Chapter 02 – Basic Classification

Lecture A2I2

Prof. Dr. Kai Höfig

Classification

- A system that can map a vector of property values onto a discrete number of classes is called a **classifier**.

measure →

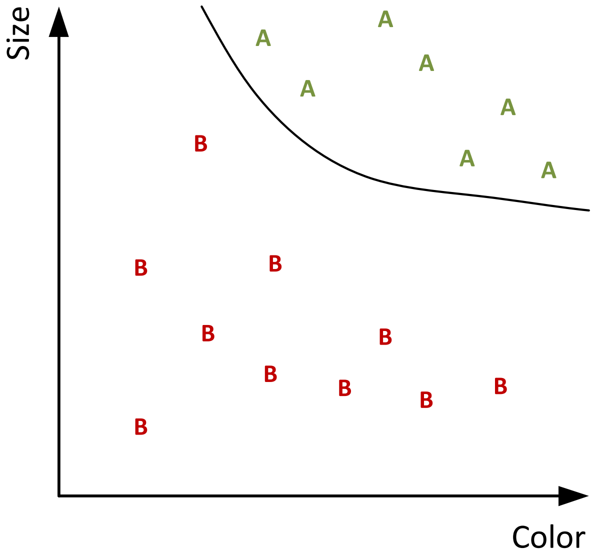
Sample	Size	Color
1	8	2
2	7	2
3	5	2
4	8	1
5	9	1
6	7	2
..	..	..

Classifier

classify →

Sample	Class
1	A
2	A
3	B
4	B
5	B
6	A
..	..

Classification is an approximation problem



- The separating line in the apple example classifies apples in two classes A and B.
- Such a (one-dimensional) line is a goal function of the classification.
- If there are *n* properties as input, the goal function is *n-1* - dimensional and called a hyperplane
- A generative model uses a probability distribution between the input and output variables.

The Golf Problem

#	Outlook	Temp	Humidity	Windy	Play
1	overcast	hot	high	yes	yes
2	overcast	cool	normal	no	yes
3	overcast	mild	high	yes	yes
4	overcast	hot	normal	no	yes
5	rainy	mild	high	no	yes
6	rainy	cool	normal	no	yes
7	rainy	cool	normal	yes	no
8	rainy	mild	normal	no	yes
9	rainy	mild	high	yes	no
10	sunny	hot	high	no	no
11	sunny	hot	high	yes	no
12	sunny	mild	high	no	no
13	sunny	cool	normal	no	yes
14	sunny	mild	normal	yes	yes

- We want to play golf and a system that supports us in making this decision using some sample data about recent decisions we made.
- What can we see in the table to support our decision?

$$P(windy|play) = \frac{P(windy \wedge play)}{P(play)}$$

$$= \frac{\frac{3}{14}}{\frac{9}{14}} = \frac{3}{9}$$

Conditional Probabilities of the Golf Problem

		Play: Yes	Play: No
Outlook	overcast	0,44	0,00
	rainy	0,33	0,40
	sunny	0,22	0,60
Temperature	hot	0,22	0,40
	cold	0,33	0,20
	mild	0,44	0,40
Humidity	high	0,33	0,80
	normal	0,67	0,20
windy	yes	0,33	0,60
	no	0,67	0,40

$$P(windy|play) = \frac{P(windy \wedge play)}{P(play)} = \frac{\frac{3}{19}}{\frac{9}{19}} = \frac{1}{3}$$

- So, we can calculate all probabilities under the condition that we played or that we have not played a golf match.
- But we want to know

$$P(play|rainy \wedge hot \wedge normal \wedge windy)$$

What do we need for classification?

- Or more precisely, we want to know if

$$P(play|rainy \wedge hot \wedge normal \wedge windy) \geq P(rest|rainy \wedge hot \wedge normal \wedge windy)$$

$$\frac{P(play) \cdot P(r,h,n,w|play)}{P(r,h,n,w)} \geq \frac{P(rest) \cdot P(r,h,n,w|rest)}{P(r,h,n,w)}$$

$$P(play) \cdot P(r,h,n,w|play) \geq P(rest) \cdot P(r,h,n,w|rest)$$

$$P(play) \cdot \frac{P(r,h,n,w,play)}{P(play)} \geq P(rest) \cdot \frac{P(r,h,n,w,rest)}{P(rest)}$$

$$P(r,h,n,w,play) \geq P(r,h,n,w,rest)$$

So, what's the big whoop?

$$
\begin{aligned}
& P(r,h,n,w,play) \\
= \quad & P(r|h,n,w,play) \cdot P(h,n,w,play) \\
= \quad & P(r|h,n,w,play) \cdot P(h|n,w,play) \cdot P(n,w,play) \\
= \quad & \ldots \\
= \quad & P(r|h,n,w,play) \cdot P(h|n,w,play) \cdot P(n|w,play) \cdot P(w|play) \cdot P(play)
\end{aligned}
$$

$$
= \quad \frac{P(r,h,n,w,play)}{P(h,n,w,play)} \cdot \frac{P(h,n,w,play)}{P(n,w,play)} \cdot \frac{P(n,w,play)}{P(w,play)} \cdot P(w|play) \cdot P(play)
$$

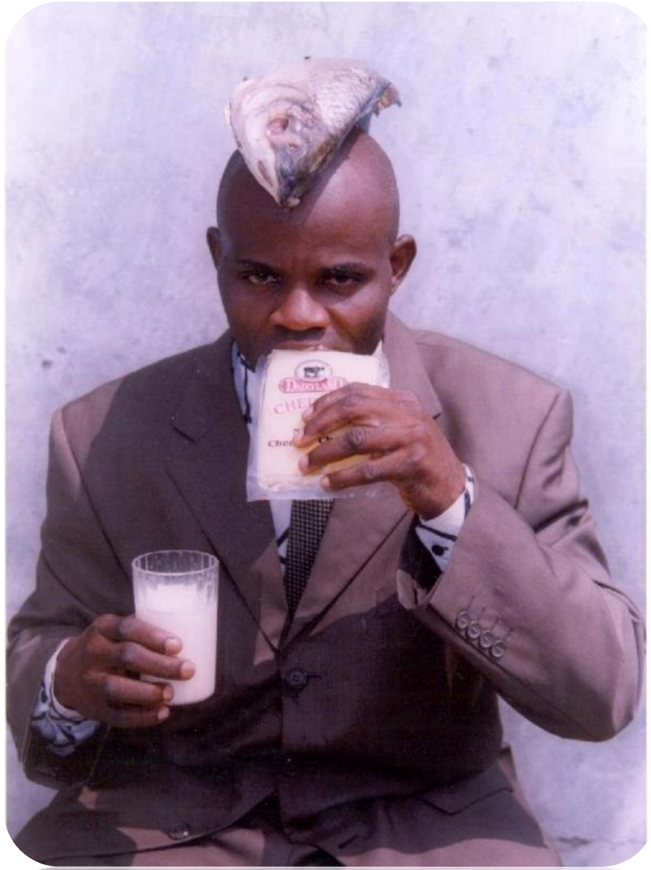
The right side we know, but for the left side, we need all the cuts!

Where to get the cuts?

- Nowhere

- For the golf game, we need vastly more samples to calculate the probability that the sun is shining, its is windy and the humidity is low at the same time and so on. Measuring all the cuts would take ages.

- For **spam classification** its is the same: we can easily measure the probability for the word *lottery* and for the word *Nigeria* and all the other words that may or may not classify spam. But measuring all the cuts would require way too much data.



Funny picture from ‚the nigeria connection‘ spam wave

https://en.wikipedia.org/wiki/Advance-fee_scam
https://www.youtube.com/watch?v=Q8I0Vip5YUw

Lets make a naïve assumption

- Well if we do not want to collect a lot of data that contains all the cuts of our variables so that we can measure the exact distribution, we can make the assumption, that all variable are independent from each other, **except for the dependent variable that we want to measure**.

- And it is pretty naïve to believe that humidity is totally independent from the outside temperature, right?

- But it makes our math much more easy, because with this assumption, it is

If A is independent from B, it is

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

$$\begin{aligned} & P(r, h, n, w, play) \\ = \quad & P(r|h, n, w, play) \cdot P(h|n, w, play) \cdot P(n|w, play) \cdot P(w|play) \cdot P(play) \\ \propto \quad & P(r|play) \cdot P(h|play) \cdot P(n|play) \cdot P(w|play) \cdot P(play) \end{aligned}$$

Hurray, Naïve Bayes Classification!

$$
\begin{aligned}
P(play|r,h,n,w) &\geq P(rest|r,h,n,w) \\
&\Leftrightarrow \\
P(r|play)\cdot P(h|play)\cdot & \quad P(r|rest)\cdot P(h|rest)\cdot \\
P(n|play)\cdot P(w|play)\cdot P(play) &\geq P(n|rest)\cdot P(w|rest)\cdot P(rest) \\
0.33\cdot 0.22\cdot 0.67\cdot 0.33\cdot 0.64 &\geq 0.4\cdot 0.4\cdot 0.2\cdot 0.6\cdot 0.36 \\
0.0103 &\geq 0.0069
\end{aligned}
$$

			Play	
			Yes	No
	Outlook	overcast	0,44	0,00
		rainy	0,33	0,40
		sunny	0,22	0,60
	Temperature	hot	0,22	0,40
		cold	0,33	0,20
		mild	0,44	0,40
	Humidity	high	0,33	0,80
		normal	0,67	0,20
	windy	yes	0,33	0,60
		no	0,67	0,40

Lets play golf in the rain on a hot windy day!

Numerical data

- So far, we used categorical data. But Naïve Bayes also works with numerical data such as *income*, *temperature* or *kph:*

$$P(X_i = x_k | C_i) = g(x_k; \mu_{i,j}, \sigma_{i,j}) = \frac{1}{\sqrt{2\pi} \cdot \sigma} \cdot e^{-\frac{(x_i - \mu)^2}{2\sigma^2}}$$

with

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i$$

and

$$\sigma = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \mu)^2)}$$