WeST People and Knowledge Networks | Institute for Web Science & Technologies

UNIVERSITÄT KOBLENZ · LANDAU

Machine Learning and Data Mining WS21/22

"6 Supervised Learning"

Dr. Zeyd Boukhers

@ZBoukhers

Institute for Web Science and Technologies

University of Koblenz-Landau

December 1, 2021

From previous lecture

- Evaluating clustering results
  - Dunn index
  - Rand index
  - Silhouette coefficient
  - Mutual information
- Clustering techniques
  - K-Means
  - Expectation Maximization
  - DBSCAN
  - Agglomerative Hierarchical Clustering
- Clustering for high-dimensional datasets
    - Dimensionality reduction
    - Superspace clustering

Agenda

- KNN
- K-D Tree
- Bayes Theorem
- Naive Bayes

k-Nearest Neighbours

k-Nearest Neighbours

- Given:
  - A classification problem with two classes ω_1 and ω_2,
  - The training feature vectors $\boldsymbol{x}_i, i = 1, \ldots, N_1$ labelled as ω_1 and the training feature vectors $\boldsymbol{x}_j, j = 1, \ldots, N_2$ labelled as ω_2, Note that $N = N_1 + N_2$,
  - An unknown feature vector $\boldsymbol{x}$,
  - A distance measure.

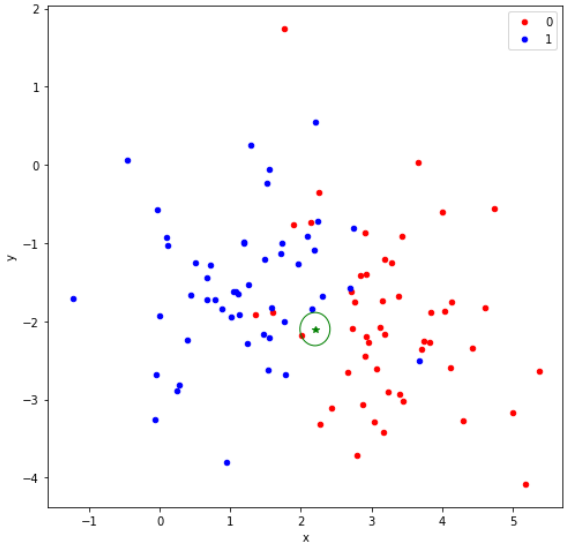
- To which class ω_1 or ω_2, does $\boldsymbol{x}$ belong?

k-Nearest Neighbours

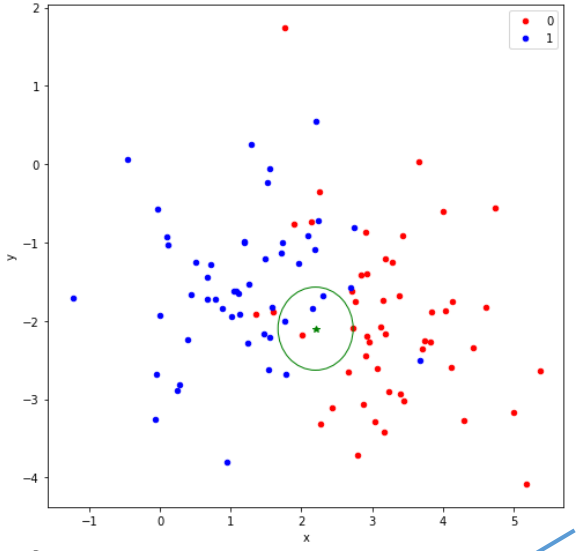
- To which class ω_1 or ω_2, does $\boldsymbol{x}$ belong?
  - Out of $\boldsymbol{x}_i, i = 1, \ldots, N,\ k$ nearest neighbours (instances represented by their feature vectors) are identified.
    - Regardless of the class label.
    - Regardless of the number of classes $|\Omega|$.
    - k should be odd (when $|\Omega| = 2$).
      - Preferably, $k\%|\Omega| \neq 0$
  - Out of the k instances, k_u that belongs to $\omega_u, u = 1,2$, are identified.
    - Clearly, $\sum_u k_u = k$.
  - $\boldsymbol{x}$ is assigned to $\omega_{\hat{u}}$, where $\hat{u} = \underset{u}{\operatorname{argmax}}\, k_u$

k-Nearest Neighbours

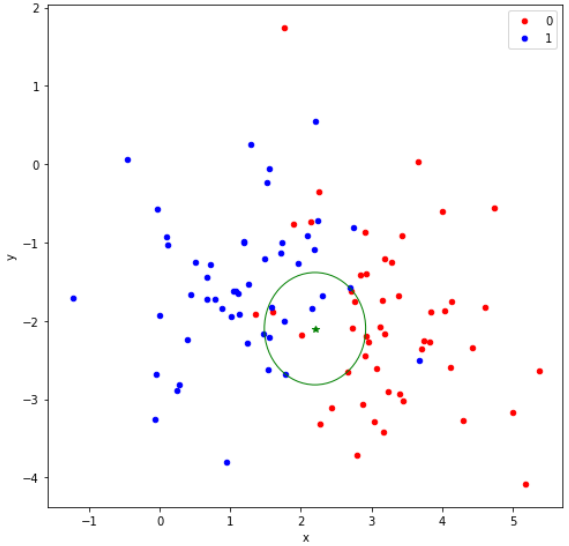
$k = 1$
$\boldsymbol{x}$ is assigned to ω_1(red)



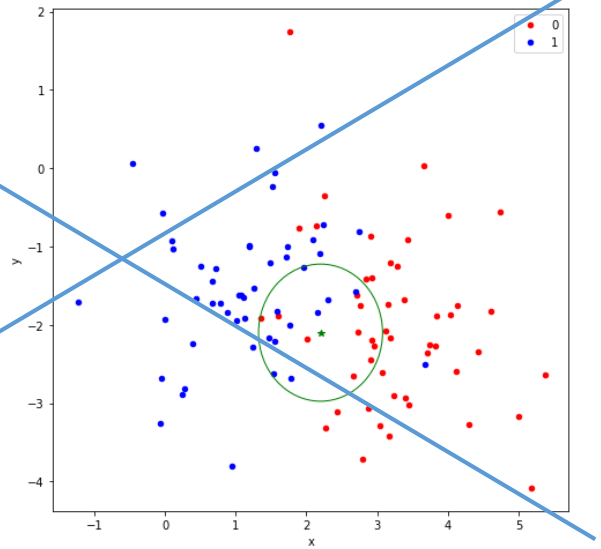
$k = 5$
$\boldsymbol{x}$ is assigned to ω_2 (blue)



$k = 11$
$\boldsymbol{x}$ is assigned to ω_2 (blue)

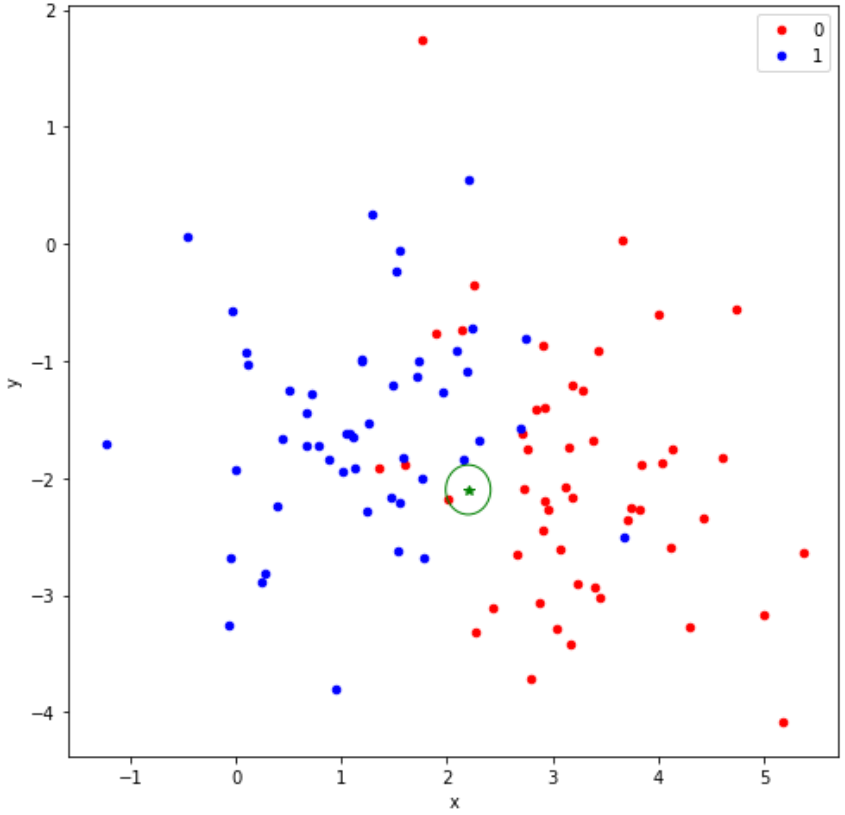


$k = 20$
$\boldsymbol{x}$ is assigned to ?
$k_0 = k_1 = 10$



k-Nearest Neighbours

- The simplest version of kNN is for $k = 1$. It is known as the *nearest neighbour* rule.
- In general, a small k means:
  - High sensitivity.
  - Little generalization.



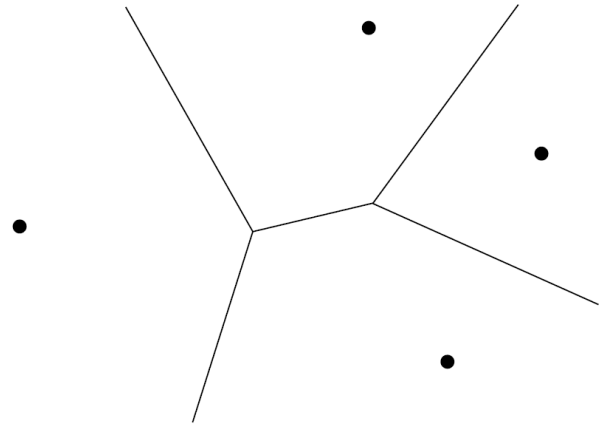
1-nearest neighbour: Voronoi tessellation

- When $k = 1$, N regions R_i of the l-dimensional space can be defined as:

$$R_i = \{\boldsymbol{x}: d(\boldsymbol{x}, \boldsymbol{x}_i) < d(\boldsymbol{x}, \boldsymbol{x}_j), i \neq j\},$$

Where $d(.)$ is a distance metric.

- This partition is known as *Voronoi tessellation*.



Source: Theodoridis, S., & Koutroumbas, K. “Pattern recognition. ” Fourth Edition, 9781597492720, 2008

k-Nearest Neighbours

- Make k larger.
  - Low sensitivity.
  - Higher generalization.
- But
  - High complexity in search of the nearest neighbours among the *N* available training samples $= O(lN + kN)$.
    - The problem is severe in high-dimensional feature spaces.
  - The closer k to N, the more k-NN is approaching the prior classifier. We will see it later!

k-NN for regression

- In case of continuous prediction, k-NN can be identically used to predict a value $\hat{y}$ for an unknown feature vector $\boldsymbol{x}$.
- Given:
  - The training feature vectors $\boldsymbol{x}_i, i = 1, \ldots, N$ associated with real-valued targets y_i,
  - An unknown feature vector $\boldsymbol{x}$,
  - A distance measure.

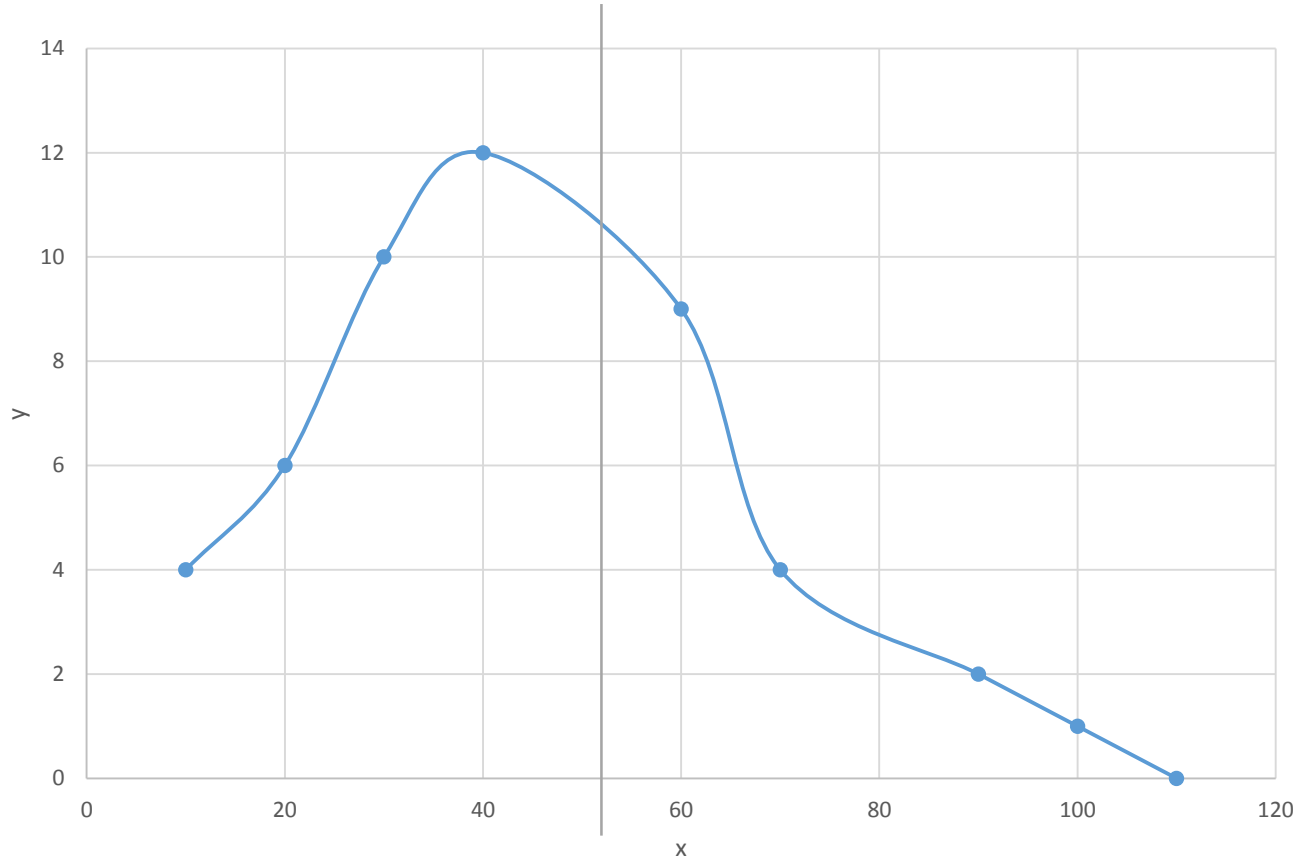
k-NN for regression

- What is the predicted value $\hat{y}$ for $\boldsymbol{x}$?
  - Out of $\boldsymbol{x}_i, i = 1, \ldots, N,\ k$ nearest neighbours (samples represented by their feature vectors) are identified.
    - Regardless of y_i.
  - $\hat{y}$ is computed as the mean of $y_u, u = 1, \ldots, k$:

$$\hat{y} = \frac{1}{k} \sum_{u=1}^{k} y_u \,.$$

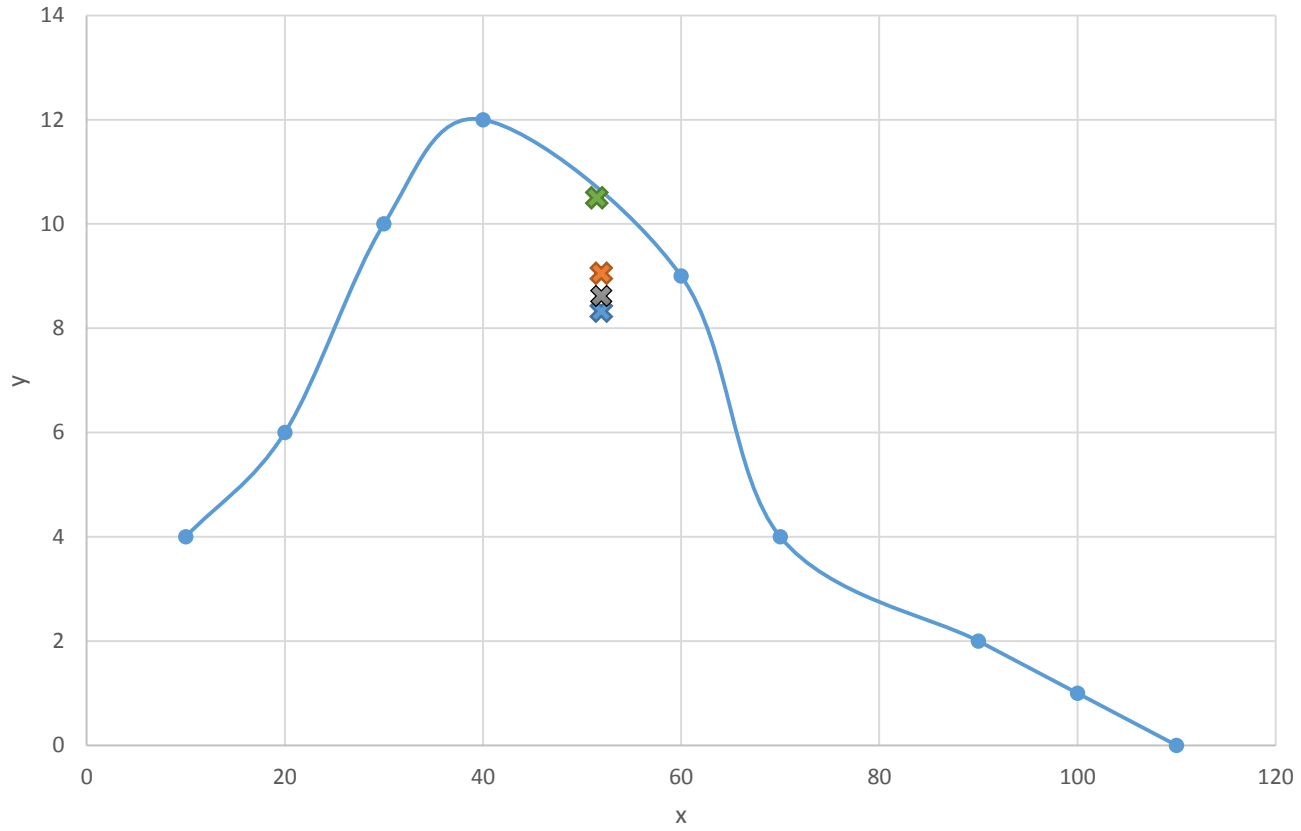
k-NN for regression

- Let's predict $\hat{y}$ for $x = 52$



k-NN for regression

- $k = 1$, $k = 2$, $k = 3$, $k = 4$



-

k-Nearest Neighbours

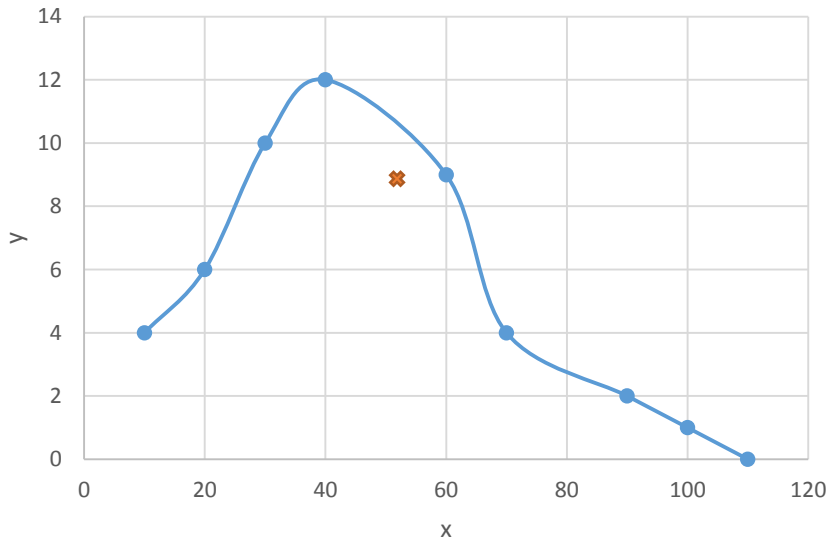
- Advantages
  - It can achieve good results when the data set (N) is large.
  - Easy to implement.

- Disadvantages
  - Its performance may degrade dramatically when the value of N is relatively small.
  - High computational cost when the data is very large.
  - Best k has to be found.
  - data-adaptive distance metric that leads to an optimal performance has to be found (when the data set is small).

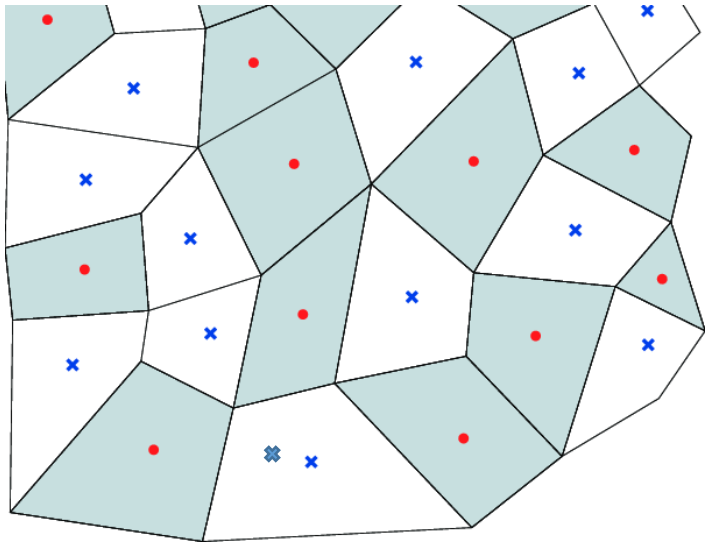
Overfitting

- When k is very small but > 1, the training samples are well modelled.
  - The error on the training set tends to be close to zero.
- When $k = 1$, the training samples are perfectly modelled.
  - The error on the training set $= 0$.

Regression



Classification



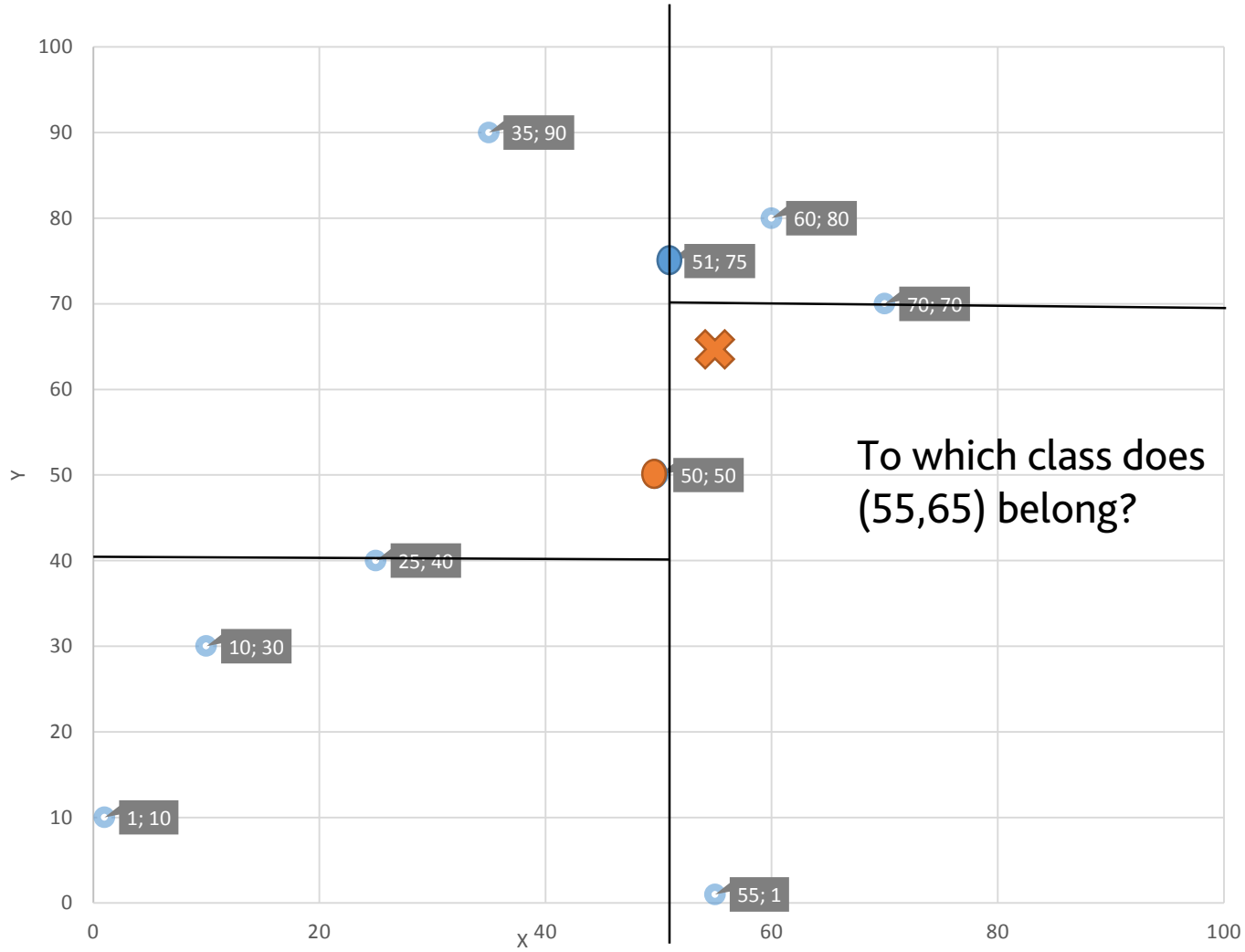
Source: Mitrovic, D. (2005). *Discrimination and Retrieval of Environmental sounds*. na.

K-D Tree

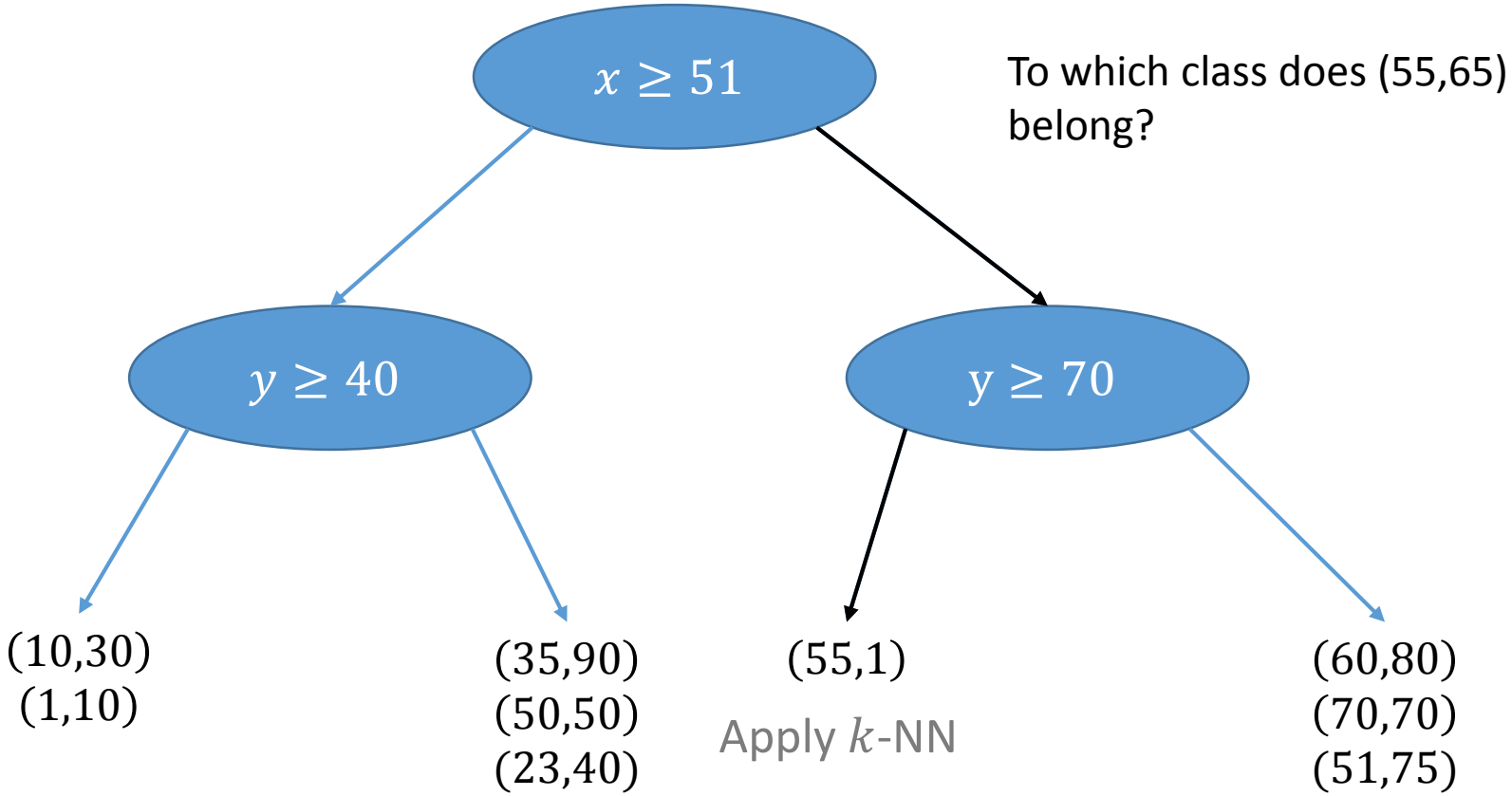
K-D Tree

- When N is very large, it is not practical to compute the distance between the unseen sample and all training samples.
- K-D Tree is an improvement over KNN that builds a data structure such as the whole data is organized as a tree.
  - Repeat until reaching the threshold:
    - Picking a random dimension (attribute) from the K dimensions.
    - Finding the median.
    - Splitting the data "evenly" w.r.t median value.
  - The threshold is the predetermined number of instances in each branch.

K-D Tree

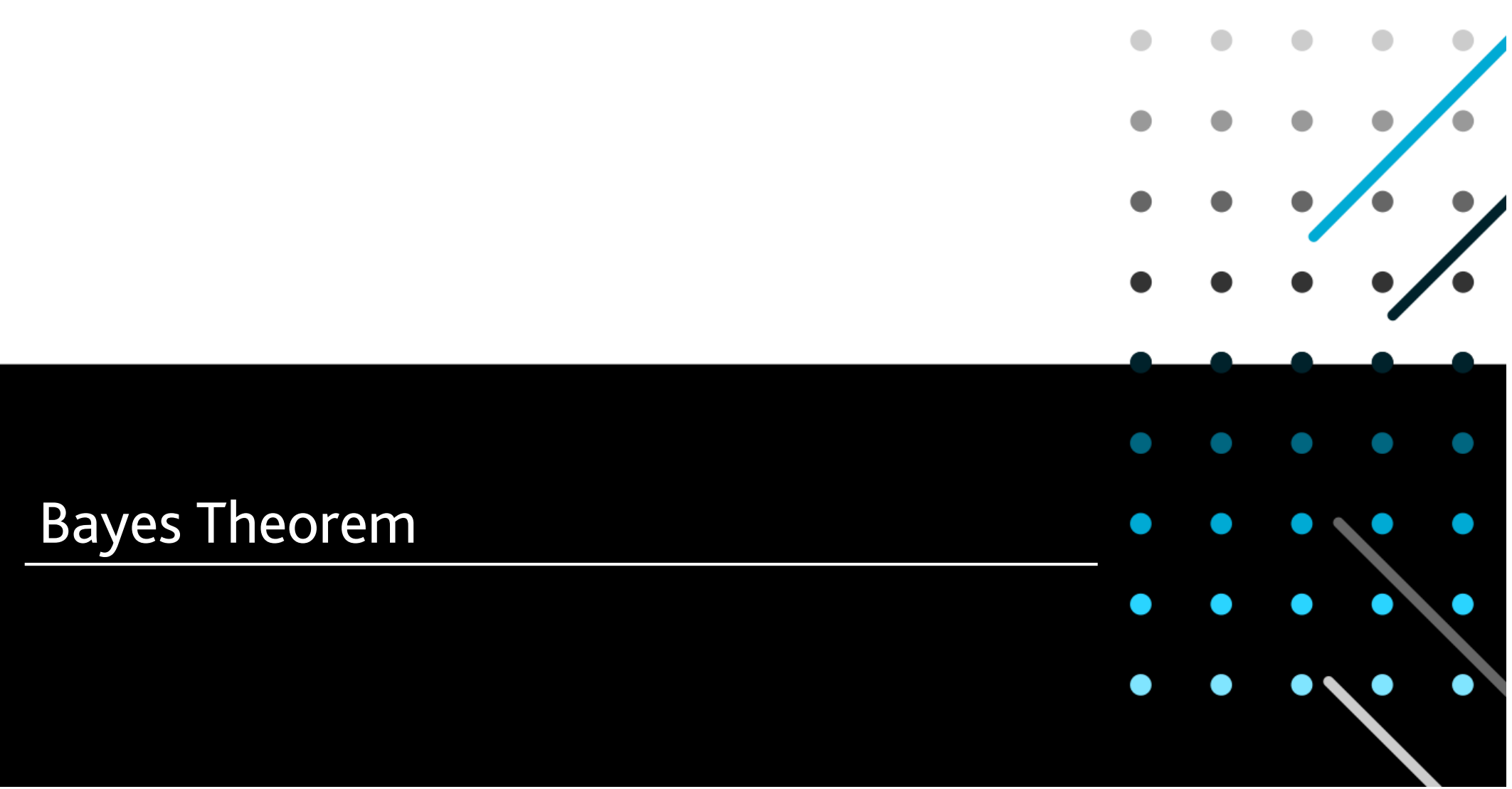

100
90
80
70
60
50
40
30
20
10
0
Y
0
20
40
60
80
100
X
35; 90
60; 80
51; 75
70; 70
50; 50
25; 40
10; 30
1; 10
55; 1
To which class does
(55,65) belong?

K-D Tree



To which class does (55,65) belong?

Bayes Theorem

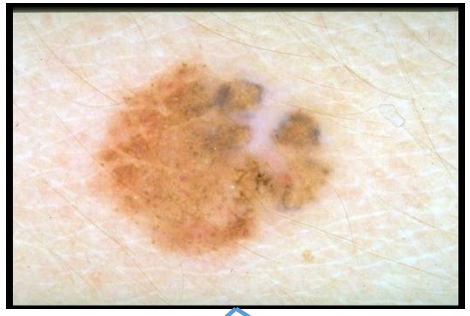


Recall

- The conditional probability of A given B:
  - $P(A|B) = \frac{P(A,B)}{P(B)}$.
- The conditional probability of B given A:
  - $P(B|A) = \frac{P(A,B)}{P(A)}$.
- Remember that:
  - $P(A|B) \neq P(B|A)$.
  - $P(A)$ and $P(B)$ are marginal probabilities.
- The joint probability of A and B:
  - $P(A,B) = P(A|B) * P(B)$.
  - $P(A,B) = P(B|A) * P(A)$.

Bayes theorem

Tumour



Benign

Benign moles have usually a single shade of brown.

Malignant (Melanoma)

A melanoma may have different shades of brown, tan or black.

Bayes theorem

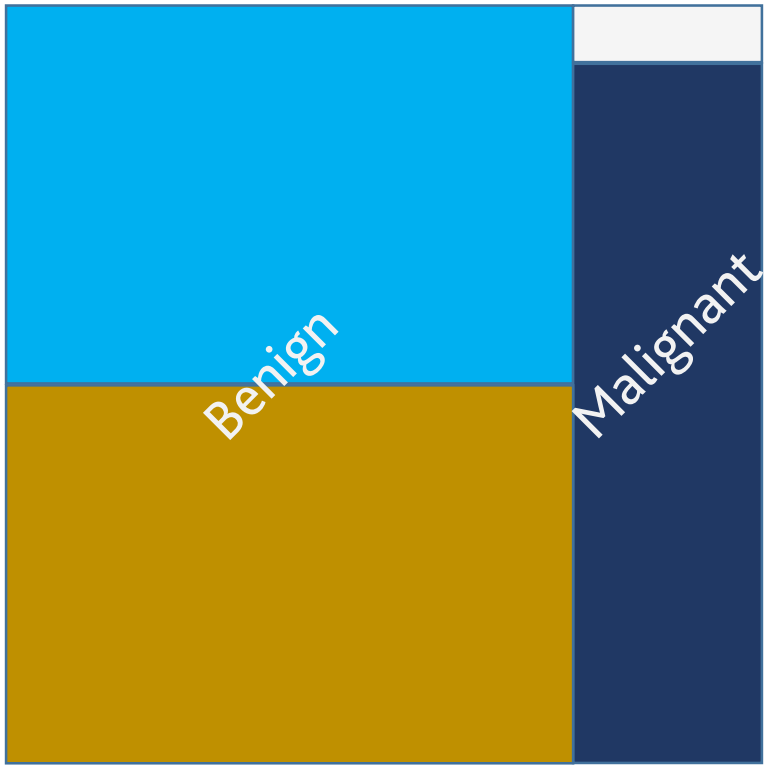
- From previous experiences;
  - Out of 120 benign cases,
    - 60 have a single shade of brown.
    - 60 have different shades of different dark colours.
  - Out of 40 malignant cases,
    - 3 have a single shade of brown.
    - 37 have different shades of different colours.
- More benign cases with different shades than malignant cases.

Bayes theorem

63 cases with a single shade (39,375%)

60 benign cases with a single shade (37,5%)

3 malignant cases with a single shade (1,875%)



120 Benign cases (75%)

40 Malignant cases (25%)

60 benign cases with different shades (37,5%)

37 malignant cases with different shades (23,125%)

97 cases with different shades (60,625%)

Bayes theorem

After obtaining the health history of the patient, it has been noticed that he/she got another malignant tumour in the past five years.

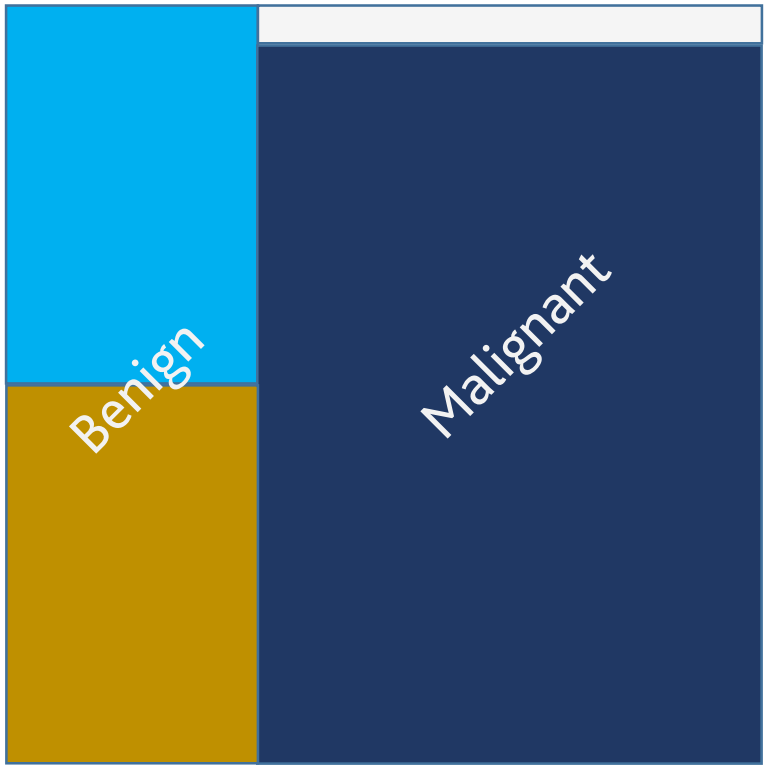
- From previous experiences with a second malignant tumour,
  - Out of 10 benign cases,
    - 5 have a single shade of brown.
    - 5 have different shades of different colours.
  - Out of 20 malignant cases,
    - 1 has a single shade of brown.
    - 19 have different shades of different colours.
- More malignant cases with different shades than benign cases.

Bayes theorem

6 cases with a single shade (20%)

5 benign cases with a single shade (16,666%)

1 malignant cases with a single shade (3,333%)



10 Benign cases (33,333%)

20 Malignant cases (66,666%)

5 benign cases with different shades (16,666%)

19 malignant cases with different shades (63,333%)

24 cases with different shades (80%)

Bayes theorem

Joint probabilities

	Benign	Malignant	Total
Single Shade	0,1666	0,0333	0,2
Different Shades	0,1666	0,6333	0,8
Total	0,3333	0,6666	$\sum$ = 1

Marginal probabilities

Bayes theorem

- $P(Benign) = 0{,}333$
- $P(Malignant) = 0{,}666$
- $P(SingleShade) = 0{,}2$
- $P(DifferentShades) = 0{,}8$
- $P(Benign, SingleShade) = 0{,}1666$

- $P(SingleShade|Benign) = \frac{\#Benign \cap SingleShade}{\#Benign} = \frac{5}{10} = \frac{P(Benign,SingleShade)}{P(Benign)} = \frac{0{,}1666}{0{,}333} = 0{,}5$
- $P(Benign|SingleShade) = \frac{\#Benign \cap SingleShade}{\#SingleShade} = \frac{5}{6} = \frac{P(Benign,SingleShade)}{P(SingleShade)} = \frac{0{,}1666}{0{,}2} = \frac{P(Benign) * P(SingleShade|Benign)}{P(SingleShade)} = \frac{0{,}333 * 0{,}5}{0{,}2} = 0{,}8333$

Bayes theorem

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

Bayes theorem

- Given:
  - A classification task of k classes; $\omega_1, \ldots, \omega_k$
  - An unknown instance represented by a feature vector $\boldsymbol{x}$
- k conditional probabilities $p(\omega_j|\boldsymbol{x}), j = 1, \ldots, k$ are formed.
- They are also referred to as *a posteriori probabilities*.

Bayes theorem

- Let's consider the two classes ω_1 and ω_2, to which all the training instances belong.
- The *a priori probabilities* $p(\omega_1)$ and $p(\omega_2)$ are assumed to be known.
  - They can be estimated as:
    - $p(\omega_u) \approx \frac{N_u}{N}, u = 1,2.$
- The class-conditional probability density functions $p(\boldsymbol{x}|\omega_u), u = 1,2$, are assumed to be known.
  - They are also referred to as *the likelihood function* of ω_u w.r.t $\boldsymbol{x}$.

Bayes theorem

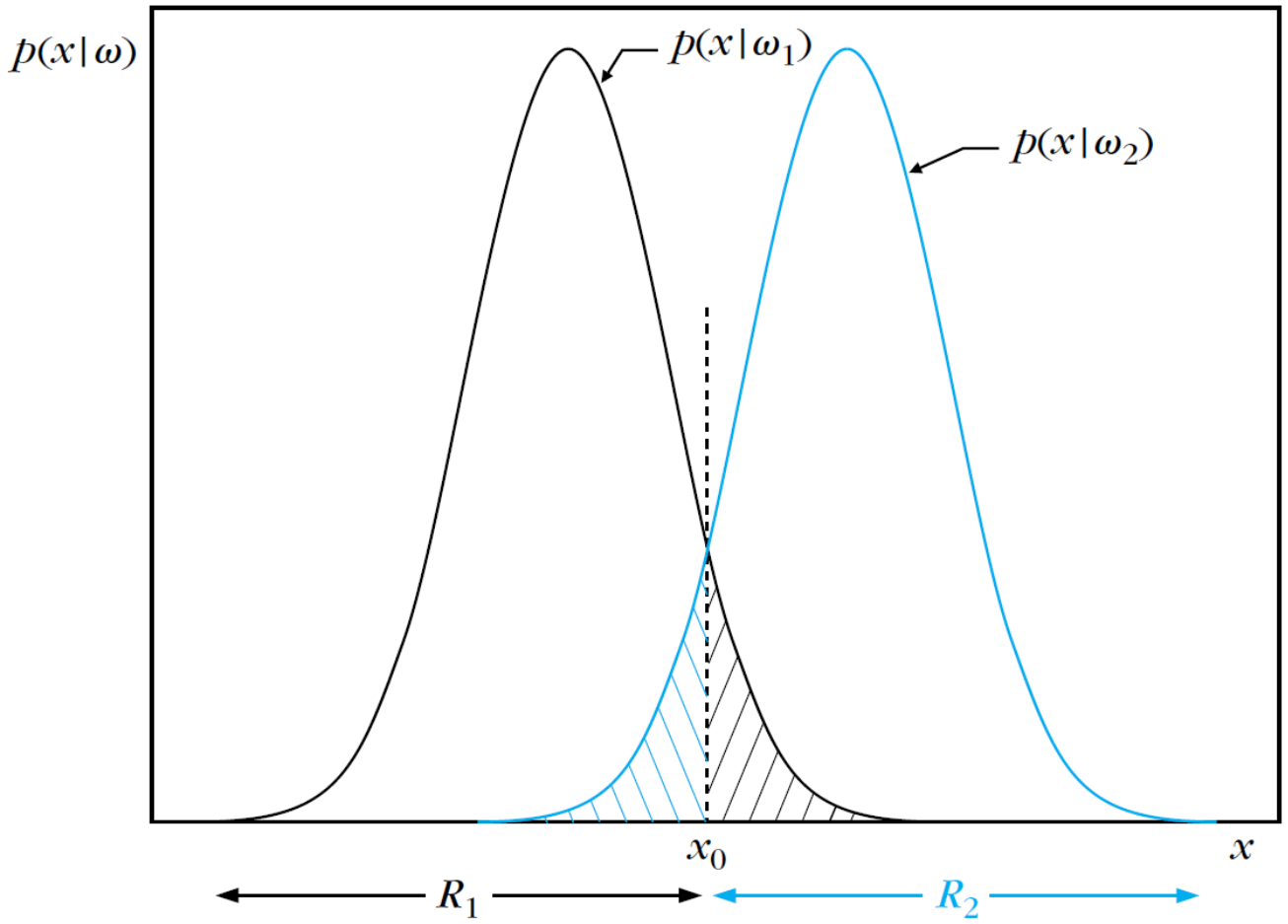
- Following the Bayes theorem:

$$P(\omega_u|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\omega_u) * P(\omega_u)}{p(\boldsymbol{x})}$$

Where $p(\boldsymbol{x}) = \sum_{u=1}^{k} p(\boldsymbol{x}|\omega_u)P(\omega_u)$.

- The *Bayes classification rule* can now be stated as:
  - If $P(\omega_1|\boldsymbol{x}) > P(\omega_2|\boldsymbol{x})$, $\boldsymbol{x}$ is classified to ω_1.
  - If $P(\omega_2|\boldsymbol{x}) > P(\omega_1|\boldsymbol{x})$, $\boldsymbol{x}$ is classified to ω_2.
- Important to note:
  - Since $p(\boldsymbol{x})$ is equal for all classes, it is not taken into consideration.
  - If the *a priori probabilities* $p(\omega_1)$ and $p(\omega_2)$ are equal, $P(\omega_u|\boldsymbol{x}) \propto p(\boldsymbol{x}|\omega_u)$

Bayes theorem



Source: Theodoridis, S., & Koutroumbas, K. “Pattern recognition. ” Fourth Edition, 9781597492720, 2008

Maximum Likelihood Estimation (MLE)

- What is the best $p(\boldsymbol{x}|\omega_u;\boldsymbol{\theta})$ that explains well the data distribution in ω_u?
  - We assume that N_u samples that belong to ω_i are drawn from a PDF, what are the parameters $\boldsymbol{\theta}$ of this PDF?
  - We know $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N_u}\}$.
- Let the PDF from which the N_u sample are drawn be: $p(\boldsymbol{x};\boldsymbol{\theta})$.
- Assuming that the samples are *statistically independent,* the joint pdf can be written as:

$$p(X;\boldsymbol{\theta}) = \prod_{i=1}^{N} p(\boldsymbol{x}_i;\boldsymbol{\theta})\,.$$

Maximum Likelihood Estimation (MLE)

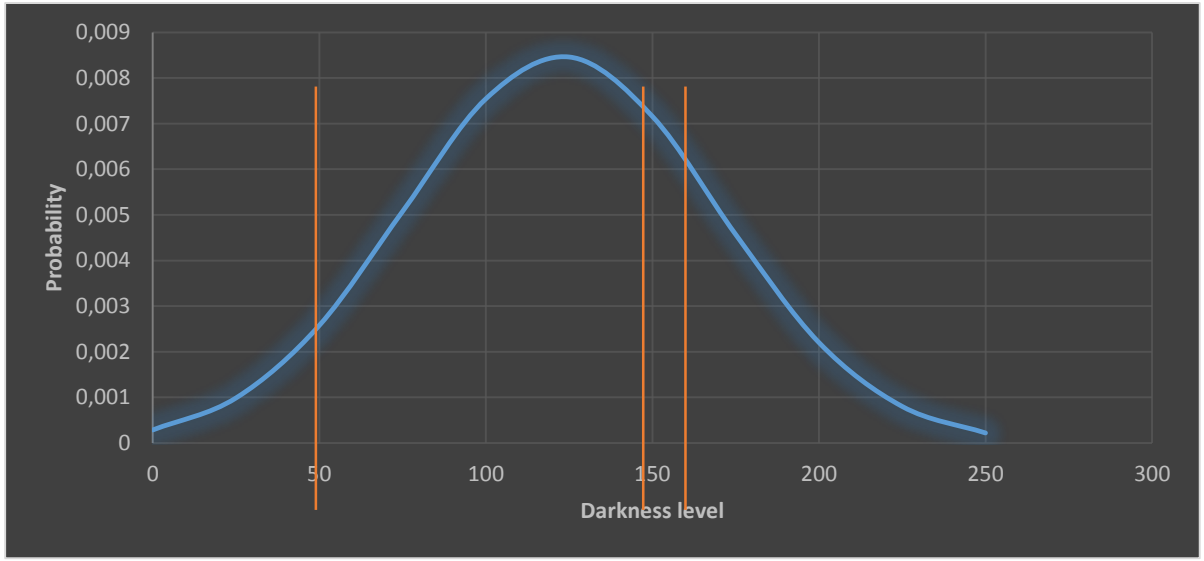
- The maximum likelihood estimation (MLE) estimates $\boldsymbol{\theta}$ so that the likelihood function $p(X; \boldsymbol{\theta})$ is maximized, that is:

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{N} p(\boldsymbol{x}_i; \boldsymbol{\theta}).$$

Maximum Likelihood Estimation (MLE)

- Let's consider the darkness level of the mole as a feature for the melanoma classification problem.
  - The scale is from 0 (black) to 255 (white).
- Considering three samples: $x_1 = 151, x_2 = 165$ and $x_3 = 52$ belonging to the benign class.

$$\mu = 122{,}666$$
$$\sigma = 47{,}1111$$



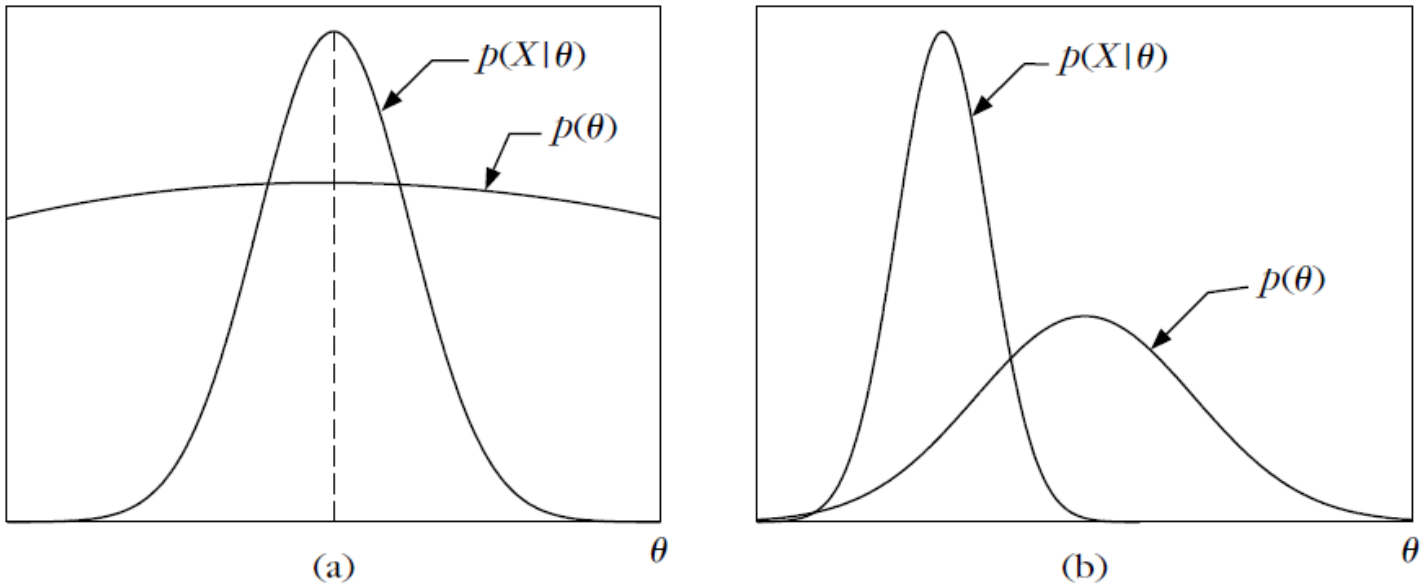
Maximum a Posteriori Probability Estimation (MAP)

- MLE is very sensitive to random variations and thus overfits the data.
- MAP regularizes this process by considering $\boldsymbol{\theta}$ as a random vector.
- Given N_u samples represented by $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N_u}\}$:

$$p(\boldsymbol{\theta}|X) = \frac{p(\boldsymbol{\theta}) * p(X|\boldsymbol{\theta})}{p(X)}$$

MLE vs. MAP

- MLE and MAP will be approximately similar is (a) as the prior $p(\theta)$ does not significantly affect the likelihood distribution $p(X|\theta)$
- In contrast, MLE and MAP will be different in (b).

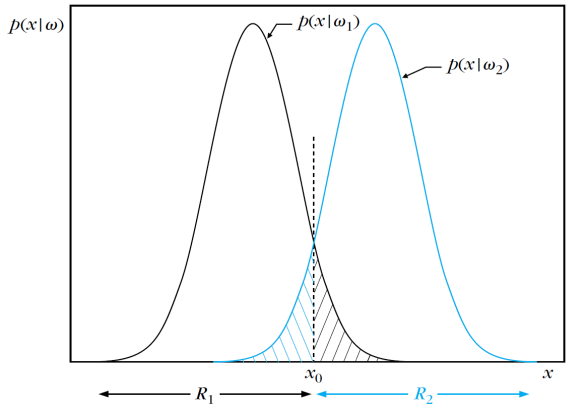


Source: Theodoridis, S., & Koutroumbas, K. “Pattern recognition. ” Fourth Edition, 9781597492720, 2008

Classification error probability

- From the figure, there is a possibility that a sample lying in R_2 but belongs to ω_1 and vice-versa.
  - The decision errors are unavoidable.
- The maximum total probability of committing an error is:

$$P_e = P(\omega_2) \int_{-\infty}^{x_0} p(\boldsymbol{x}|\omega_2) d\boldsymbol{x} + P(\omega_1) \int_{x_0}^{+\infty} p(\boldsymbol{x}|\omega_1) d\boldsymbol{x}$$



Naïve Bayes

Naïve bayes

- For good estimates of the PDFs, the number of training samples N must be large enough.
- If N could be regarded as a sufficient number of samples to obtain satisfactory estimates of a pdf in a one-dimensional $(l = 1)$ space,
  - N^l would be required for an l-dimensional space.

✓It is assumed that individual features $x_j, j = 1, \ldots, l$ are statistically independent.

Naïve Bayes

- Under this assumption, $p(\boldsymbol{x}|\omega_u)$ becomes:

$$p(\boldsymbol{x}|\omega_u) = \prod_{j=1}^{l} p(x_j|\omega_u), \qquad u = 1, \dots k$$

- Remember that $p(x_j|\omega_u) = \frac{p(\boldsymbol{x},\omega_u)}{p(\omega_u)}$
- Now $l * N$ samples (instead of N^l) would be enough in order to obtain good estimates.
- Naïve Bayes classifier assigns a sample represented by the feature vector $\boldsymbol{x} = [x_1, x_2, \dots, x_l]^T \in \mathbb{R}^l$ to the class:

$$\gamma(\boldsymbol{x}) = \arg\max_{\omega_u} P(\omega_u) * \prod_{j=1}^{l} p(x_j|\omega_u), \qquad i = 1, \dots, k$$

Naïve Bayes

- Assuming the attributes are categorical, what if, for example, $p(x_2|\omega_u) = 0$, $u = 1, \dots, k$?
  - $P(\omega_u) * \prod_{j=1}^{l} p(x_j|\omega_u) = 0$
- Remember that for categorical features:
  - $p(x_j|\omega_u) = \frac{p(x,\omega_u)}{p(\omega_u)} = \frac{|x_j \cap \omega_u|}{|\omega_u|}$, where $|\omega_u|$ is the cardinality of the set of training samples belonging toω_i.
- Use Laplace smoothing:
  - $p(x_j|\omega_u) = \frac{|x_j \cap \omega_u|+1}{|\omega_u|+|l_j|}$, where $|l_j|$ is the number of values the jth attribute can take.
- Or Generalized additive smoothing (Lidstone):
  - $p(x_j|\omega_u) = \frac{|x_j \cap \omega_u|+\lambda}{|\omega_u|+|l_j|*\lambda}$, where λ is a hyperparameter.

Smoothing is applied only on categorical attributes

Naïve Bayes

$$p(\omega_u|\boldsymbol{x}) \propto P(\omega_u) * \prod_{j=1}^{l} p(x_j|\omega_u), \qquad u = 1, \dots, k$$

- Multiplying many small values < 1

✓It is more convenient to work with an equivalent function $g_i(.)$, for example:

- $g_i(\boldsymbol{x}) \equiv \log\big(P(\omega_u|\boldsymbol{x})\big) \propto \log(P(\omega_u)) * \sum_{j=1}^{l} \log\big(p(x_j|\omega_u)\big).$

Summary

Summary

- KNN
- K-D Tree
- Bayes Theorem
- Naive Bayes

Thank you!

Zeyd Boukhers

E-mail: Boukhers@uni-koblenz.de
Phone: +49 (0) 261 287-2765
Web: Zeyd.Boukhers.com

University of Koblenz-Landau
Universitätsstr. 1
56070 Koblenz