LECTURER: TAI LE QUY

MACHINE LEARNING

SUPERVISED LEARNING

Introduction to Machine Learning	1
Regression	2
Basic Classification Techniques	3
Support Vector Machines	4
Decision & Regression Trees	5

UNIT 3

BASIC CLASSIFICATION TECHNIQUES

STUDY GOALS

- Understand the **concept of classification** and when to use it.
- Evaluate the **prediction performance** of a classification model.
- Apply **two** very popular classification **models** using **Python**.

EXPLAIN SIMPLY

1. Name three different distance measures?
2. Explain the use of the confusion matrix.
3. Explain what the "k" in k-Nearest Neighbors stands for.

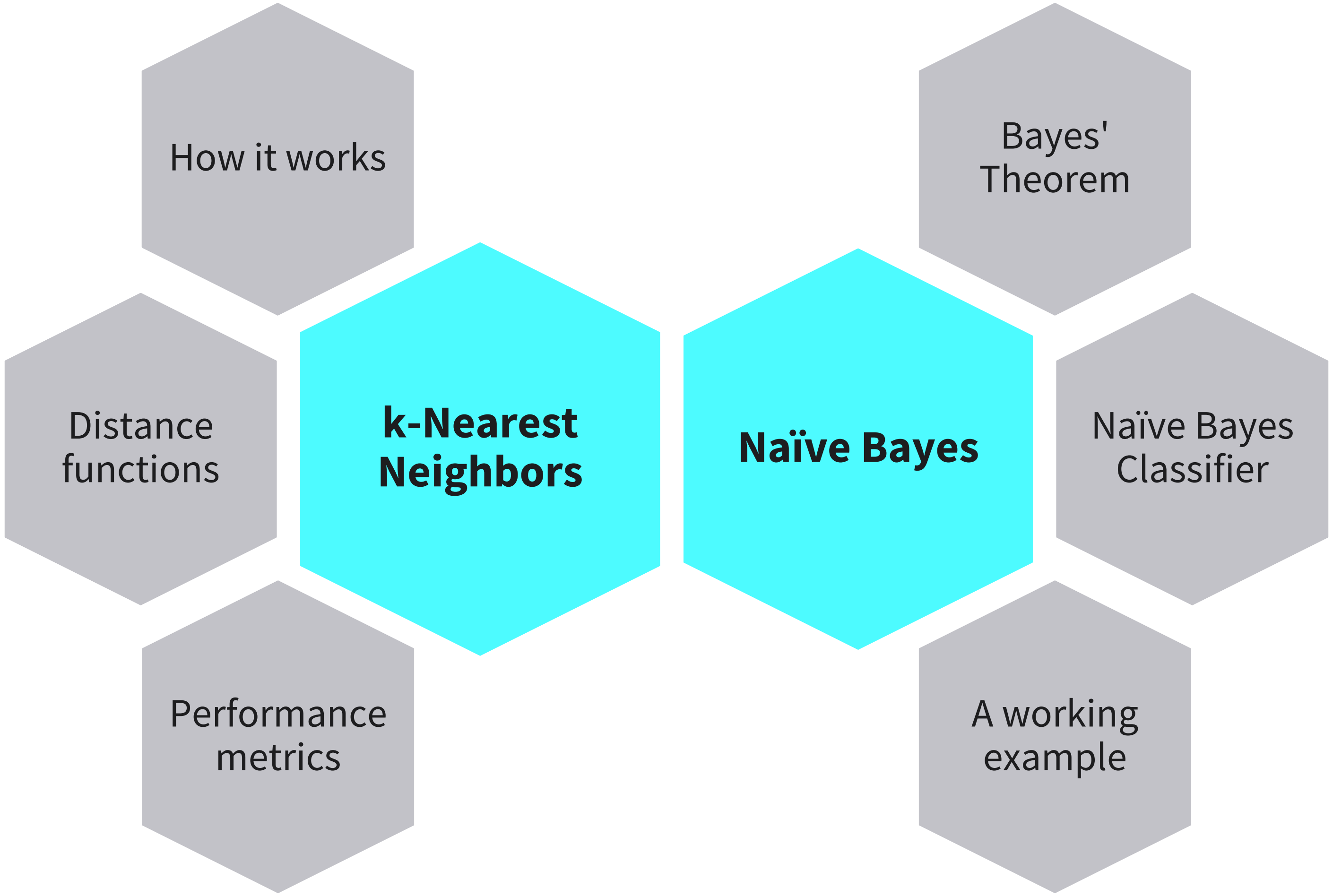
- The classification algorithms learn the characteristics of the classes provided in the training dataset and can then categorize a previously unseen observation by assigning a class label to it based on its feature values
- Binary and multi-class classifiers

Source of image: Course book

- Eager learners
  - Construct a classification model (based on training set)
  - Learn models are ready and eager to classify previously unseen instances
  - E.g., decision tress, SVMs, MNB, neural networks
- Lazy learners
  - Simply store training data (with labels) and wait until a new instance arrives that needs to be classified
  - No model is constructed
  - Known also as **instance-based learners**
  - E.g., kNN

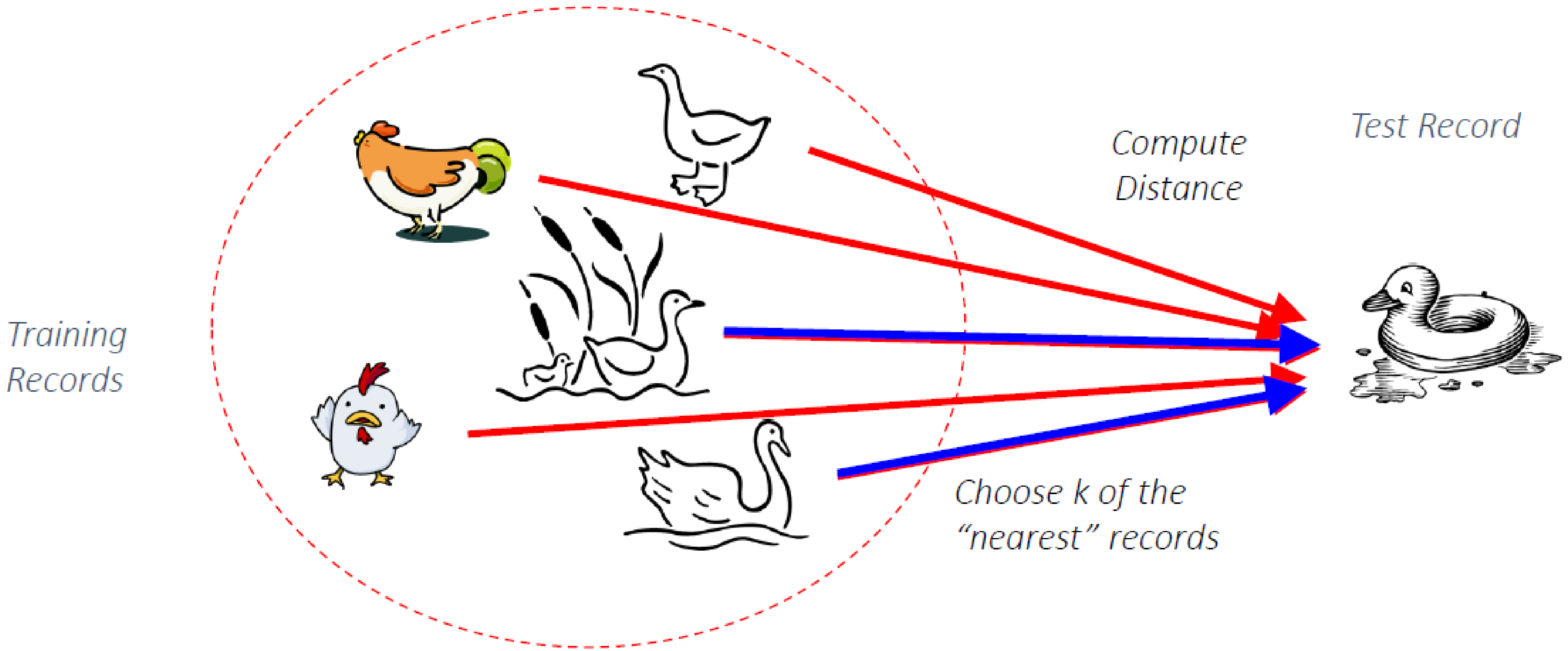
UNIT CONTENT

Img. 1: Basic classification techniques



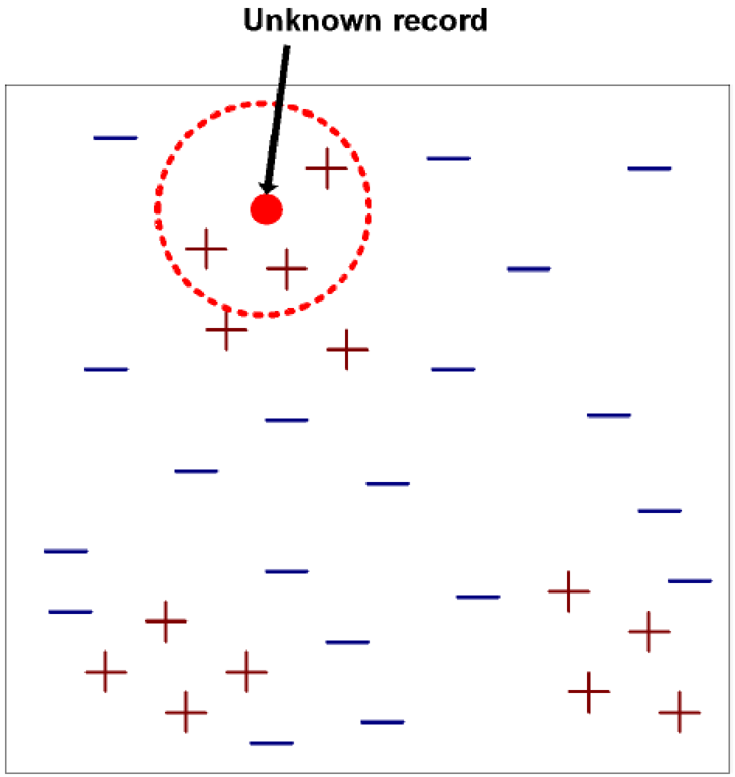
Source of image 1: Christian Müller-Kett, 2021.

- Basic idea
  - Basic idea: If it walks like a duck, quacks then it’s probably duck



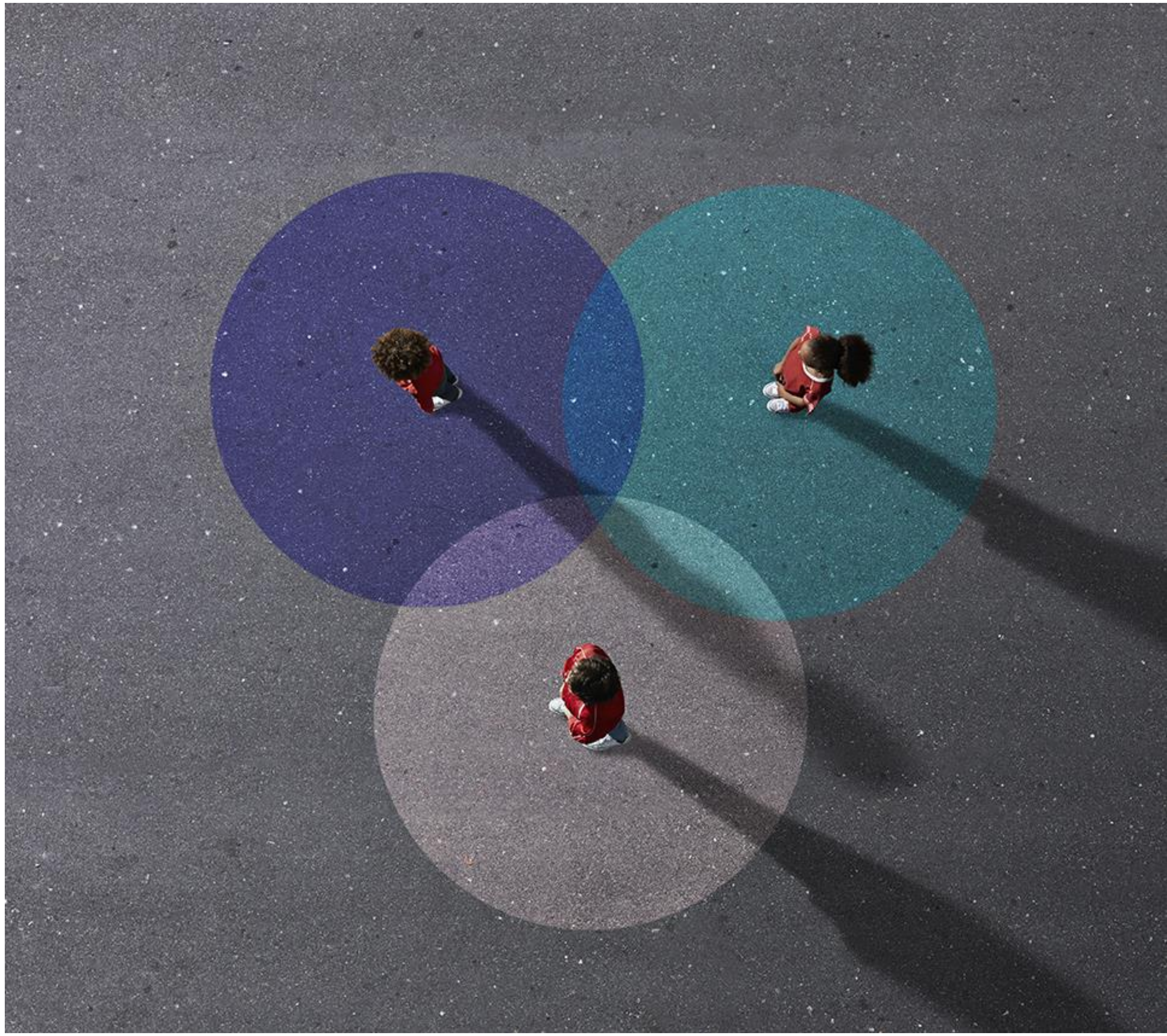
Source of image: Ntoutsi, 2021

- Input:
  - A training set D (with known class labels)
  - A **distance measure** to compute the distance between two instances
  - The **number of neighbor k**
- Classification:
  - Given a new unknown instance X
  - Compute distance to the training records
  - Identify the k nearest neighbors
  - Use the class labels of the k nearest neighbor to determine the class label of X (e.g., majority vote)
- Complexity: O(|D|) for each new instance



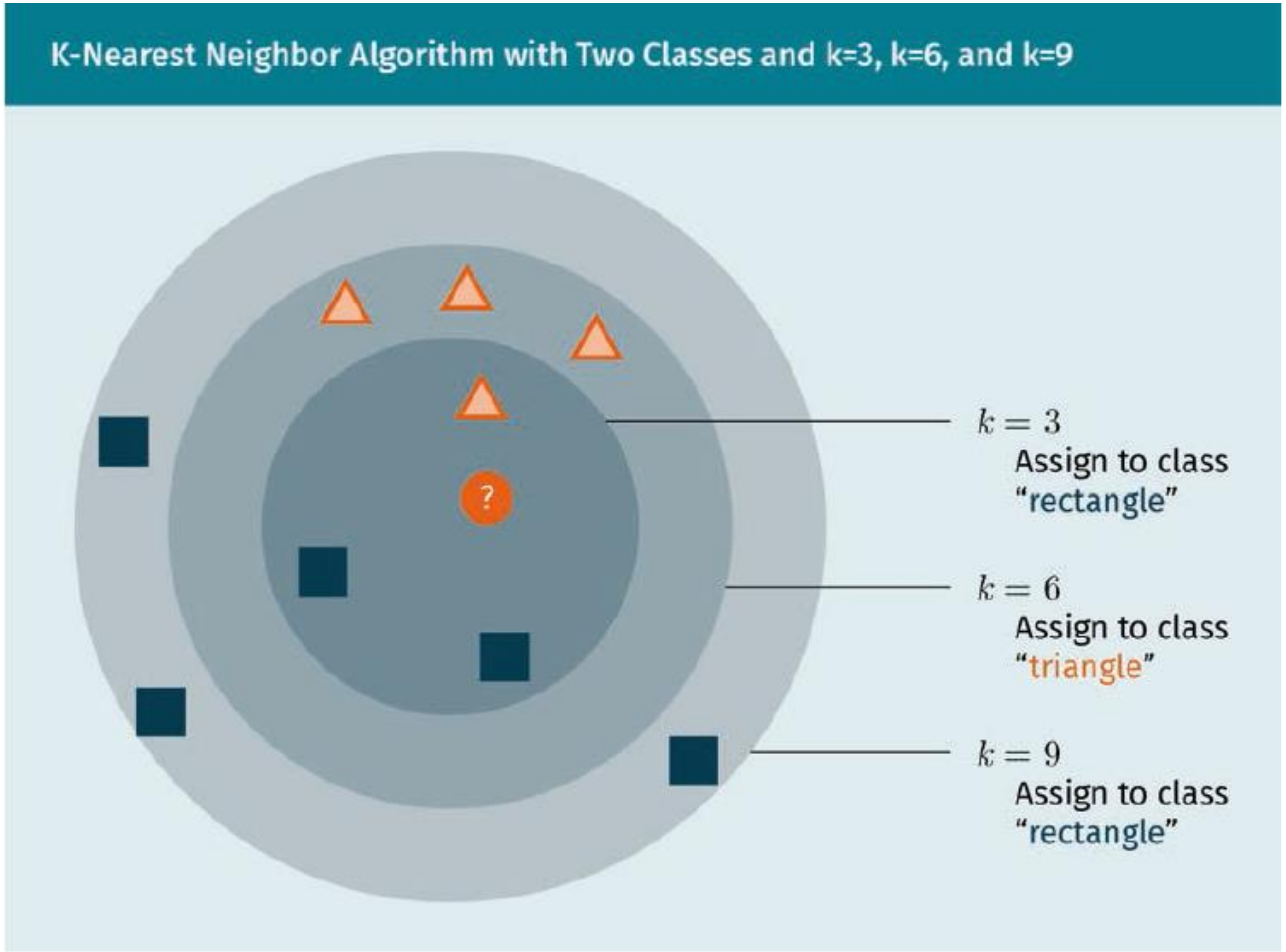
1. **Lazy learner**
2. Calculate **distances** to all samples
3. **Sort** distances in ascending order
4. Choose **k neighbors**
5. Assign the predicted class as the **neighbors' majority**

Img. 2: Neighbors



Source of image 2: Microsoft Archive

K-NEAREST NEIGHBOR (KNN)



Source of iamge: Course book

- Euclidean

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2}$$

- Manhattan

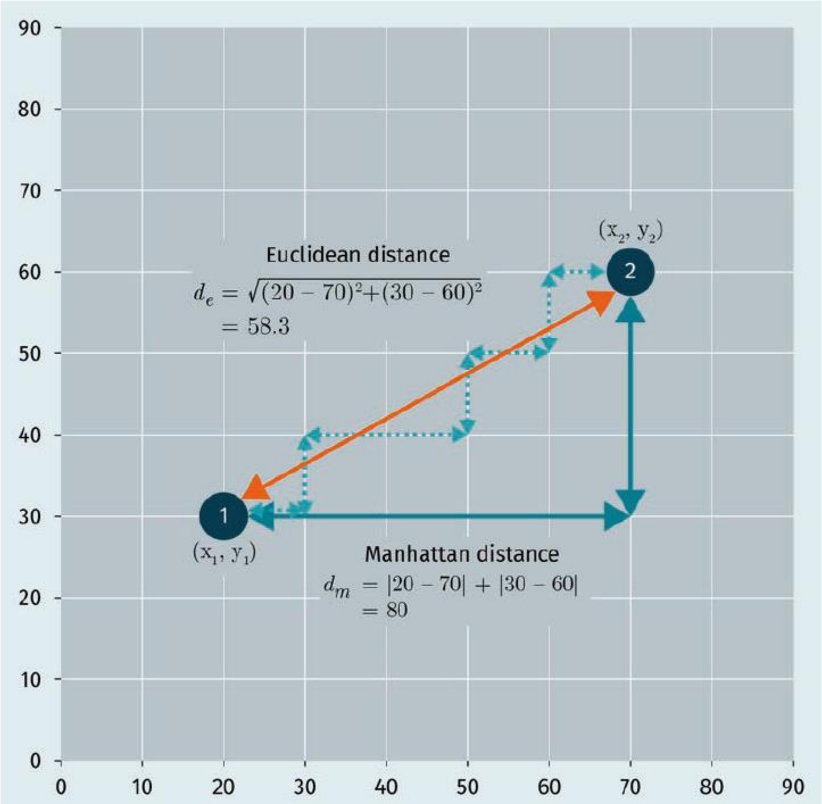
$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

- Cosine
  - Cosine similarity

$$s(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|} = \frac{\sum_{i=0}^{n-1} x_i y_i}{\sqrt{\sum_{i=0}^{n-1} (x_i)^2} \times \sqrt{\sum_{i=0}^{n-1} (y_i)^2}}$$

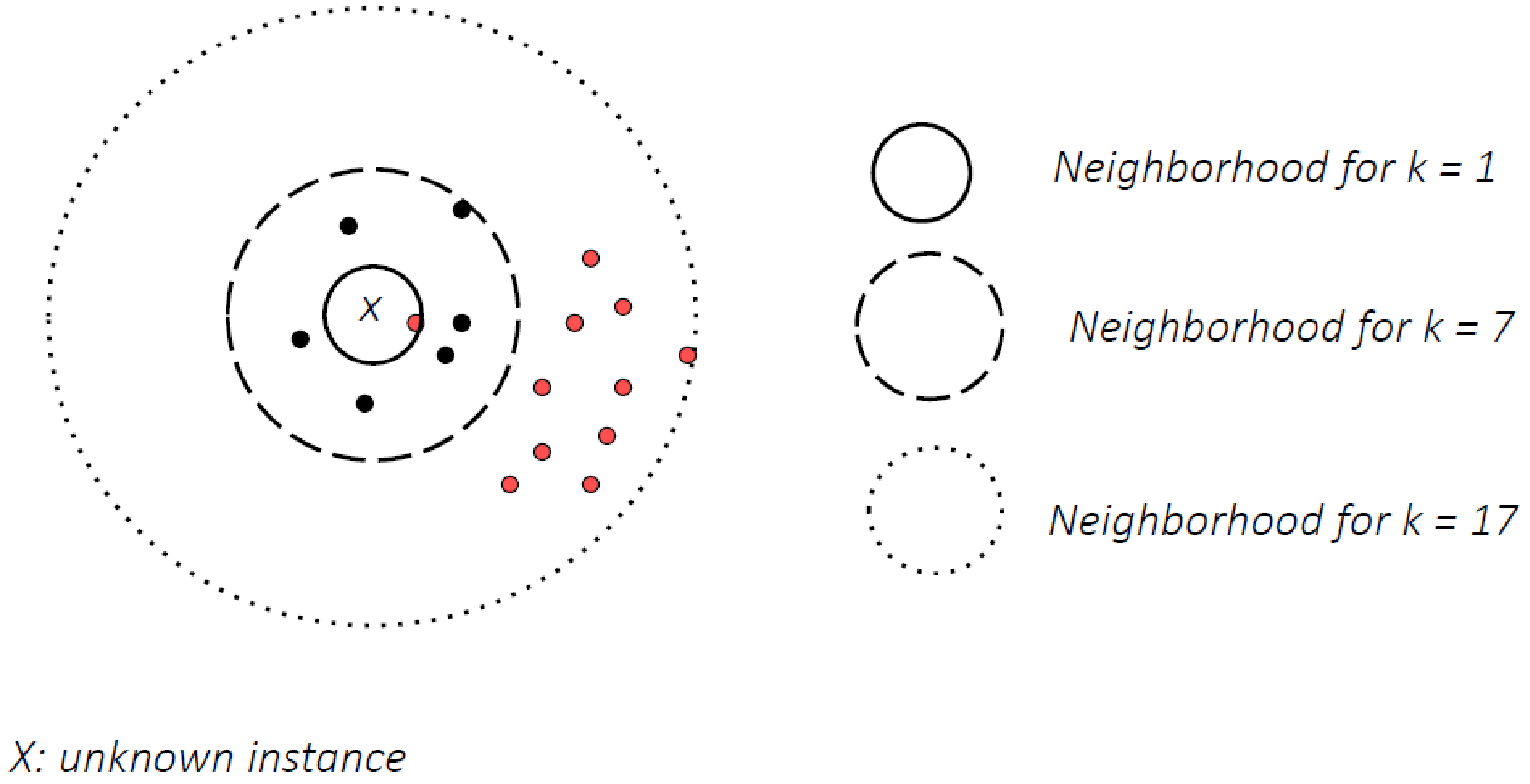
  - Cosine distance

$$d(\mathbf{x}, \mathbf{y}) = 1 - s(\mathbf{x}, \mathbf{y})$$



Source of figure: course book

CHOOSING A VALUE FOR K



- too small k: high sensitivity to outliers
- too large k: many objects from other classes in the resulting neighborhood
- average k: highest classification accuracy

Source of image: Ntoutsi, 2021

- Confusion matrix
- Measures:
  - Accuracy:

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN}$$

	Predicted classes		
Actual classes		Positive	Negative
	Positive	True positive (TP)	False negative (FN)
	Negative	False positive (FP)	True negative (TN)

  - Recall

$$\text{Recall}=\frac{TP}{TP+FN}$$

–$$\boldsymbol{Sensitivity} = \boldsymbol{Recall} = \frac{TP}{TP+FN}$$

–$$\boldsymbol{Specificity} = \frac{TN}{TN+FP}$$

  - Precision

$$\text{Precision}=\frac{TP}{TP+FP}$$

  - F1-score

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Classes are imbalanced:

e.g.,

- Consider a test set of 10000 instances:

9990 instances belong to class $c1$

10 instances belong to class $c2$

- Assuming a model M that predicts everything to be of class $c1$

Example:

Predicted class

Actual class

classes	buy_computer = yes	buy_computer = no	total
buy_computer = yes	6954	46	7000
buy_computer = no	412	2588	3000
total	7366	2634	10000

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN} \qquad \text{Precision}=\frac{TP}{TP+FP} \qquad \text{Recall}=\frac{TP}{TP+FN} \qquad F1=\frac{2\times Precision\times Recall}{Precision+Recall}$$

RECEIVER OPERATING CHARACTERISTIC CURVE (ROC CURVE)

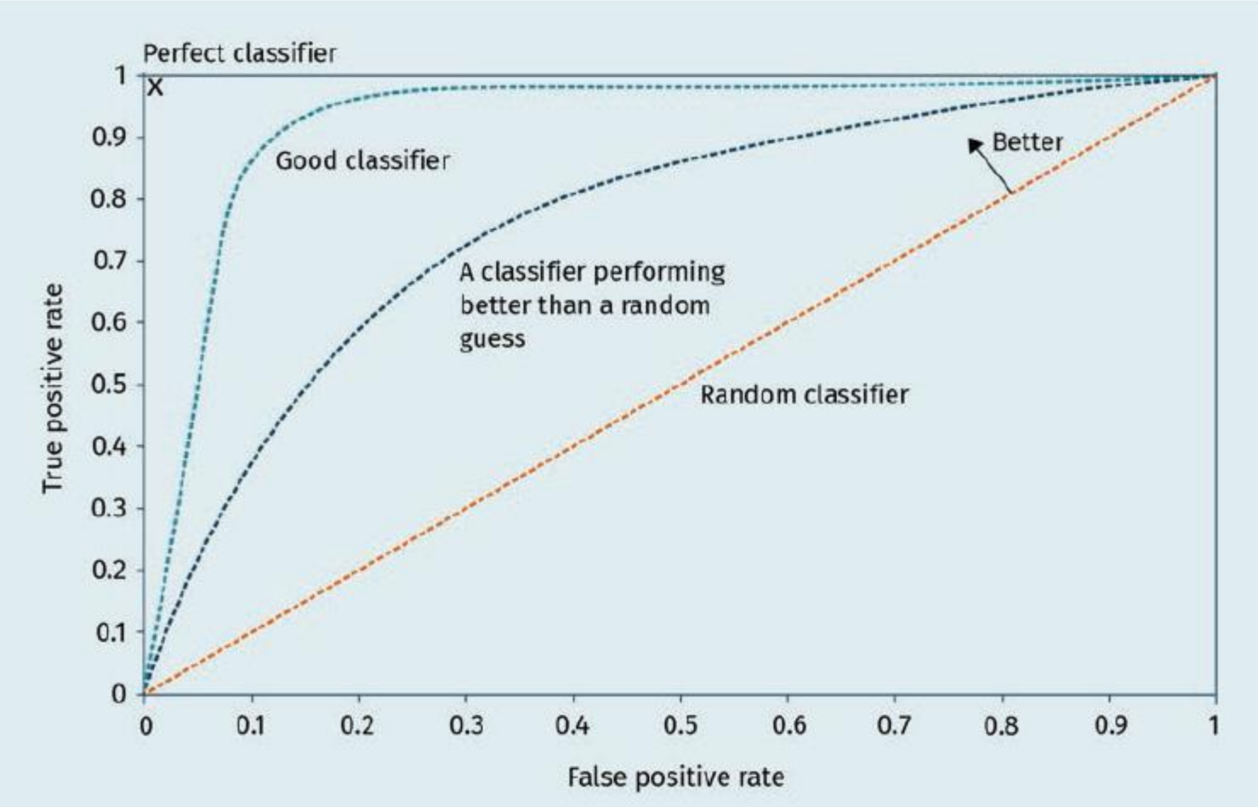
		Predicted classes	
		Positive	Negative
Actual classes	Positive	True positive (TP)	False negative (FN)
	Negative	False positive (FP)	True negative (TN)

- ROC curve (receiver operating characteristic curve): performance of a classification model at all classification thresholds.
- Two parameters:
  - True Positive Rate (recall)

$$TPR = \frac{TP}{TP + FN}$$

  - False Positive Rate

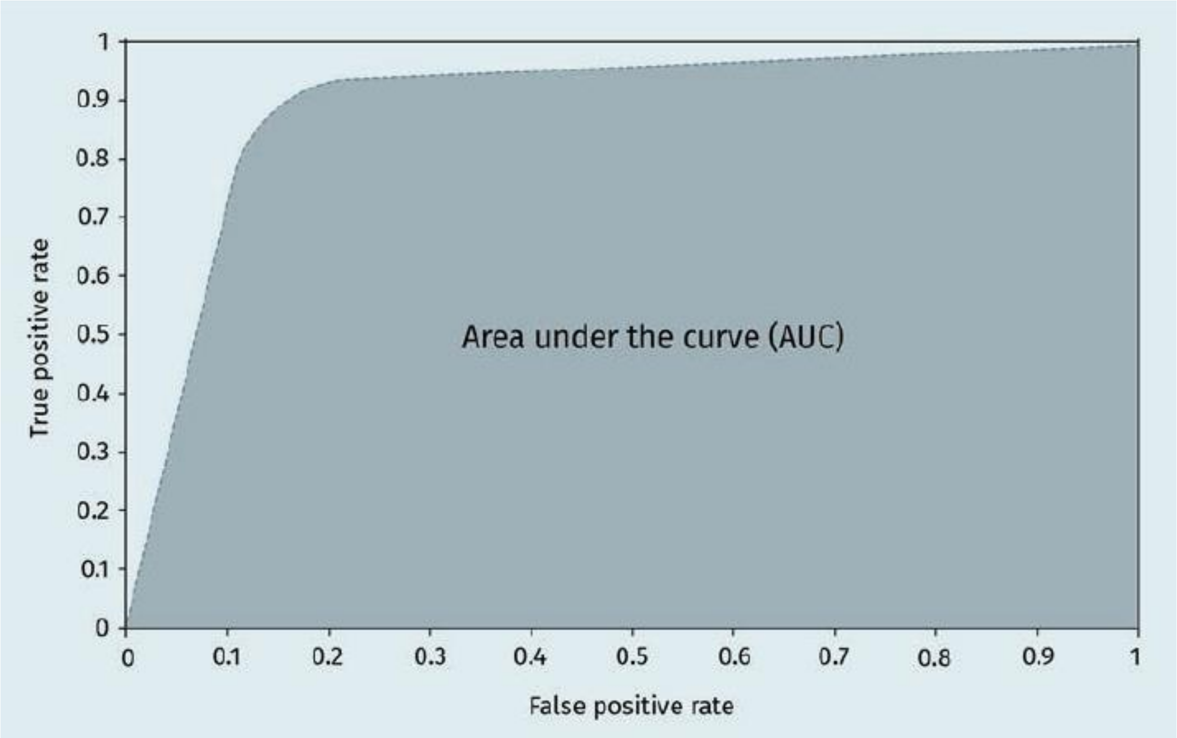
$$FPR = \frac{FP}{FP + FN}$$



Source of image: Course book

AREA UNDER THE CURVE (AUC)

- AUC measures the entire two-dimensional area underneath the entire ROC curve (think integral calculus) from (0,0) to (1,1).



- Bayesian classifiers is a probabilistic approach to inference and bases itself on “the assumption that the quantities of interest are governed by probability distributions and that optimal decisions can be made by reasoning about these probabilities together with observed data” (Mitchell, 1997, p. 154).
- Naïve Bayes (NB) classifier is based on **Bayes’ theorem**
- Assumption: the distributions of the features are independent of each other

- Probability of **observing something (A)**, ***given*** that **something else (B)** was observed
- We call this "**posterior**" probability
- We know the "**prior**" probability of B
- $p(A|B) = \frac{p(B|A) \cdot p(A)}{p(B)}$

Img. 4: Probability

Source of image 4: Microsoft Archive
Source of text: Mitchell, 1997

- Notations

  h $\in H$: a hypothesis

  D: training data

  P(h): initial probability that hypothesis h is true

  P(D): prior probability that the data D will be observed

  P(D|h): probability of observing the training data D given that hypothesis h is true

- Bayes theorem

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

- Naïve Bayes classifier attempts to find the most probable hypothesis h from the set of all possible hypotheses H given the training data D.
- The result is the maximum a posteriori hypothesis, i.e., the hypothesis with maximum probability

NAÏVE BAYES CLASSIFIER

Example

Frequency Table		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3

Likelihood Table		Play Golf		
		Yes	No	
Outlook	Sunny	3/9	2/5	5/14
	Overcast	4/9	0/5	4/14
	Rainy	2/9	3/5	5/14
		9/14	5/14	

$$P(x \mid c) = P(Sunny \mid Yes) = 3/9 = 0.33$$

$$P(x) = P(Sunny) = 5/14 = 0.36$$

$$P(c) = P(Yes) = 9/14 = 0.64$$

Posterior Probability: $P(c \mid x) = P(Yes \mid Sunny) = 0.33 \times 0.64 \div 0.36 = 0.60$

Frequency Table		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3

		Play Golf		
		Yes	No	
Outlook	Sunny	3	2	5
	Overcast	4	0	4
	Rainy	2	3	5
		9	5	14

$$P(x \mid c) = P(Sunny \mid No) = 2/5 = 0.4$$

$$P(x) = P(Sunny) = 5/14 = 0.36$$

$$P(c) = P(No) = 5/14 = 0.36$$

Posterior Probability: $P(c \mid x) = P(No \mid Sunny) = 0.40 \times 0.36 \div 0.36 = 0.40$

Likelihood

Class Prior Probability

$$P(c \mid x) = \frac{P(x \mid c)\,P(c)}{P(x)}$$

Posterior Probability

Predictor Prior Probability

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

Source of image: https://www.saedsayad.com/naive_bayesian.htm

NAÏVE BAYES CLASSIFIER

Frequency Table

		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3

		Play Golf	
		Yes	No
Humidity	High	3	4
	Normal	6	1

		Play Golf	
		Yes	No
Temp.	Hot	2	2
	Mild	4	2
	Cool	3	1

		Play Golf	
		Yes	No
Windy	False	6	2
	True	3	3

Likelihood Table

		Play Golf	
		Yes	No
Outlook	Sunny	3/9	2/5
	Overcast	4/9	0/5
	Rainy	2/9	3/5

		Play Golf	
		Yes	No
Humidity	High	3/9	4/5
	Normal	6/9	1/5

		Play Golf	
		Yes	No
Temp.	Hot	2/9	2/5
	Mild	4/9	2/5
	Cool	3/9	1/5

		Play Golf	
		Yes	No
Windy	False	6/9	2/5
	True	3/9	3/5

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

Outlook	Temp	Humidity	Windy	Play
Rainy	Cool	High	True	?

$$P(Yes \mid X) = P(Rainy \mid Yes) \times P(Cool \mid Yes) \times P(High \mid Yes) \times P(True \mid Yes) \times P(Yes)$$

$$P(Yes \mid X) = 2/9 \times 3/9 \times 3/9 \times 3/9 \times 9/14 = 0.00529 \qquad 0.2 = \frac{0.00529}{0.02057 + 0.00529}$$

$$P(No \mid X) = P(Rainy \mid No) \times P(Cool \mid No) \times P(High \mid No) \times P(True \mid No) \times P(No)$$

$$P(No \mid X) = 3/5 \times 1/5 \times 4/5 \times 3/5 \times 5/14 = 0.02057 \qquad 0.8 = \frac{0.02057}{0.02057 + 0.00529}$$

Source of image: https://www.saedsayad.com/naive_bayesian.htm

NAÏVE BAYES CLASSIFIER - APPLIED EXAMPLE

```
…

from sklearn.naive_bayes import GaussianNB

…

model = GaussianNB()

model.fit(training_data)

…

y_pred = model.predict(testing_data)
```

Img. 5: Code

Source of image 5: Microsoft Archive

REVIEW STUDY GOALS

- Understand the **concept of classification** and when to use it.
- Evaluate the **prediction performance** of a classification model.
- Apply **two** very popular classification **models** using **Python**.

SESSION 3

TRANSFER TASK

A start-up that sells **sustainable products in smaller stores** has been very successful in recent years. As a result, more stores are to be opened worldwide.

Also, the **number of products increases**. To **quickly assess** whether new products are **sustainable**, you and your team of Data Scientists are tasked with training a simple machine learning model. Given several product characteristics, this model should classify new products as sustainable or not.

You have a **couple of weeks to train** the model. Once the model is trained, **predictions should be obtained quickly** as these are integrated into the product order software to be used in real-time. Choose an appropriate algorithm and justify your choice.

Please present your results.

The results will be discussed in plenary.

1. What is the common name for the boundary learned by a classifier to separate the individual classes?
   a) cluster boundary
   b) classification boundary
   c) separation boundary
   d) decision boundary

2. What does the “k” stand for in the k-nearest neighbor algorithm?
   a) the number of kernels
   b) the number of neighbors
   c) the number of classes
   d) the number of learning cycles

3. Which one of the following is not a metric for a classification model?
   a) RMSE
   b) ROC curve
   c) confusion matrix
   d) precision

LIST OF SOURCES

Boehmke, B., & Greenwell, B. (2019). *Hands-on machine learning with R*. Chapman & Hall.

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill.