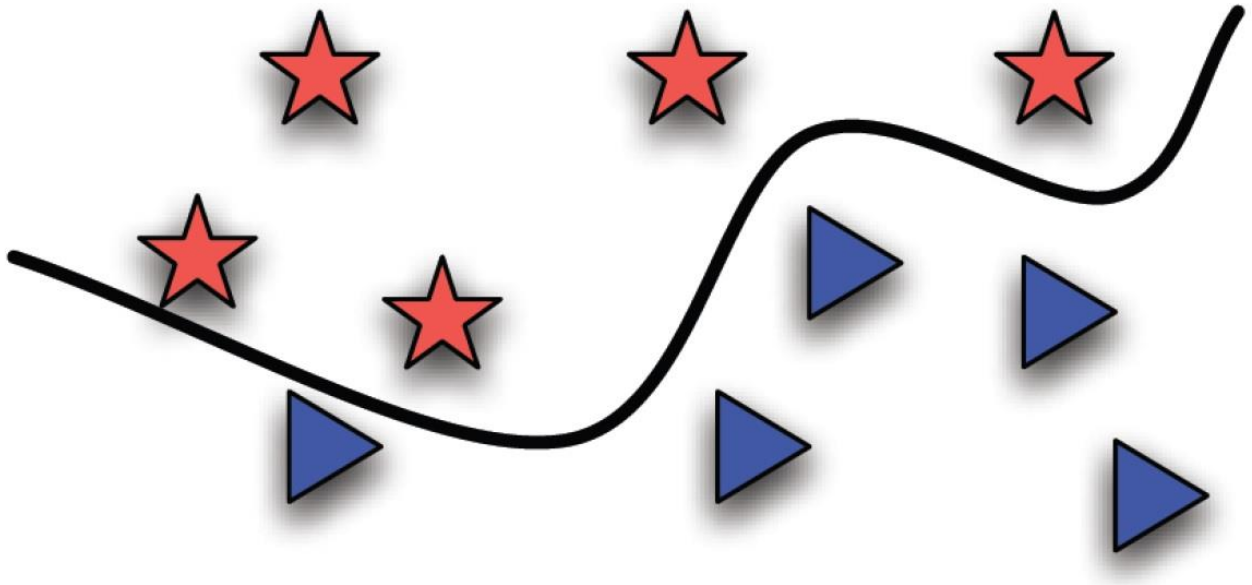
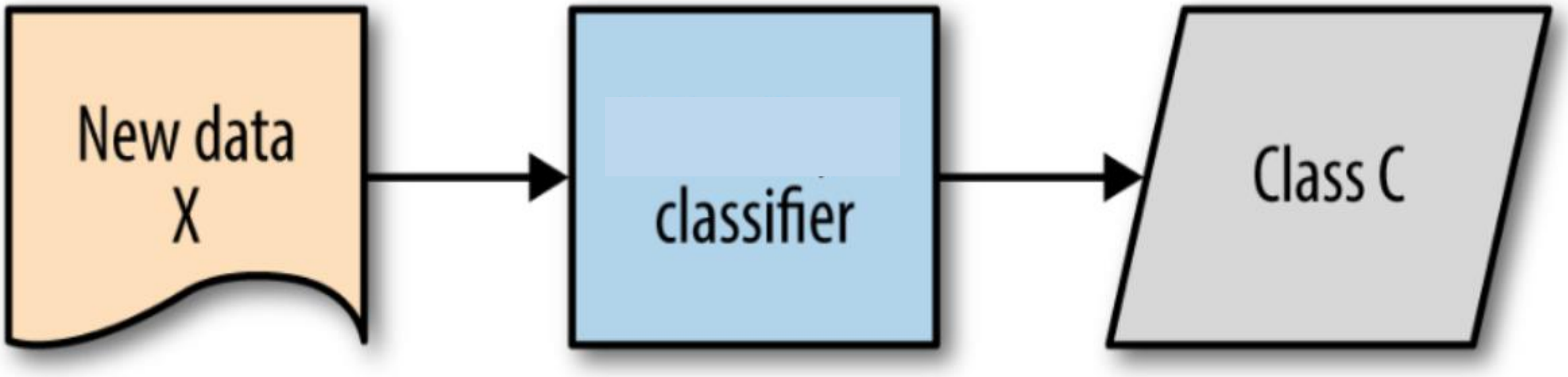
Python and Data science
Zaouche-Dahmani Djaouida

Classifications of data



Classifications



Introduction to Bayes

- Naive Bayes classifiers have been especially popular for text classification, and are a traditional solution for problems such as spam detection.
- In the context of text classification, where features may be word counts, features may follow a **multinomial distribution**. In other cases, where features are continuous, they may follow a **Gaussian distribution**.
- The goal of any probabilistic classifier is, with features x_0 ,…, x_n and classes C_0 , …, C_k, to determine the probability of the features occurring in each class, and to return the most likely class. Therefore, for each class, we want to be able to calculate is:

$$P(C_i \mid x_0, \ldots, x_n).$$

Introduction to Bayes

- Naive Bayes is a probabilistic technique for constructing classifiers.

- The characteristic assumption of the naive Bayes classifier is to consider that the value of a particular feature is independent of the value of any other feature, given the class variable.

- Despite the oversimplified assumptions mentioned previously, naive Bayes classifiers have good results in complex real-world situations. An advantage of naive Bayes is that it only requires a small amount of training data to estimate the parameters necessary for classification and that the classifier can be trained incrementally.

Introduction to Bayes

- Supervised learning.
- Probabilistic point of view : estimating the probability for a data record to belong to a class C.

Introduction to Bayes

- Given two events A and B, we have :

$$Pr(A/B) = \frac{Pr(B/A).Pr(A)}{Pr(B)}.$$

- where $Pr(X/Y)$ is the probability of X given Y.

Introduction to Bayes

class y

The probability: X belongs to y

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)}$$

$$P(y|x_1, ..., x_n) = \frac{P(x_1|y)P(x_2|y)...P(x_n|y)}{P(x_1)P(x_2)...P(x_n)}$$

The probability P is propotional to The numerator, since the denominator is constant

$$X = (x_1, x_2, x_3, ..., x_n)$$

Independent data

Introduction to Bayes

- Given a data record A defined by the values a_1, ..., a_p for the attributes A_1, ..., A_p.
- Given m classes C_1, ..., C_m.
- We want to calculate :
  - the probability that d belongs to the class C_i where $C_i \in \{C_1, ..., C_m\}$.
  - that is, $Pr(C_i/A)$.

Introduction to Bayes

$$P(y|x_1, ..., x_n) \propto P(y) \prod_{i=1}^{n} P(x_i|y)$$

$$y = argmax_y P(y) \prod_{i=1}^{n} P(x_i|y)$$

Data $(x_1,\ldots,x_n)$ belongs to class y, where $P(y/x_1,\ldots,x_n)$ is the highest probability.

Example (train data)

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Example

The weather data, with counts and probabilities

outlook			temperature			humidity			windy			play	
	yes	no		yes	no		yes	no		yes	no	yes	no
sunny	2	3	hot	2	2	high	3	4	false	6	2	9	5
overcast	4	0	mild	4	2	normal	6	1	true	3	3		
rainy	3	2	cool	3	1								
sunny	2/9	3/5	hot	2/9	2/5	high	3/9	4/5	false	6/9	2/5	9/14	5/14
overcast	4/9	0/5	mild	4/9	2/5	normal	6/9	1/5	true	3/9	3/5		
rainy	3/9	2/5	cool	3/9	1/5								

A new day

outlook	temperature	humidity	windy	play
sunny	cool	high	true	?

Introduction to Bayes

Likelihood of yes

= P(sunny/yes) * P(cool/yes) * P(high/yes) * P(true/yes) *P (yes)

$$= \frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{9}{14} = 0.0053$$

Introduction to Bayes

Likelihood of no

= P(sunny/no) * P(cool/no) * P(high/no) * P(true/no) *P (no)

$$= \frac{3}{5} \times \frac{1}{5} \times \frac{4}{5} \times \frac{3}{5} \times \frac{5}{14} = 0.0206$$

Introduction to Bayes

If outlook is sunny, the temperature is cool, and the humidity is high and it is windy

The prediction is no play

Exercise

- Suppose we have the following training set D :

A	B	C
m	b	t
m	s	t
g	q	t
h	s	t
g	q	t
g	q	f
g	s	f
h	b	f
h	q	f
m	b	f

Exercise

- Let us take the following data record d :
  - $A = m$ and $B = q$.
- Let us calculate the probabilities that this example belongs to t and to f.