Tutorial 9

COMP90014 Algorithm for Bioinformatics

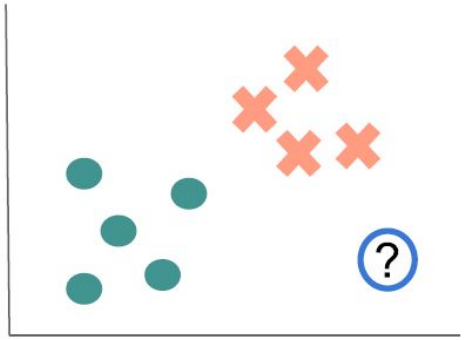
Semester 2, 2025

Supervised vs Unsupervised

Supervised

Infers a mapping function between inputs and outputs ***given*** labelled training data.

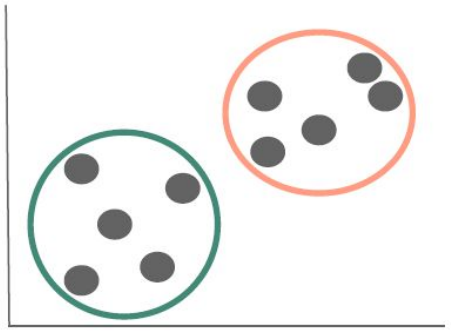
➢ Classify a new observation



Unsupervised

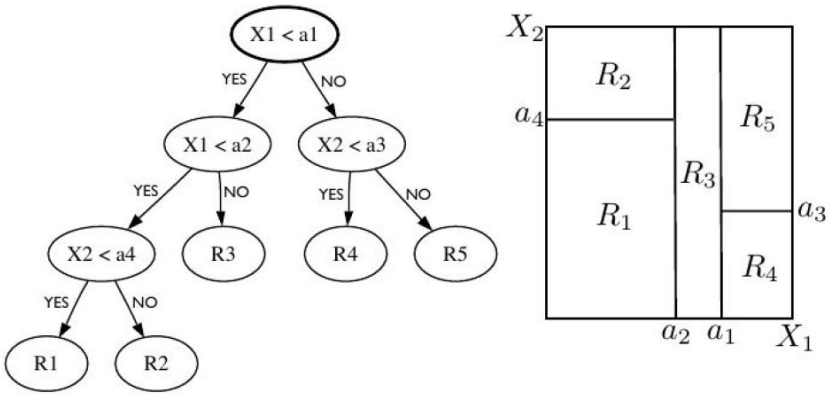
Finds implicit/hidden patterns in data ***without*** pre-existing labels.

➢ Identify clusters



Decision Trees

Recursive binary splitting is a greedy heuristic!



1. Choose a decision point yielding best purity
2. Partition data into corresponding subsets
3. Reiterate with resulting subsets
4. Stop when regions are approximately pure

Impurity in classification

- misclassification
- Gini impurity: probability of incorrectly classifying a randomly chosen data point

Impurity in regression

- mean squared error

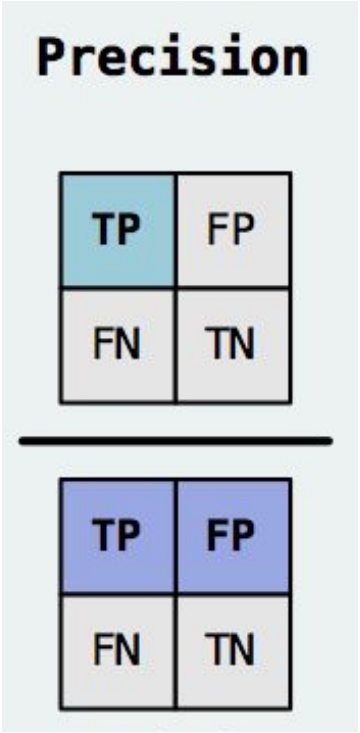
$$F(R) = \sum_{x_i \in R} (y_i - \langle y \rangle)^2$$

Precision

Spam filter (10 spam messages, 90 not spam)

	Spam	**Not spam**
Pred. spam	1 (TP)	0 (FP)
Pred. not-spam	9 (FN)	90 (TN)

- What's the **precision** for the spam class?
  - **100%**



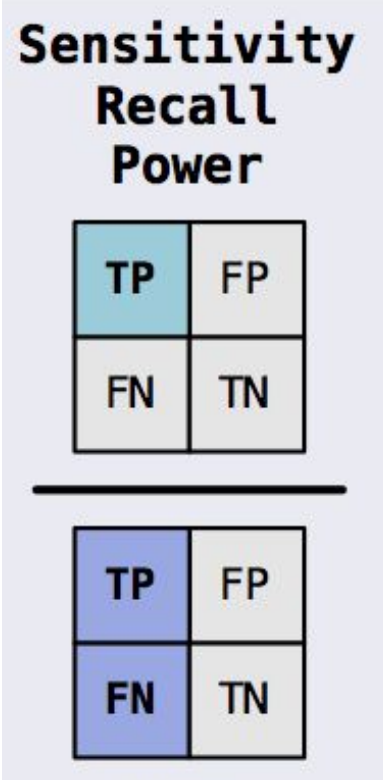
https://www.bluemontlabs.com/images/statistical-classification-metrics.png

Recall or Sensitivity

Spam filter (10 spam messages, 90 not spam)

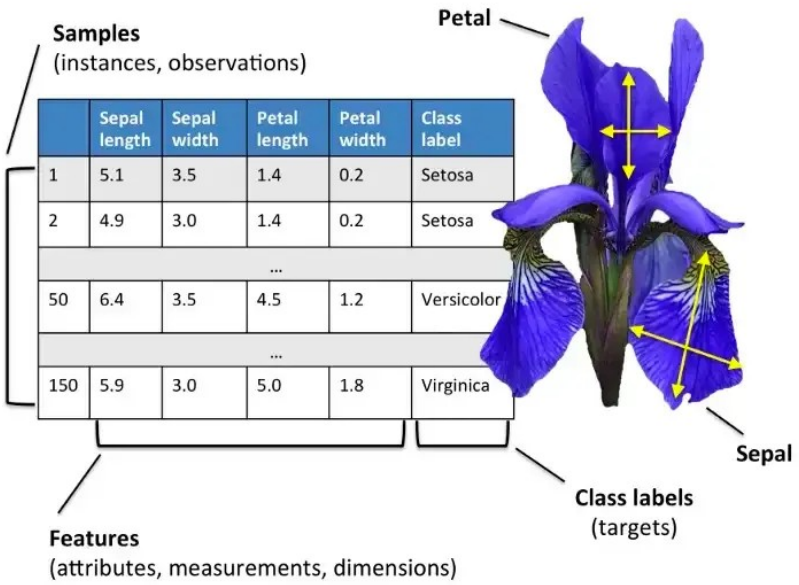
	Spam	**Not spam**
Pred. spam	10 (TP)	90 (FP)
Pred. not-spam	0 (FN)	0 (TN)

- What's the **recall** for the spam class?
  - **100%**



https://www.bluemontlabs.com/images/statistical-classification-metrics.png

Iris Dataset



	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

https://eminebozkus.medium.com/exploring-the-iris-flower-dataset-4e000bcc266c