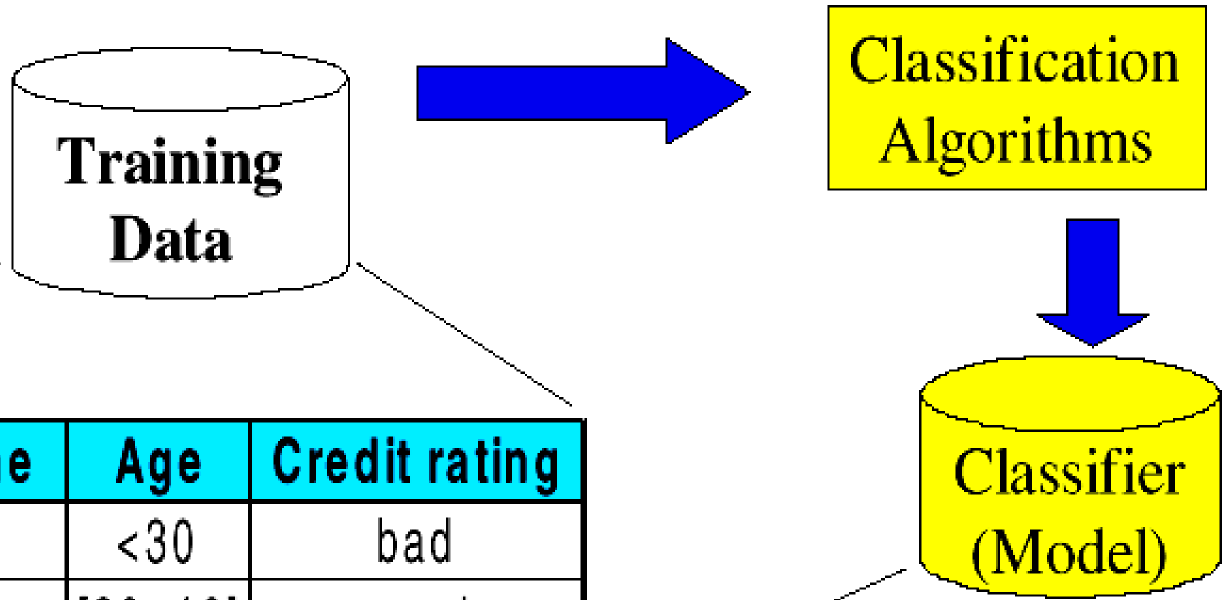
2. Model Evaluation/Testing:

- **Estimate Accuracy rate of the model based on a test/ validation set**
- **Accuracy rate:** the percentage of test set samples that are correctly classified by the model

3. Model Use:

- **The model is used to classify unseen instances, i.e. assign the class labels**

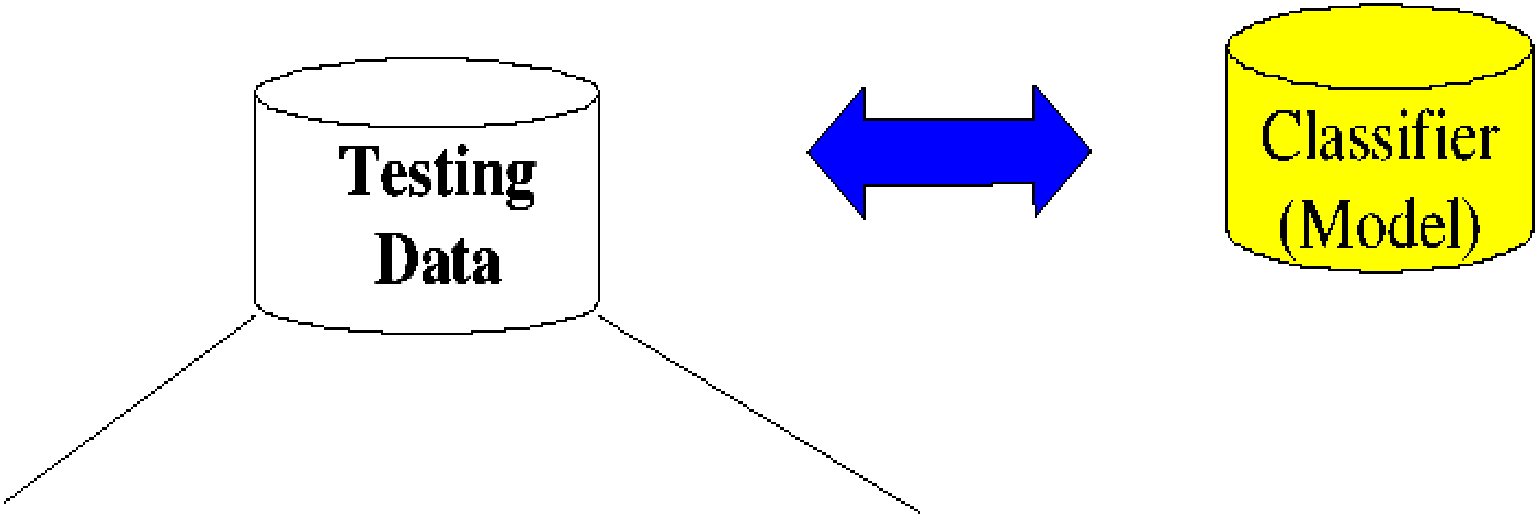
Training: Model Construction



Name	Income	Age	Credit rating
Bruce	Low	<30	bad
Dave	Medium	[30..40]	good
William	High	<30	good
Marie	Medium	>40	good
Anne	Low	[30..40]	good
Chris	Medium	<30	bad

IF Income = 'High'
OR Age > 30
THEN CreditRating = 'Good'

Testing: Model Evaluation



Name	Income	Age	Credit rating
Tom	Medium	<30	bad
Jane	High	<30	bad
Wei	High	>40	good
Hua	Medium	[30..40]	good

How accurate is the model?

IF Income = 'High'
OR Age > 30
THEN CreditRating = 'Good'

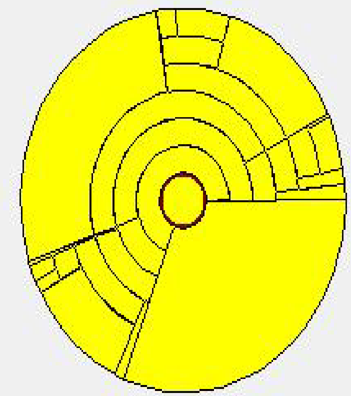
Choosing the best model

- Several models are built for 1 classification task.
  - Using several subsets of the dataset
  - Using several sets of features
  - Using several algorithms
    - Several parameters
- Which one to use for predictions?

  Estimate the Accuracy of the model
  - The known label of the test sample is compared with the classified result from the model
  - **Accuracy rate** is the percentage of test set samples that are correctly classified by the model
  - **Test set should be independent of Training set**
    - Helps to identify **over-fitting.**

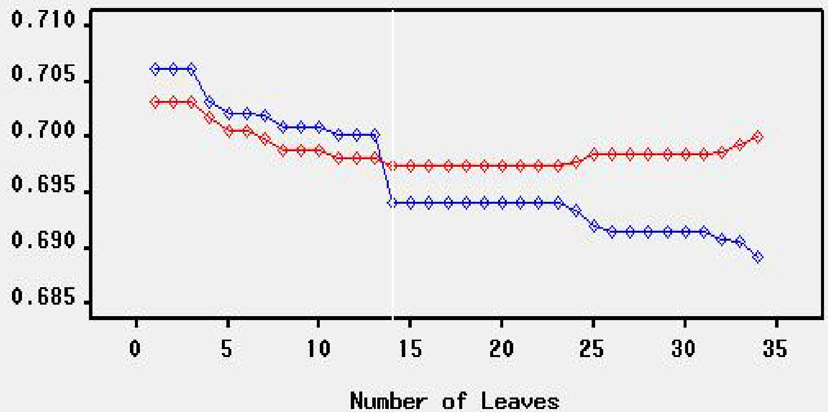
Overfitting: An Example

SOURCE	STAT	CRASH_COUNT_TARGET	==> 1	==> 2	==> 3	==> 4	==> 5	==> 6	==> 7
TRAIN	N	1	1713	52	164	24	0	11	0
TRAIN	N	2	877	64	181	36	0	13	0
TRAIN	N	3	559	46	204	32	0	8	0
TRAIN	N	4	426	21	143	37	0	11	0
TRAIN	N	5	267	21	112	20	0	12	0
TRAIN	N	6	212	14	82	33	0	14	0
TRAIN	N	7	175	17	95	25	0	4	0
TRAIN	N	8	154	8	40	16	0	2	0
TRAIN	N	9	75	0	55	11	0	0	0
TRAIN	N	10	76	0	66	11	0	0	0
TRAIN	N	11	42	3	23	14	0	0	0
TRAIN	N	12	37	0	26	24	0	0	0
TRAIN	N	13	21	0	7	0	0	0	0
TRAIN	N	14	9	0	10	11	0	5	0
TRAIN	N	15	25	13	12	0	0	0	0



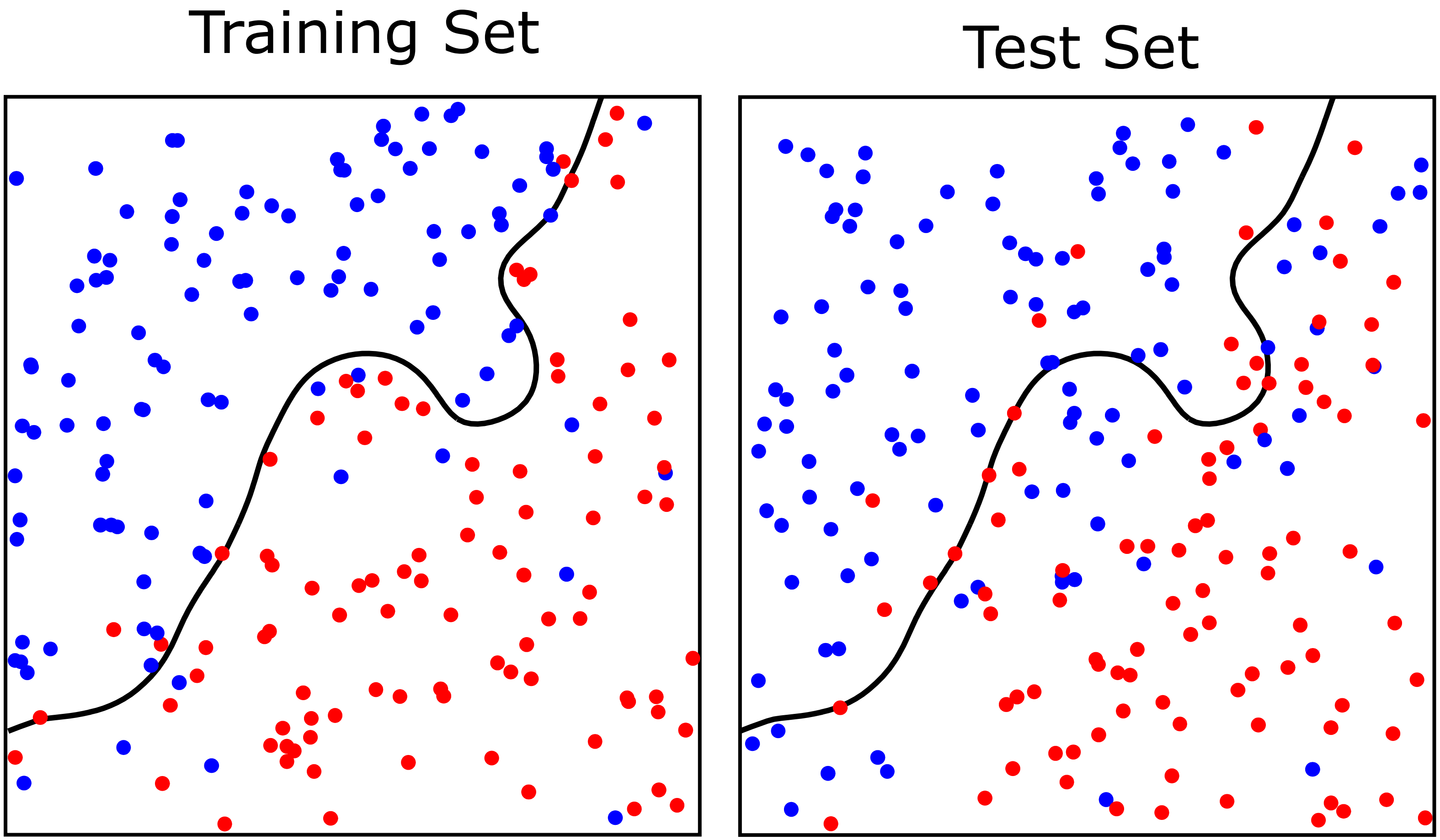
Blue Line; Training Error
Red Line: Test Error

Leaves	Training	Validation
14	0.6942	0.6975
15	0.6942	0.6975
16	0.6942	0.6975
17	0.6942	0.6975
18	0.6942	0.6975
19	0.6942	0.6975
20	0.6942	0.6975
21	0.6942	0.6975
22	0.6942	0.6975
23	0.6942	0.6975
24	0.6934	0.6977
25	0.6921	0.6985

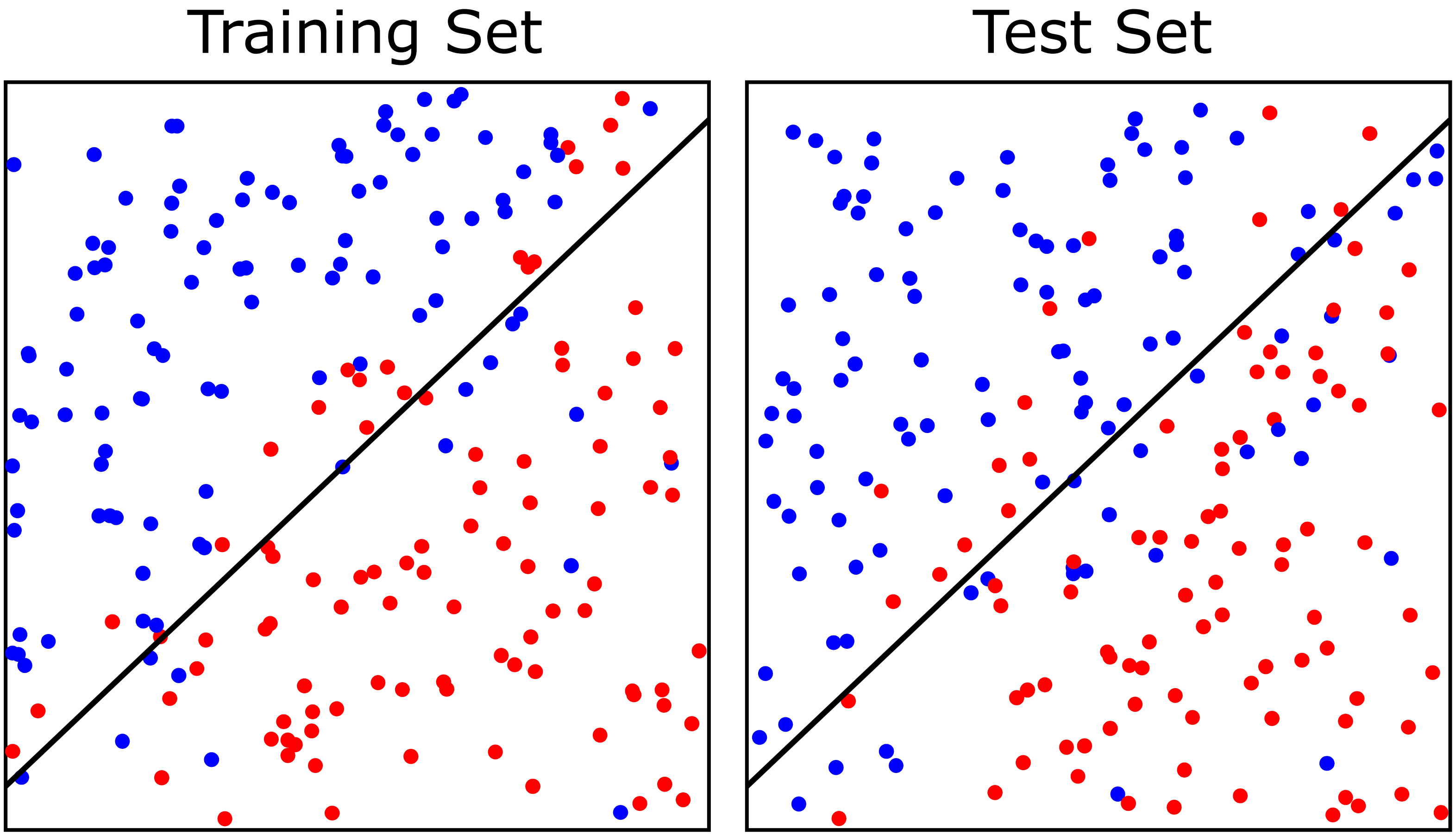


Log - (Untitled) | Editor - Untitled1 | Results - Tree | Assessment Colors

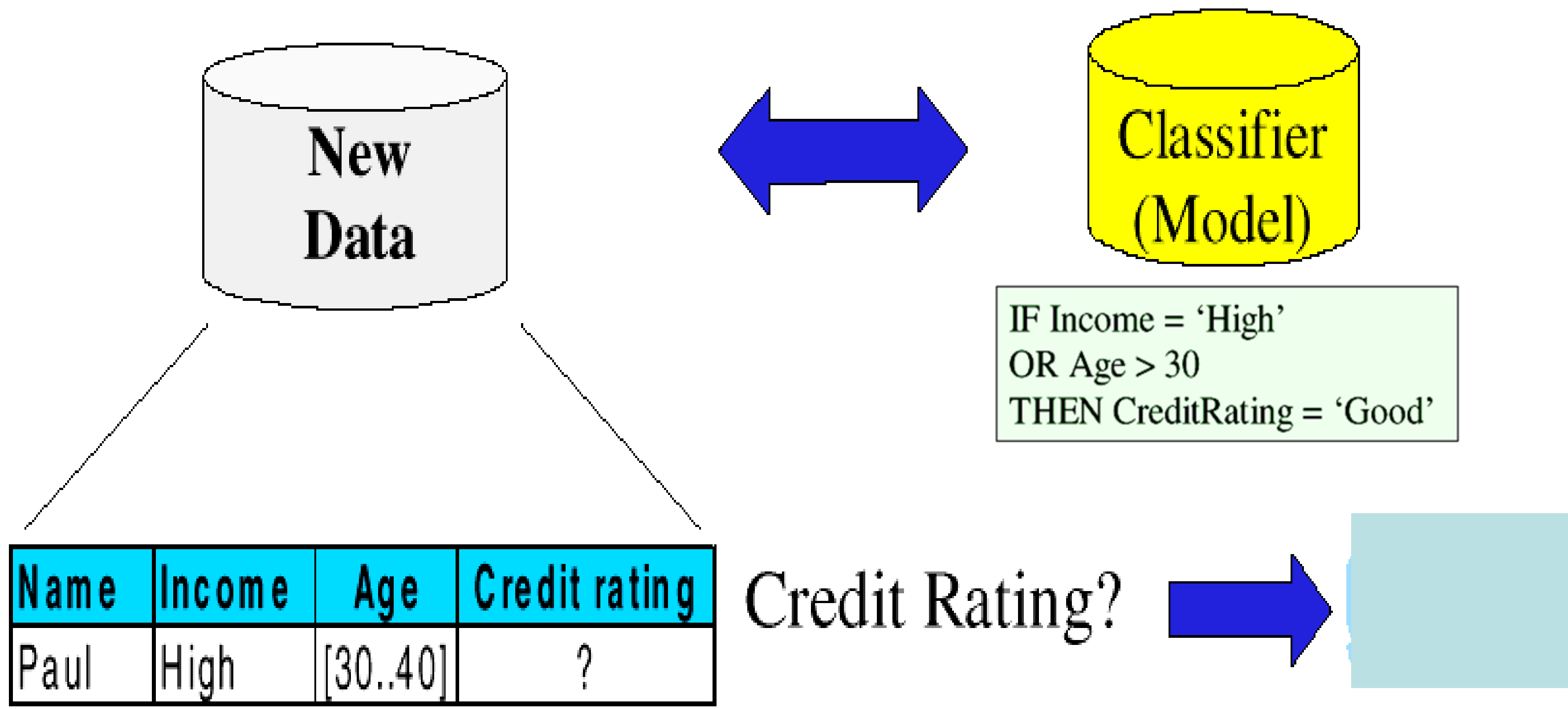
Overfitting



Approximate Fitting

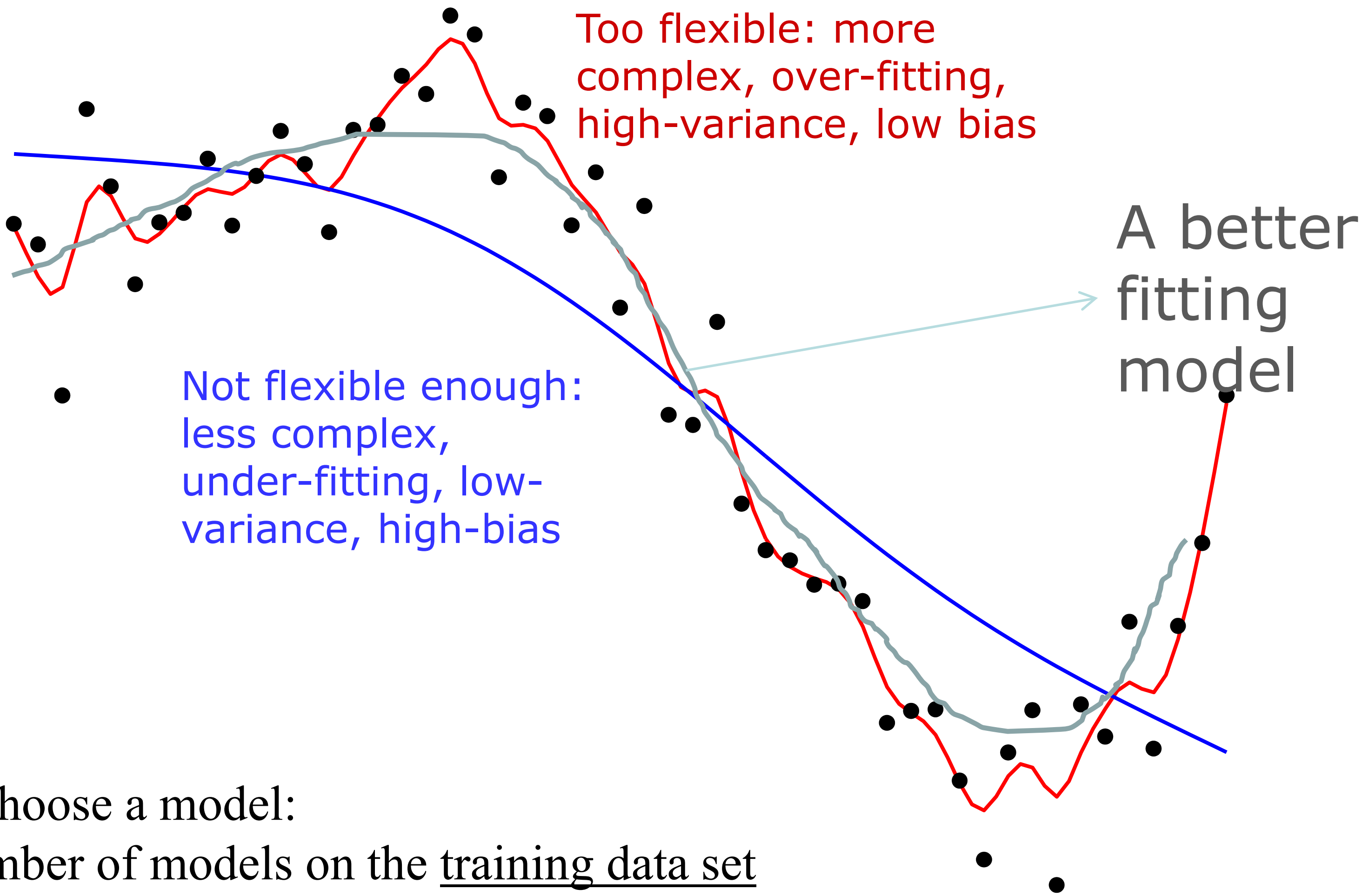


Model Use: Classification



Name	Income	Age	Credit rating
Paul	High	[30..40]	?

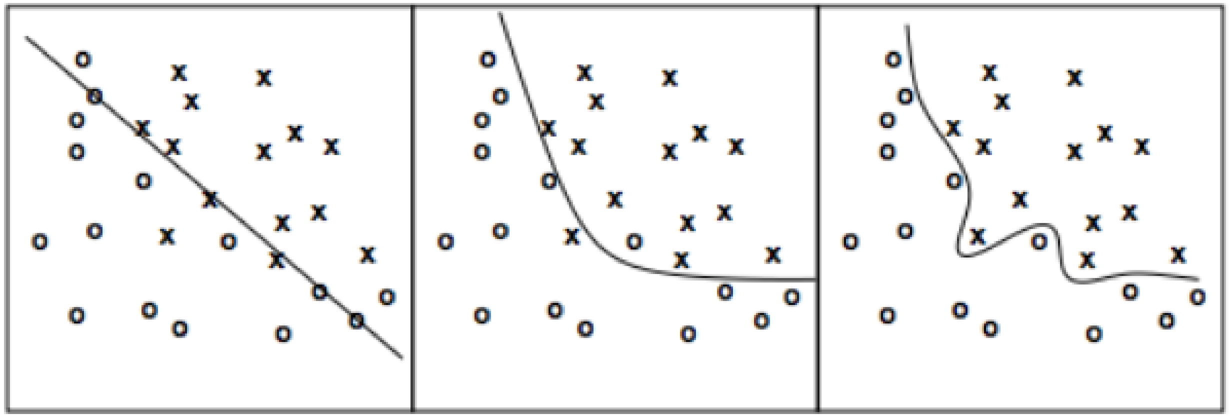
Model Complexity



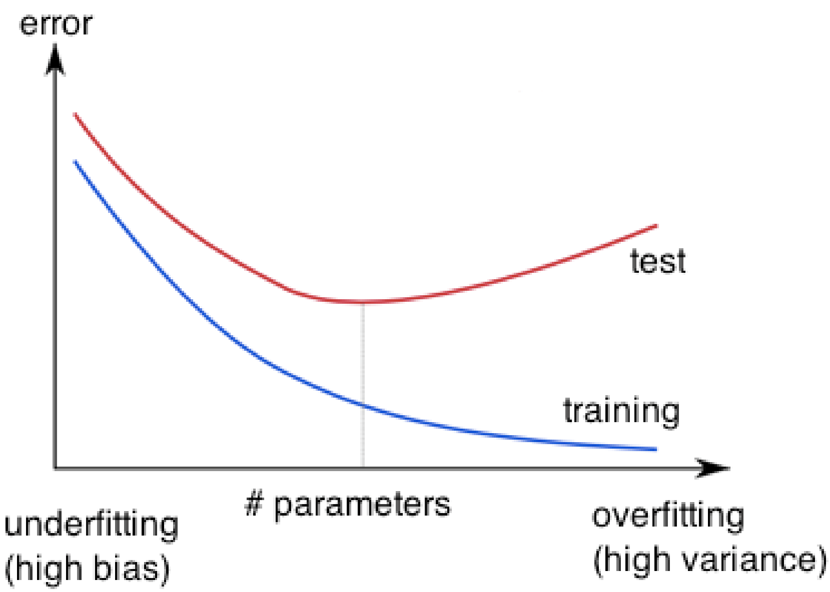
Strategy to choose a model:

- Build a number of models on the training data set
- Select the model that performs best on the validation data set
- Use the test data to estimate generalization

Overfitting: Summary



inadequate good compromise over-fitting



https://www.neuraldesigner.com/images/learning/selection_error.svg

http://wiki.bethanycrane.com/overfitting-of-data

<u>**Overfitting**</u>: Learned hypothesis may **fit** the training data very well, even outliers (**noise**) but fail to **generalize** to new examples (test data)

Bias is the simplifying assumptions made by the **model** to make the target function easier to approximate.

Variance is the amount that the estimate of the target function will change given different training data.

Classification Vs Regression

Two Types of Predictive Modeling

1. Classification (Predict categorical labels)

Task: Find a *model* for the class attribute as a function of the values of other attributes.

- Predict categorical labels
  - Event/no event (binary target)
  - Class label (multi-class problem – Nominal or Ordinal)

Example of class labels:

- Eligibility of clients for a loan (‘YES’ ‘NO)
- Course of treatment for a patient (‘A’ ‘B’ or ‘C’)
- Topic of a document (‘sport’ ‘politics’ ‘entertainment’)
- Rating of a product (‘good’ ‘average’ ‘bad’)
- Urgency of an email (‘urgent’ ‘non-urgent’)
- Anomaly of a transaction (‘anomalous’ ‘normal’)
- Personality of a person (‘introvert’ ‘extravert’ ‘both’)

Regresseion Modelling

2. Regression (or value prediction)

Task: Find a *model* for the continuous attribute as a function of the values of other attributes.

- Predict continuous labels
  - Age or Amount

Example:

- Approved Loan Amount ($$$$)
- Crash rate prediction
- House price prediction

Classification and Regression Prediction

- Classification learning: class is **Binary** or **Nominal** or **Ordinal**
-

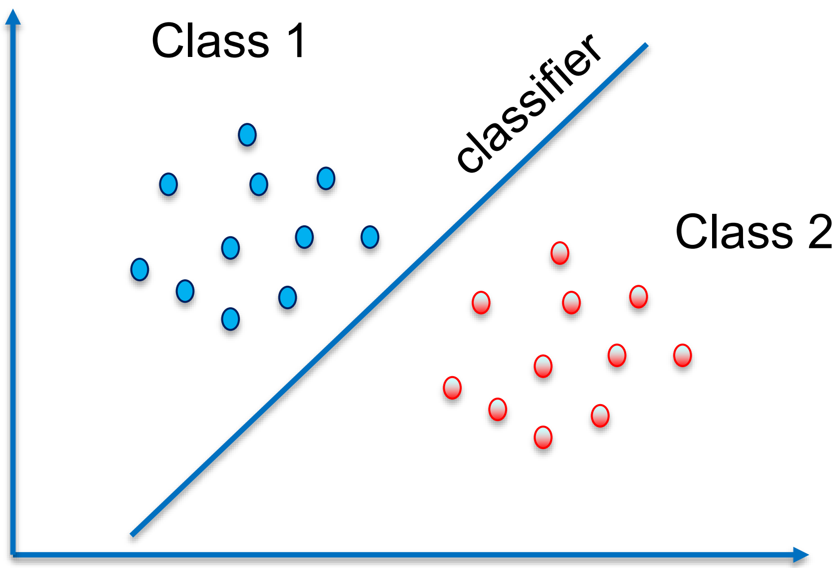
Outlook	Temperature	Humidity	Windy	**Play-time**
Sunny	Hot	High	False	**YES**
Sunny	Hot	High	True	**NO**
Overcast	Hot	High	False	**YES**

- Regression learning: class is **Numeric (Interval)**
  - Each training sample is provided with a target value that is continuous.

Outlook	Temperature	Humidity	Windy	**Play-time**
Sunny	Hot	High	False	**5**
Sunny	Hot	High	True	**0**
Overcast	Hot	High	False	**55**

Various Classification Algorithms Introduction

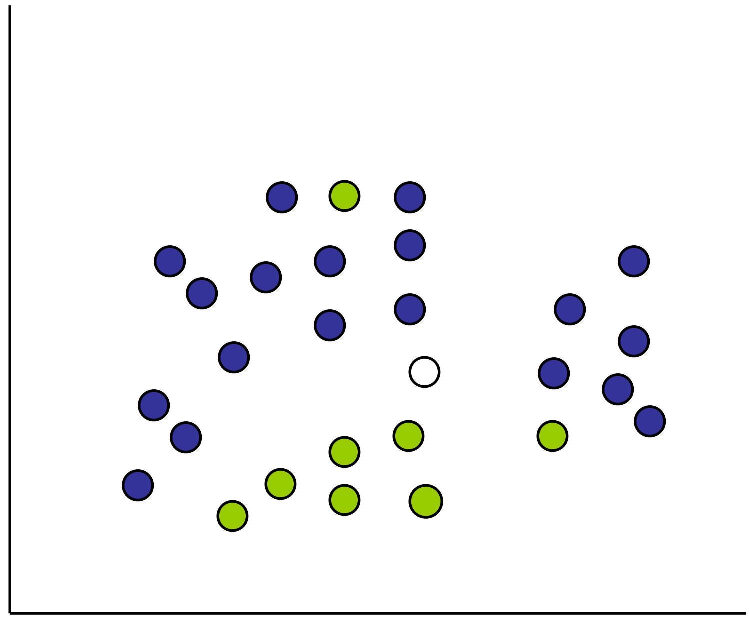
Classification learning



Data set: $\{(x_1,y_1),\ldots,(x_n,y_n)\}$, $(x,y) \sim \mathcal{D}$
x: feature vectors
y: binary label in {-1, +1} representing which class x belongs to
Learning: train classifier f(x) on data
Prediction: use f(x) to predict the label for arbitrary x

Classification: An Example

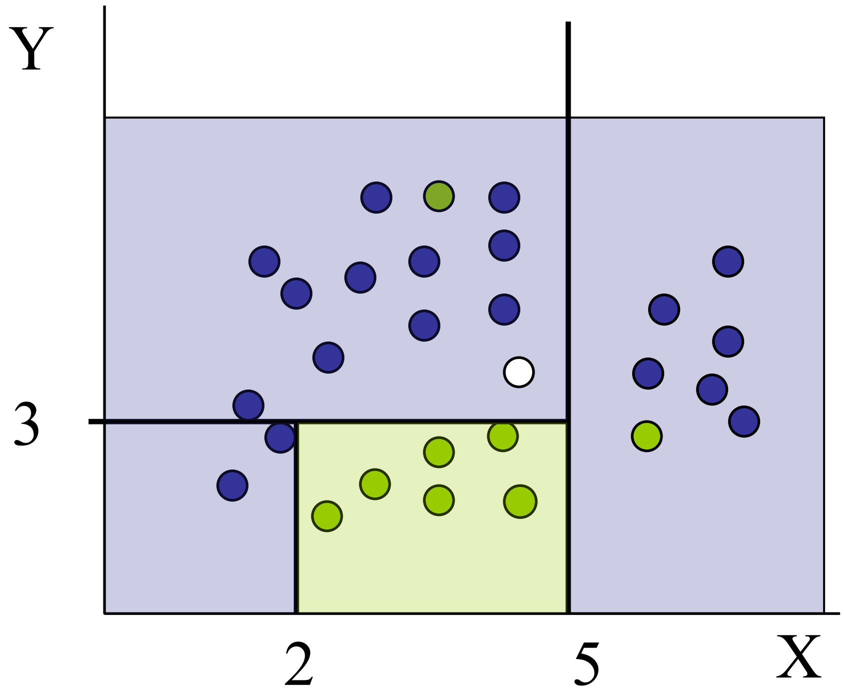
Learn a model for predicting the instance class from the pre-labelled instances



Each point is a multi-dimensional instance that includes several attributes.

Given a set of points from classes Blue ● and Green ●
What is the class of new point ○?

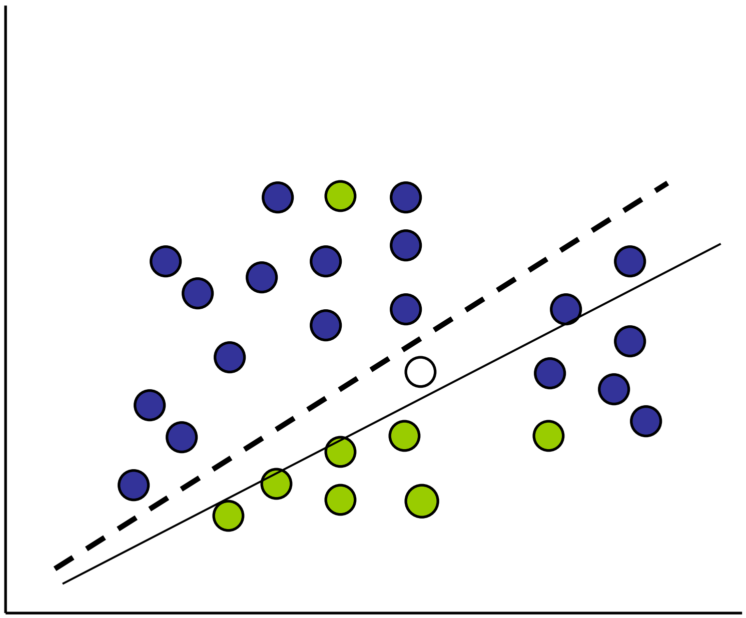
Classification: Decision Trees



if X > 5 then Blue
else if Y > 3 then Blue
else if X > 2 then Green
else Blue

- Piecewise constant approximation of decision regions
- Symbolic if-then rules
- Linear/non-linear, continuous/categorical model of decision regions

Classification: Logistic Regression

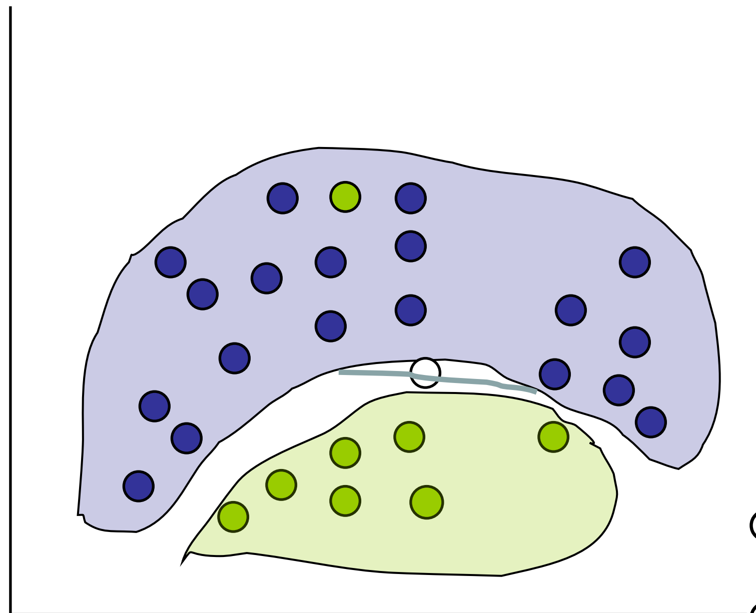


- Linear Regression

  $w_0 + w_1 x + w_2 y >= 0$
- Regression computes w_i from data to minimize squared error to 'fit' the data
- Find the "best" line (linear function $y=f(X)$) to explain the data.
- Not flexible enough

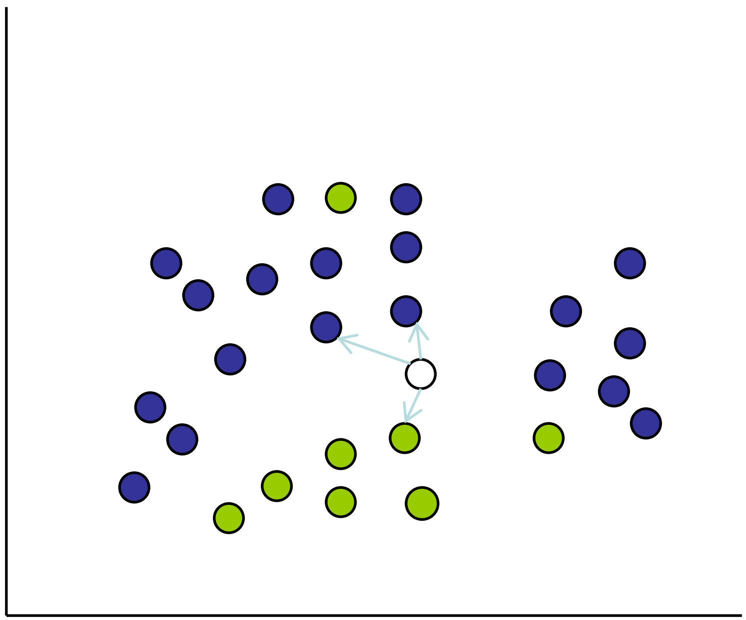
Classification: Neural Nets

- Linear/non-linear, continuous/categorical model of decision regions
- A number of parameters such as a set of weight matrices

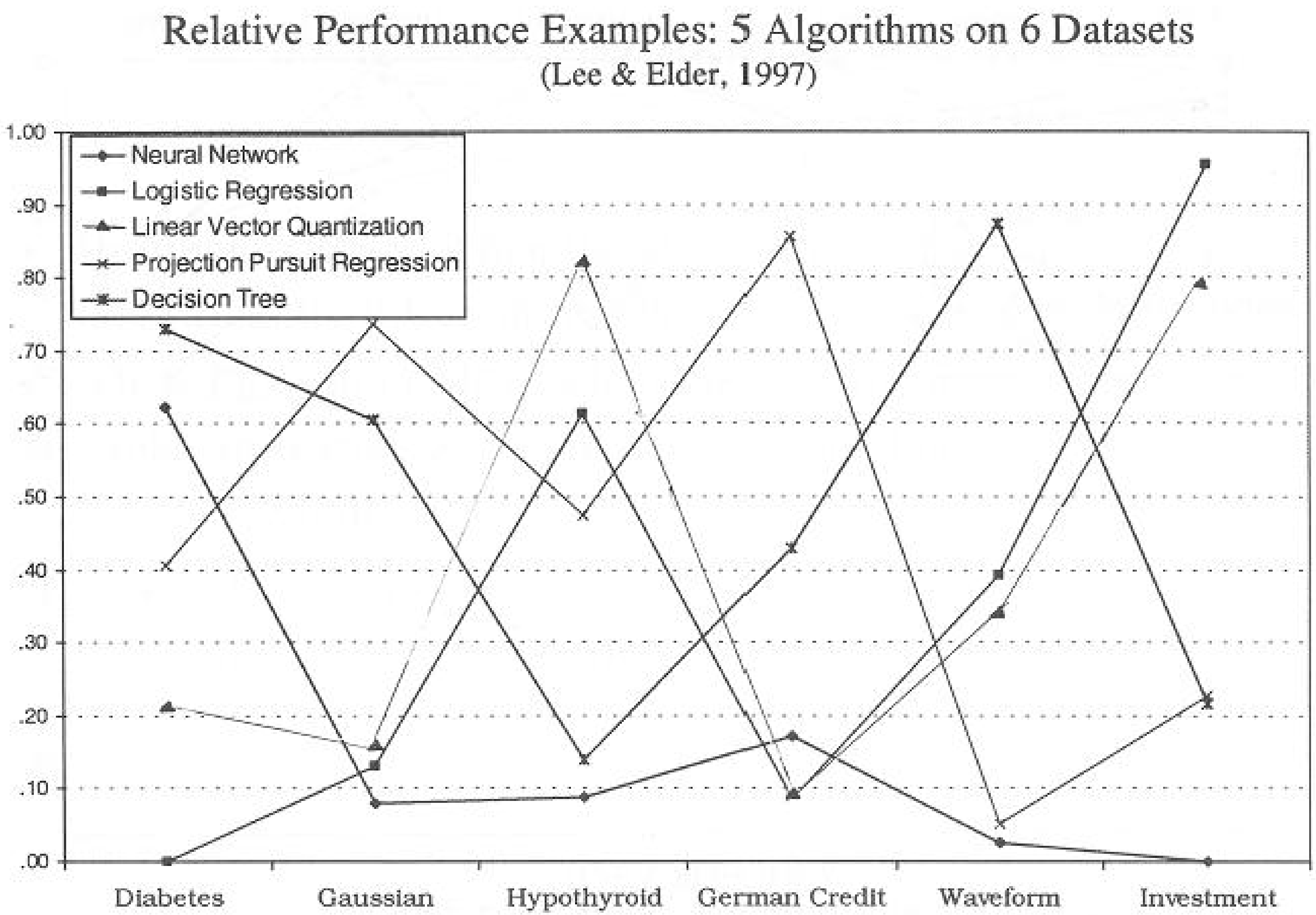


- Can select more complex regions
- Can be more accurate
- Can overfit the data – find patterns in random noise

Classification: Nearest Neighbor



- Does not make a model
- Learns localised decision regions from data
- A metric space based on proximity – calculates the distance between the query point and data points
- Chooses nearest neighbors and makes decisions based on neighbors' outcome
- Sensitive to data errors



Relative Performance Examples: 5 Algorithms on 6 Datasets
(Lee & Elder, 1997)
Neural Network
Logistic Regression
Linear Vector Quantization
Projection Pursuit Regression
Decision Tree
1.00
.90
.80
.70
.60
.50
.40
.30
.20
.10
.00
Diabetes
Gaussian
Hypothyroid
German Credit
Waveform
Investment

Various Classification Algorithms

Comparison: Evaluation Measures

Comparing Classification Algorithms

- Model goodness:
  - Predictive Accuracy: Ability of the model to correctly predict the class label of new data
  - RMSE: Root Mean Square error
  - AUC: **A**rea **U**nder (ROC) **C**urve
- Speed
  - Computation cost involved in generating and using the model
- Robustness
  - Ability of the model to make a correct prediction in the presence of noise and errors in the data
- Scalability
  - Ability to construct the model efficiently with the large amounts of data
- Interpretability
  - Level of understanding and insight provided by the model

Confusion Matrix

		Predicted class	
		Yes	No
Actual class	Yes	TP: True positive	FN: False negative
	No	FP: False positive	TN: True negative

- Machine Learning methods aim to minimize FP+FN
- TPR (True Positive Rate): TP / (TP + FN)
- FPR (False Positive Rate): FP / (TN + FP)

- A confusion matrix can also be generalized to multi-class.

Classification measures

- Precision: Proportion of all positive predictions by the model that are correct.
  - measures how many positive predictions are actual positive observations.

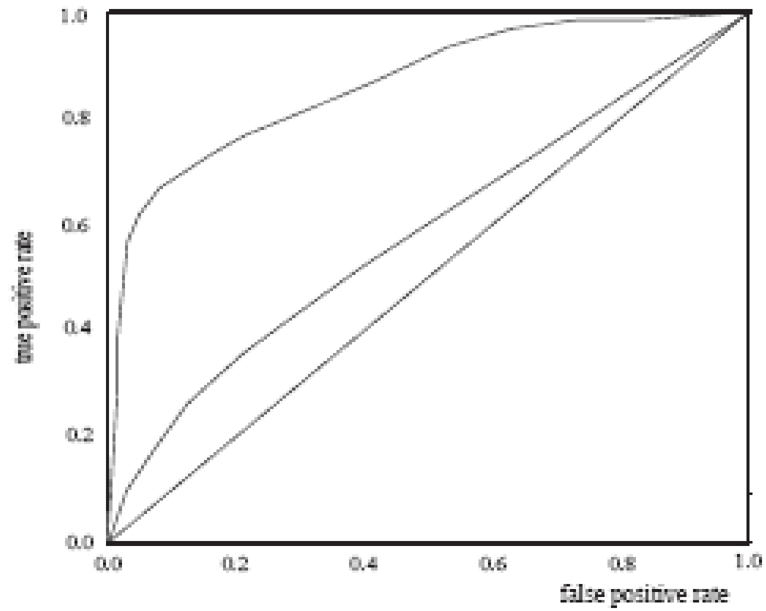
    Precision or Accuracy = TP/(TP+FP)
- Recall: Proportion of all real positive observations that are correct.
  - measures how many actual positive observations are predicted correctly.

    Recall or Coverage or Sensitivity = TP/(TP+FN) = TPR
- F1: The harmonic mean (average) of precision and recall.

    F-measure=(2×recall×precision)/(recall + precision)
- Specificity: Proportion of all negative predictions that are correct.

    Specificity = TN / (FP + TN) = 1 – FPR
- AUC (<u>A</u>rea <u>U</u>nder the ROC <u>C</u>urve)
  - measures how well predictions are ranked, rather than their absolute values.

AUC and ROC Curves



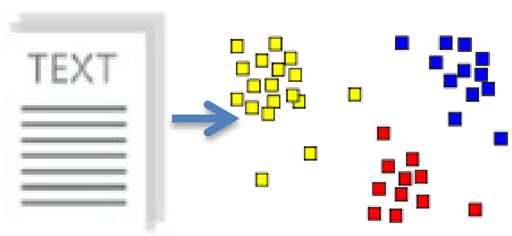
- ROC (Receiver Operating Characteristics) curves: for visual comparison of classification models
- The Area Under the ROC Curve (AUC) is a measure of the accuracy of the model
  - Shows the trade-off between the true positive rate and the false positive rate
  - true positive: Positive instances that are correctly classified as positive
  - false positive: Negative instances that are incorrectly classified as positive
- The closer to the diagonal line (i.e., AUC = approx. 0.5), the less accurate is the model
- A model with perfect accuracy will have an AUC of 1.0, which means the model predicts true positives 100% correctly.

Summary

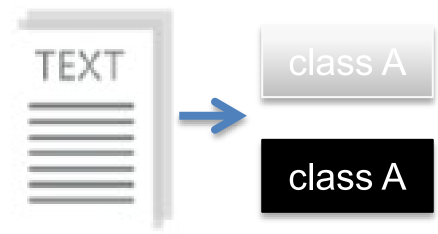
Recap: Types of Learning

Unsupervised: Discover **patterns** in **unlabeled** data
Example: *cluster* similar documents based on text

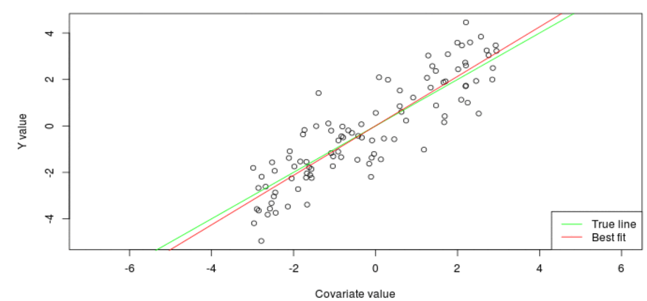
Supervised: Learning with a **labeled training** set
Example: (1) email *classification* with already labeled emails
(2) loan amount prediction (*regression*) using the historical data



Clustering



Classification



Regression

Reinforcement learning: learn to **act** based on **feedback/reward**: *win or lose*
Example: learn to play Go

http://mbjoseph.github.io/2018/11/27/measure.html

Final Remarks

- Predictive modelling is a supervised learning method
  - Due to its use of target attribute information.
  - Algorithms vary as how they use this target information
- Predictive Modelling includes three steps
  - Training; Testing; Classification
  - Training should avoid **overfitting**

References

- Data Mining techniques and concepts by Han J et al, 2011.
- Discovering Data Mining, by Cabena, et al., 1997.
- Predictive Data Mining, by Weiss and Indurkhya, 1999.