Naive Bayes Classifier

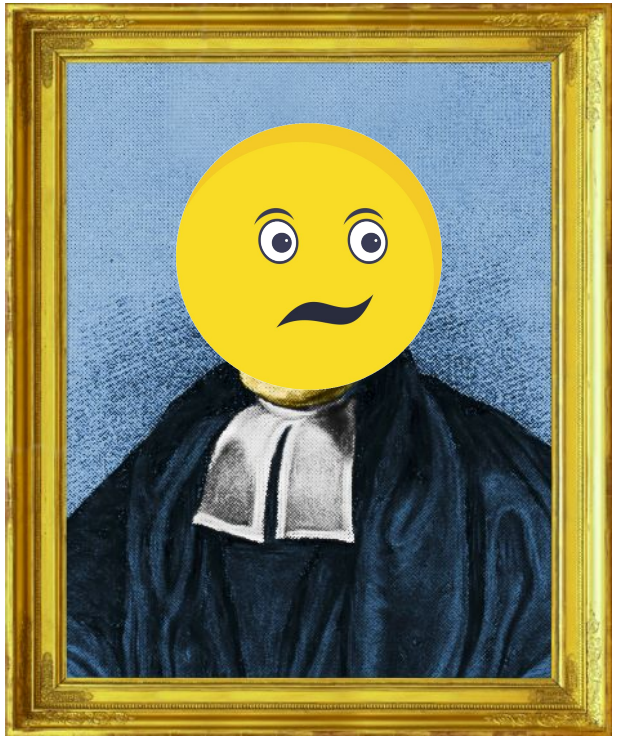
for *nominal* data

Naive Bayes classifiers

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}, \text{ if } P(B) \neq 0.$$

Thomas Bayes

(1701 – 1761)

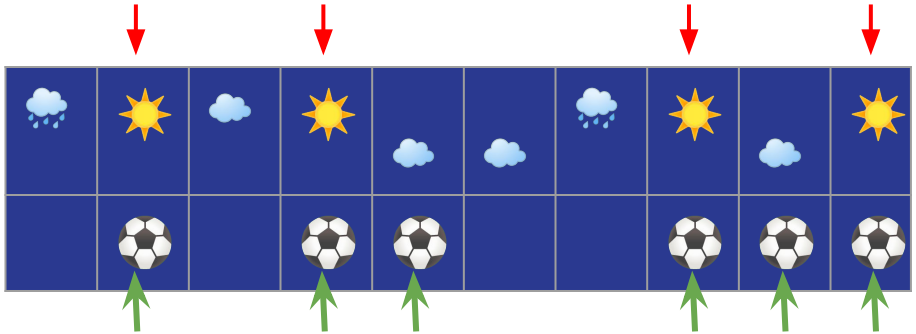


Conditional Probabilities

Prob. of A given B

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

Prob. of Soccer Game given that is Sunny



P(⚽ | ☀️) = P (⚽ ∩ ☀️) / P(⚽)

P(⚽ | ☀️) = 4/10 / 6/10

Proof of Bayes' Theorem

Prob. of A given B

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

Prob. of B given A

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

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}, \text{ if } P(B) \neq 0.$$

Naive Bayes Classification Algorithm

Goal Given the values of a certain number of attributes, find the most probable class.

Making use of Bayes' theorem…

$$\arg\max_{c_j \in C} P(c_j | a_1, a_2, \ldots, a_n)$$

$$\arg\max_{c_j \in C} \frac{P(a_1, a_2, \ldots, a_n | c_j) P(c_j)}{P(a_1, a_2, \ldots, a_n)}$$

Constant

$$\arg\max_{c_j \in C} P(a_1, a_2, \ldots, a_n | c_j) P(c_j)$$

Naive Bayes Classification Algorithm

Assumption Conditional independence of the values of the attributes given the class.

$$P(a_1, a_2, \ldots, a_n | c_j) = P(a_1|c_j) * P(a_2|c_j) * \ldots * P(a_n|c_j)$$

Creating the model We need to estimate the probabilities of:

- Each class: $P(c_j)$
- Of each class given a certain value of an attribute: $P(a_i|c_j)$

Prediction The chosen class must maximize the expression: $P(c_j) \prod P(a_i|c_j)$

Prediction

The chosen class must maximize the expression: $P(c_j) \prod P(a_i|c_j)$

outlook	temp	humid	wind	sport
sunny	hot	high	weak	no
sunny	hot	high	strong	no
overcast	hot	high	weak	yes
rainy	mild	high	weak	yes
rainy	cool	normal	weak	yes
rainy	cool	normal	strong	no
overcast	cool	normal	strong	yes
sunny	mild	high	weak	no
sunny	cool	normal	weak	yes
rainy	mild	normal	weak	yes
sunny	mild	normal	strong	yes
overcast	mild	high	strong	yes
overcast	hot	normal	weak	yes
rainy	mild	high	strong	no

Considering the following values for the attributes, what is the most probable class?

x = {sunny, cool, high, strong}

P(yes) = 9/14	P(no) = 5/14
P(sunny|yes) = 2/9	P(sunny|no) = 3/5
P(cool|yes) = 3/9	P(cool|no) = 1/5
P(high|yes) = 3/9	P(high|no) = 4/5
P(strong|yes) = 3/9	P(strong|no) = 3/5
0.0053 or 0.5%	0.02 or 2%

Prediction

The chosen class must maximize the expression: $P(c_j) \prod P(a_i|c_j)$

outlook	temp	humid	wind	sport
sunny	hot	high	weak	no
sunny	hot	high	strong	no
overcast	hot	high	weak	yes
rainy	mild	high	weak	yes
rainy	cool	normal	weak	yes
rainy	cool	normal	strong	no
overcast	cool	normal	strong	yes
sunny	mild	high	weak	no
sunny	cool	normal	weak	yes
rainy	mild	normal	weak	yes
sunny	mild	normal	strong	yes
overcast	mild	high	strong	yes
overcast	hot	normal	weak	yes
rainy	mild	high	strong	no

PROBLEM: what if these were the attributes…?

x = {overcast, cool, high, strong}

P(yes) = 9/14	P(no) = 5/14
P(overcast|yes) = 4/9	P(overcast|no) = 0/5
P(cool|yes) = 3/9	P(cool|no) = 1/5
P(high|yes) = 3/9	P(high|no) = 4/5
P(strong|yes) = 3/9	P(strong|no) = 3/5
0.0105 or 1%	0%

Solution

Laplace smoothing technique

$$P(x) = \frac{n_x + \alpha}{total + \alpha * nvals}$$

$$P(x) = \frac{0 + \alpha}{total + \alpha \times nvals}$$

- n_x = number of occurrences of a value
- α = smoothing parameter
- nvals = number of possible values for the attribute

Laplace smoothing

$$P(x) = \frac{n_x + \alpha}{total + \alpha * nvals}$$

outlook	temp	humid	wind	sport
sunny	hot	high	weak	no
sunny	hot	high	strong	no
overcast	hot	high	weak	yes
rainy	mild	high	weak	yes
rainy	cool	normal	weak	yes
rainy	cool	normal	strong	no
overcast	cool	normal	strong	yes
sunny	mild	high	weak	no
sunny	cool	normal	weak	yes
rainy	mild	normal	weak	yes
sunny	mild	normal	strong	yes
overcast	mild	high	strong	yes
overcast	hot	normal	weak	yes
rainy	mild	high	strong	no

x = {overcast, cool, high, strong}

α = 1

$P(\text{yes}) = \frac{9+1}{14+1*2}$	$P(\text{no}) = \frac{5+1}{14+1*2}$

Laplace smoothing

$$P(x) = \frac{n_x + \alpha}{total + \alpha * nvals}$$

outlook	temp	humid	wind	sport
sunny	hot	high	weak	no
sunny	hot	high	strong	no
overcast	hot	high	weak	yes
rainy	mild	high	weak	yes
rainy	cool	normal	weak	yes
rainy	cool	normal	strong	no
overcast	cool	normal	strong	yes
sunny	mild	high	weak	no
sunny	cool	normal	weak	yes
rainy	mild	normal	weak	yes
sunny	mild	normal	strong	yes
overcast	mild	high	strong	yes
overcast	hot	normal	weak	yes
rainy	mild	high	strong	no

x = {overcast, cool, high, strong}

α = 1

$P(yes) = \frac{10}{16}$	$P(no) = \frac{6}{16}$
$P(overcast\|yes) = \frac{4+1}{9+1*3}$	$P(overcast\|no) = \frac{0+1}{5+1*3}$

Laplace smoothing

$$P(x) = \frac{n_x + \alpha}{total + \alpha * nvals}$$

outlook	temp	humid	wind	sport
sunny	hot	high	weak	no
sunny	hot	high	strong	no
overcast	hot	high	weak	yes
rainy	mild	high	weak	yes
rainy	cool	normal	weak	yes
rainy	cool	normal	strong	no
overcast	cool	normal	strong	yes
sunny	mild	high	weak	no
sunny	cool	normal	weak	yes
rainy	mild	normal	weak	yes
sunny	mild	normal	strong	yes
overcast	mild	high	strong	yes
overcast	hot	normal	weak	yes
rainy	mild	high	strong	no

x = {overcast, cool, high, strong}
α = 1

P(yes) = 10/16	P(no) = 6/16
P(overcast|yes) = 5/12	P(overcast|no) = 1/8
P(cool|yes) = 4/12	P(cool|no) = 2/8
P(high|yes) = 4/11	P(high|no) = 5/7
P(strong|yes) = 4/11	P(strong|no) = 4/7
0.011 or 1%	0.0047 or 0.47%

Implementation