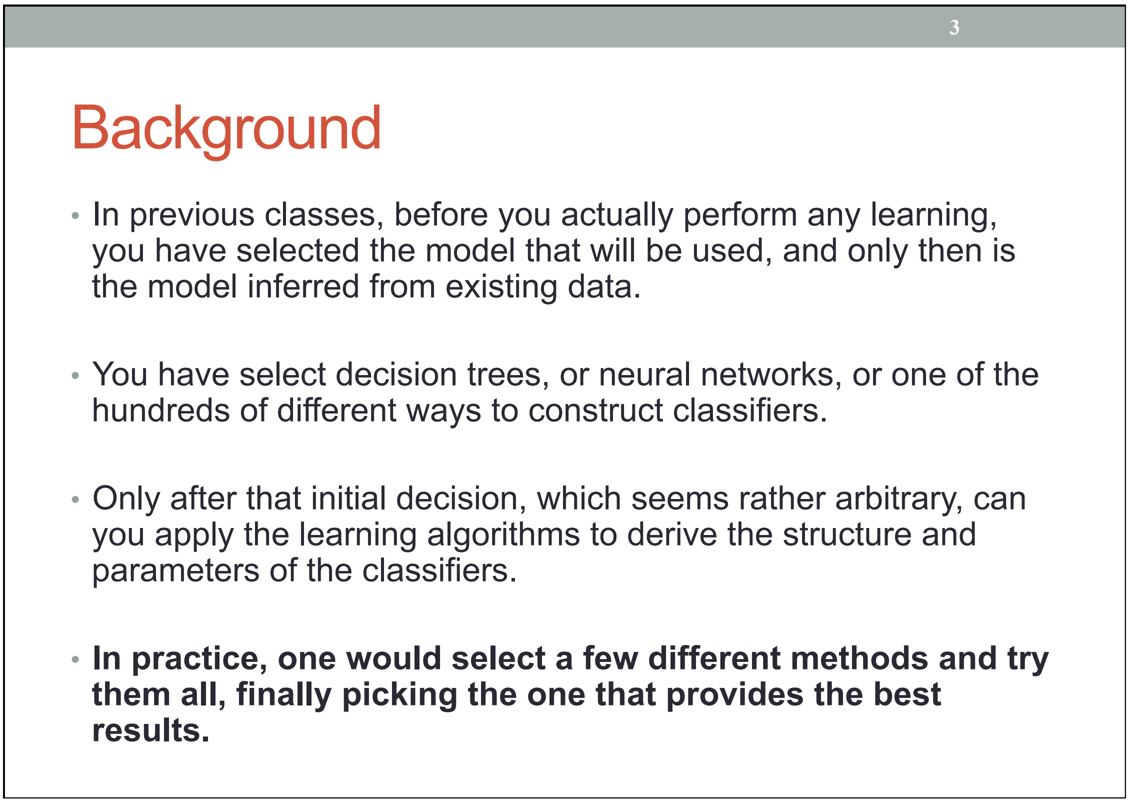
NAÏVE BAYES CLASSIFIER

Ana Teresa Freitas
Adapted from "Digital Minds", Arlindo Oliveira

Computational Biology
2015/2016

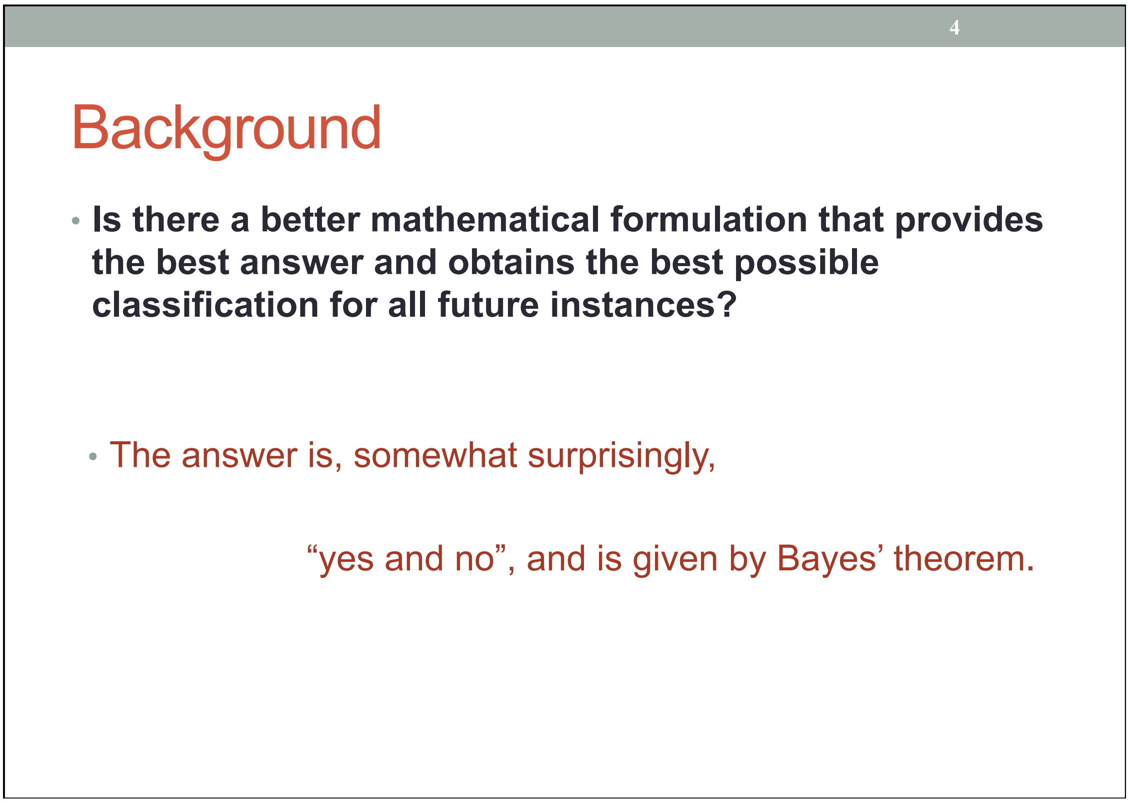
Outline

- Background
- Probability Basics
- Bayes' Theorem
- Naïve Bayes
  - Principle and Algorithm
  - Example: Play Tennis
- Relevant Issues



Background

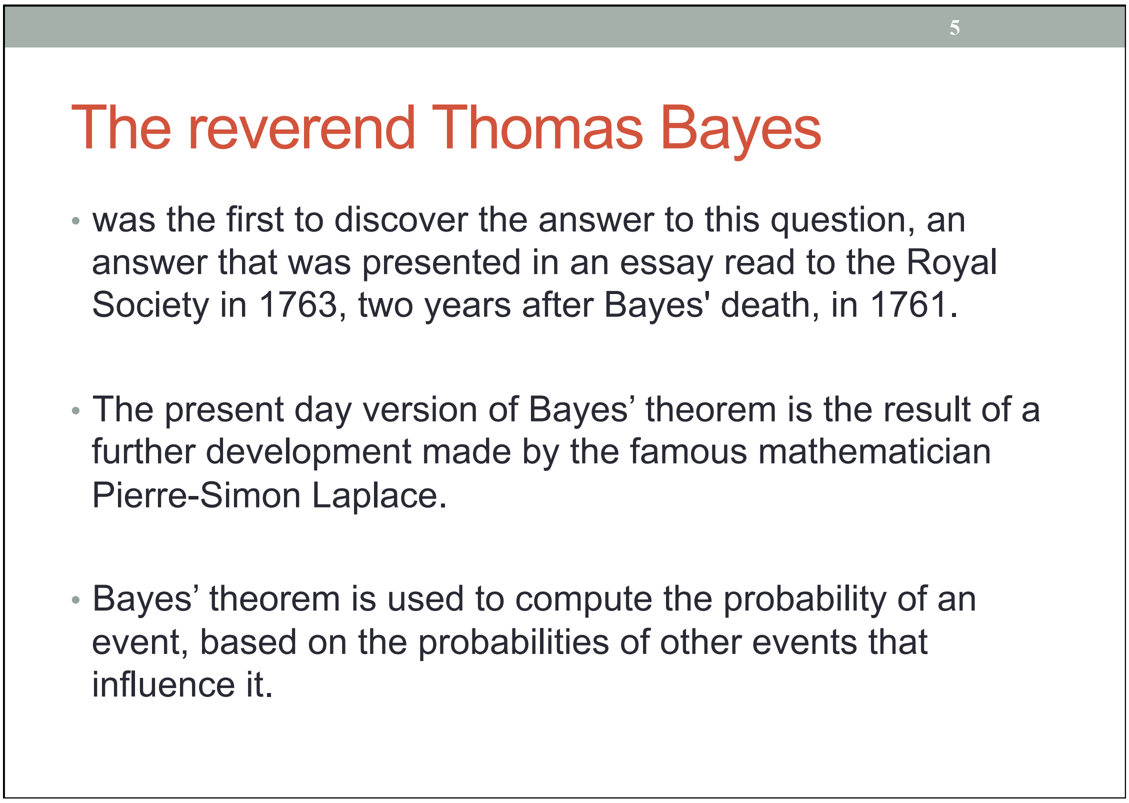
- In previous classes, before you actually perform any learning, you have selected the model that will be used, and only then is the model inferred from existing data.
- You have select decision trees, or neural networks, or one of the hundreds of different ways to construct classifiers.
- Only after that initial decision, which seems rather arbitrary, can you apply the learning algorithms to derive the structure and parameters of the classifiers.
- **In practice, one would select a few different methods and try them all, finally picking the one that provides the best results.**



Background

- **Is there a better mathematical formulation that provides the best answer and obtains the best possible classification for all future instances?**
  - The answer is, somewhat surprisingly,

“yes and no”, and is given by Bayes’ theorem.

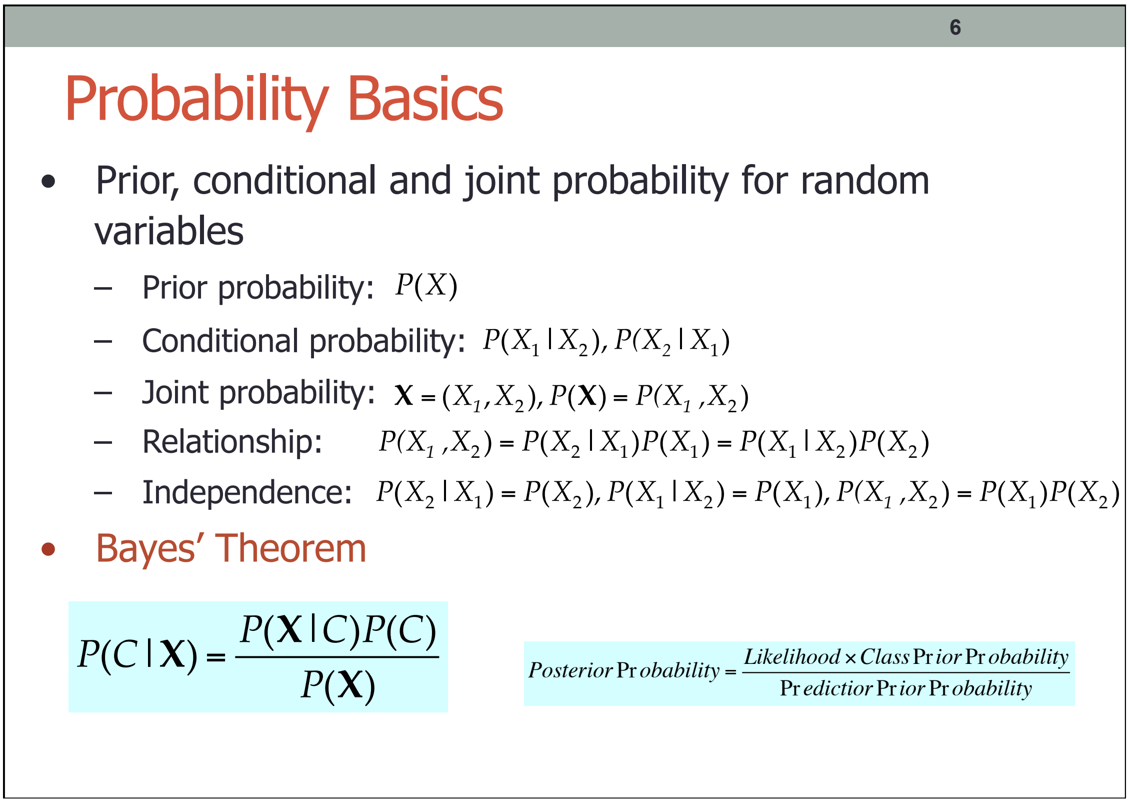


The reverend Thomas Bayes

- was the first to discover the answer to this question, an answer that was presented in an essay read to the Royal Society in 1763, two years after Bayes' death, in 1761.

- The present day version of Bayes’ theorem is the result of a further development made by the famous mathematician Pierre-Simon Laplace.

- Bayes’ theorem is used to compute the probability of an event, based on the probabilities of other events that influence it.



Probability Basics

- Prior, conditional and joint probability for random variables
  - Prior probability: $P(X)$
  - Conditional probability: $P(X_1 | X_2), P(X_2 | X_1)$
  - Joint probability: $\mathbf{X} = (X_1, X_2), P(\mathbf{X}) = P(X_1, X_2)$
  - Relationship: $P(X_1, X_2) = P(X_2 | X_1)P(X_1) = P(X_1 | X_2)P(X_2)$
  - Independence: $P(X_2 | X_1) = P(X_2), P(X_1 | X_2) = P(X_1), P(X_1, X_2) = P(X_1)P(X_2)$
- Bayes’ Theorem

$$P(C | \mathbf{X}) = \frac{P(\mathbf{X} | C)P(C)}{P(\mathbf{X})}$$

$$\text{Posterior Probability} = \frac{\text{Likelihood} \times \text{Class Prior Probability}}{\text{Predictior Prior Probability}}$$

Bayes' Theorem

HIV global prevalence = 0,008
Test with 95% Specificity and Sensitivity

P(T|HIV) = 95%
P(~T|~HIV) = 95%

Perform a first and the result is positive.
What is the probability of having HIV?

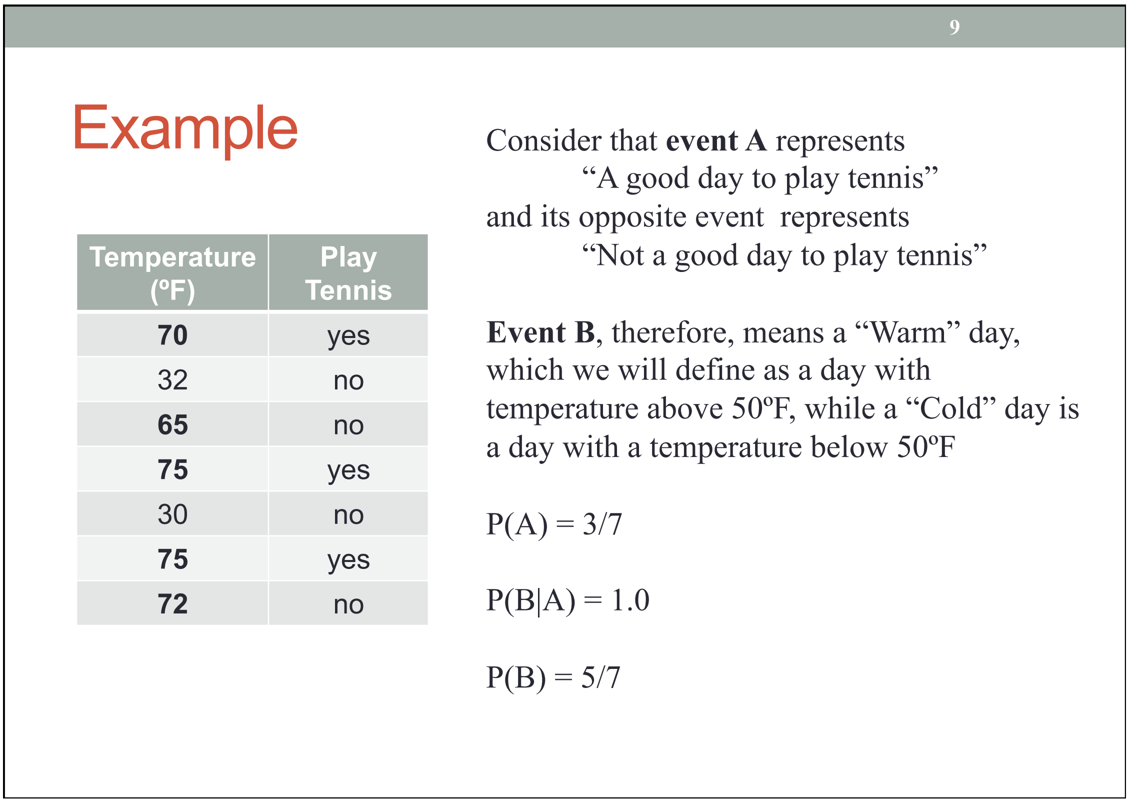
Bayes' Theorem

P(HIV|T) ≈ P(T|HIV) x P(HIV) = 0,95 x 0,008 = 0,0076
P(~HIV|T) ≈ P(T|~HIV) x P(~HIV) = 0,05 x 0,992 = 0,0496 (6,5x)

HIV global prevalence = 0,008
Test with 99% Specificity and Sensitivity

P(T|HIV) = 99%
P(~T|~HIV) = 99%

P(HIV|T) ≈ P(T|HIV) x P(HIV) = 0,99 x 0,008 = 0,00792
P(~HIV|T) ≈ P(T|~HIV) x P(~HIV) = 0,01 x 0,992 = 0,00992



Example

Temperature (ºF)	Play Tennis
70	yes
32	no
65	no
75	yes
30	no
75	yes
72	no

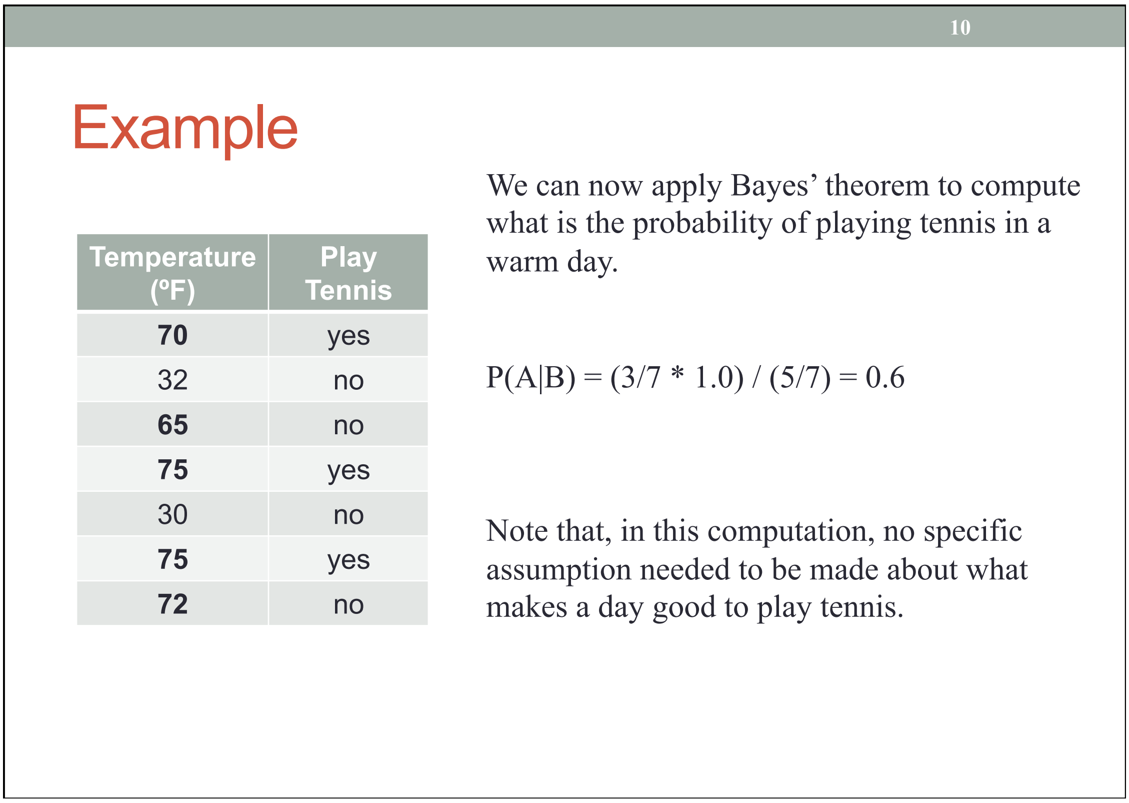
Consider that **event A** represents
“A good day to play tennis”
and its opposite event represents
“Not a good day to play tennis”

Event B, therefore, means a “Warm” day, which we will define as a day with temperature above 50ºF, while a “Cold” day is a day with a temperature below 50ºF

P(A) = 3/7

P(B|A) = 1.0

P(B) = 5/7



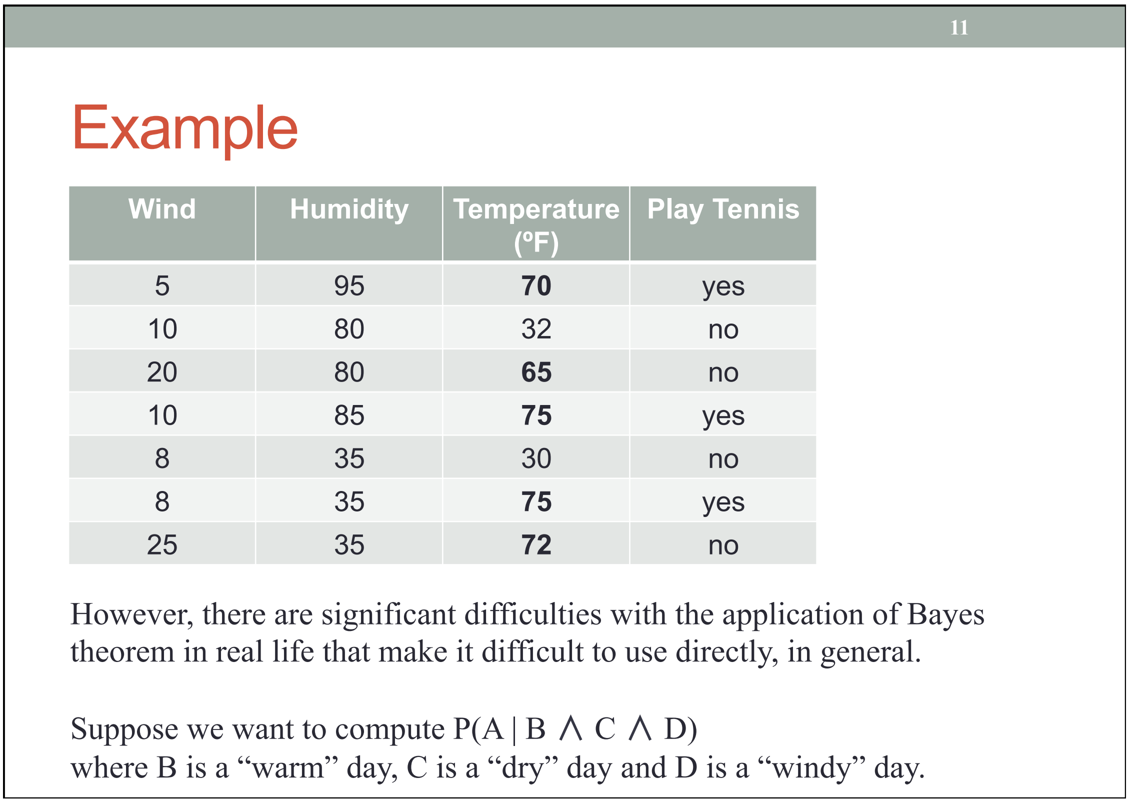
Example

Temperature (ºF)	Play Tennis
70	yes
32	no
65	no
75	yes
30	no
75	yes
72	no

We can now apply Bayes’ theorem to compute what is the probability of playing tennis in a warm day.

P(A|B) = (3/7 * 1.0) / (5/7) = 0.6

Note that, in this computation, no specific assumption needed to be made about what makes a day good to play tennis.

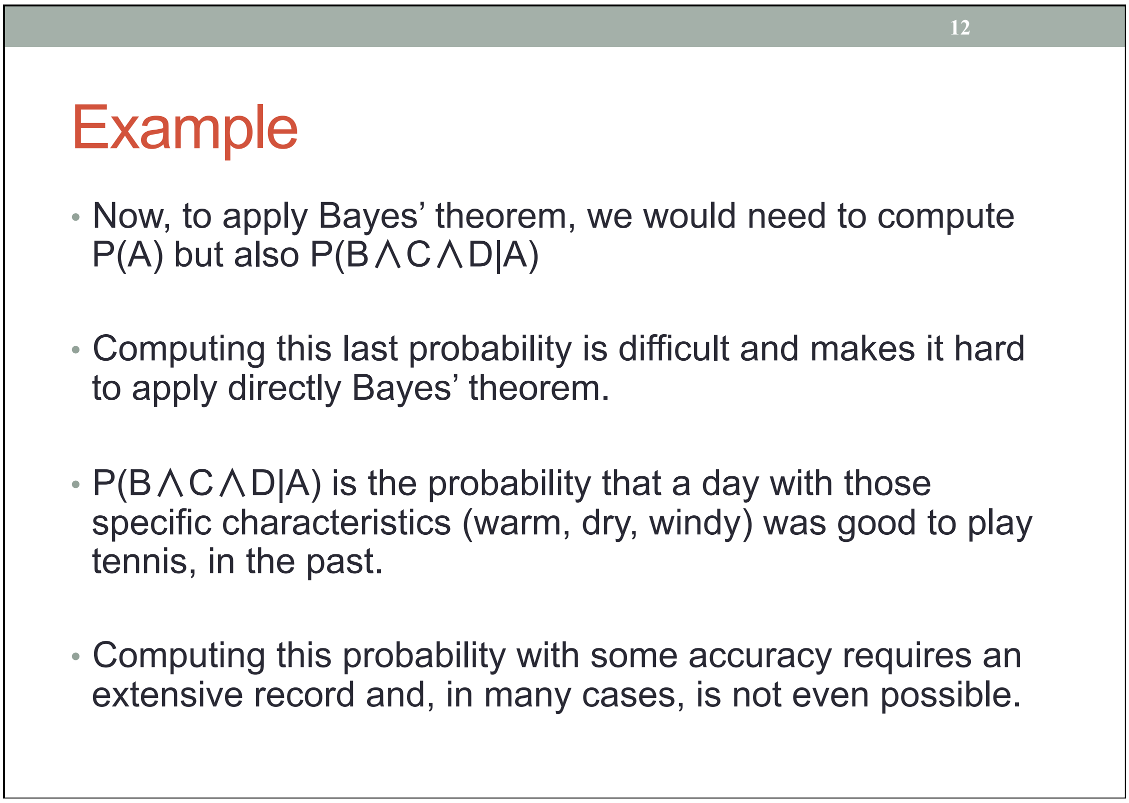


Example

Wind	Humidity	Temperature (ºF)	Play Tennis
5	95	**70**	yes
10	80	32	no
20	80	**65**	no
10	85	**75**	yes
8	35	30	no
8	35	**75**	yes
25	35	**72**	no

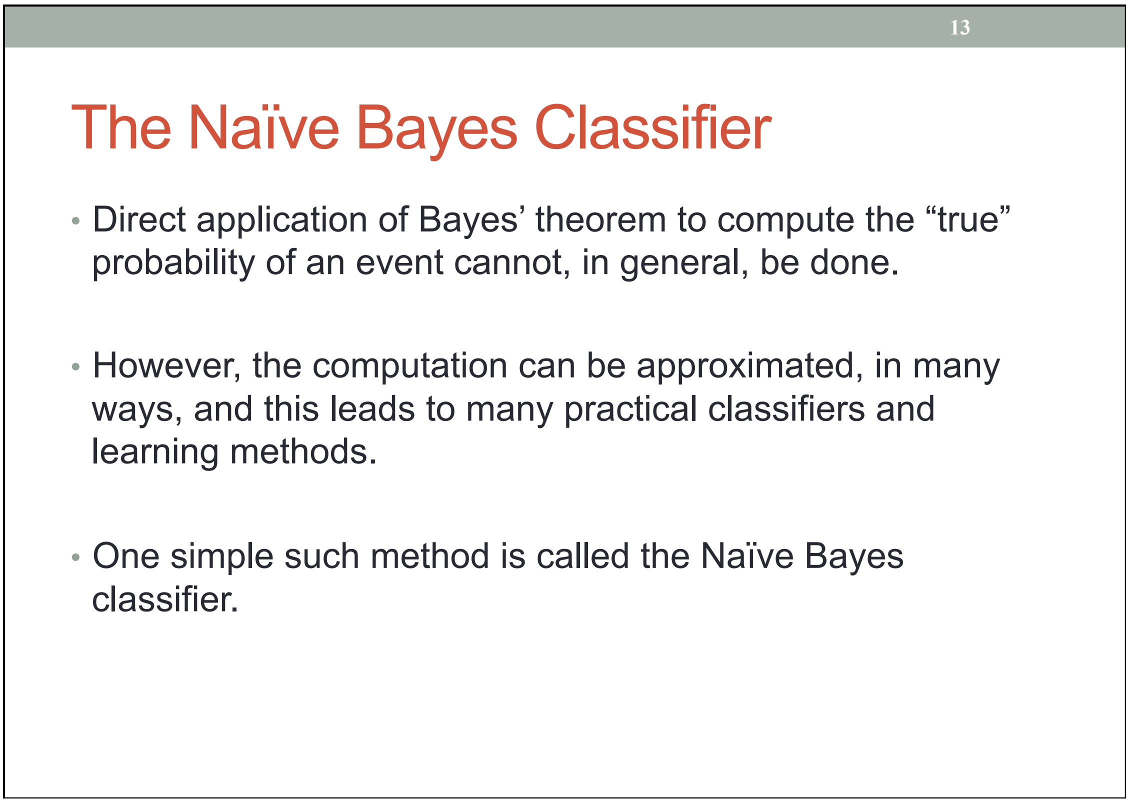
However, there are significant difficulties with the application of Bayes theorem in real life that make it difficult to use directly, in general.

Suppose we want to compute P(A | B ∧ C ∧ D)
where B is a “warm” day, C is a “dry” day and D is a “windy” day.



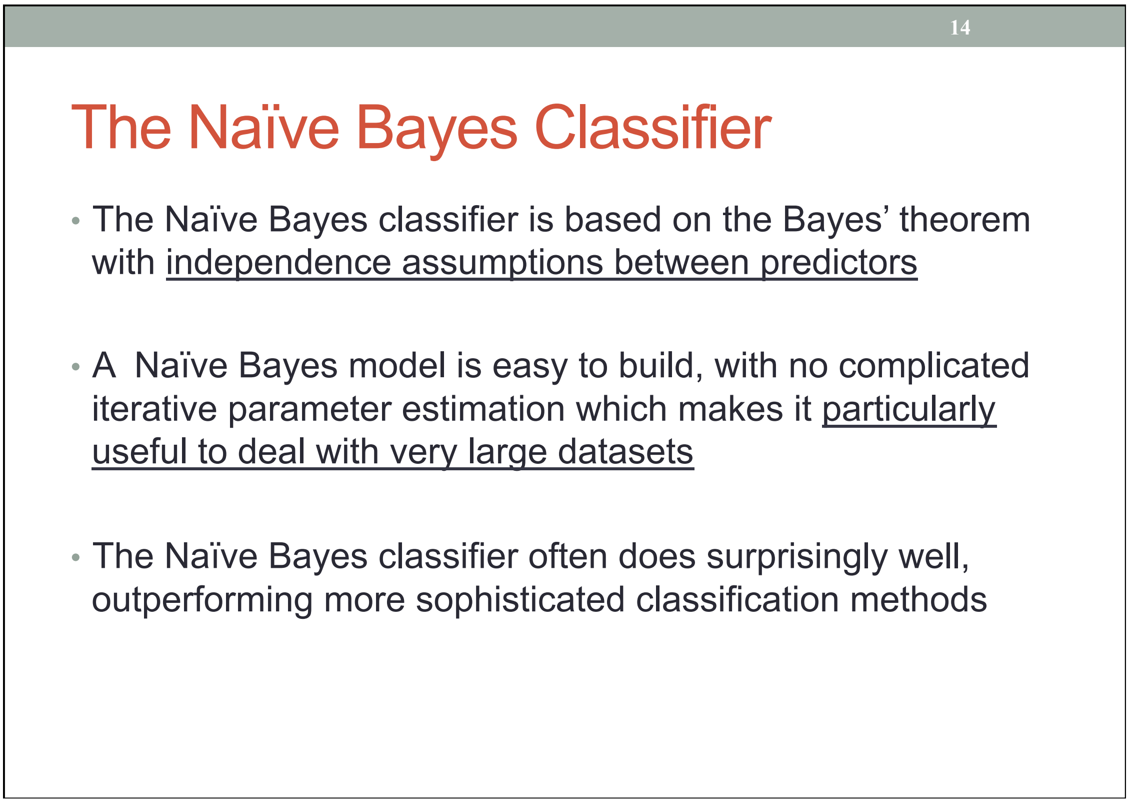
Example

- Now, to apply Bayes’ theorem, we would need to compute P(A) but also P(B∧C∧D|A)
- Computing this last probability is difficult and makes it hard to apply directly Bayes’ theorem.
- P(B∧C∧D|A) is the probability that a day with those specific characteristics (warm, dry, windy) was good to play tennis, in the past.
- Computing this probability with some accuracy requires an extensive record and, in many cases, is not even possible.



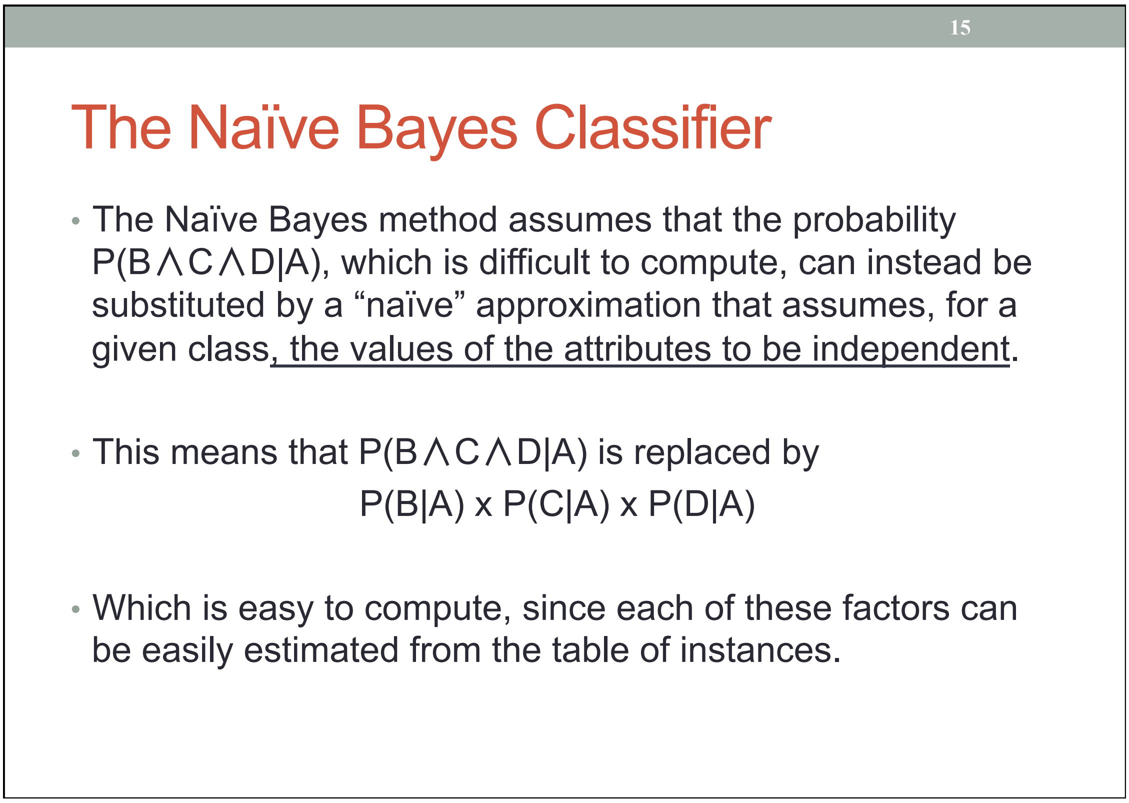
The Naïve Bayes Classifier

- Direct application of Bayes' theorem to compute the "true" probability of an event cannot, in general, be done.
- However, the computation can be approximated, in many ways, and this leads to many practical classifiers and learning methods.
- One simple such method is called the Naïve Bayes classifier.



The Naïve Bayes Classifier

- The Naïve Bayes classifier is based on the Bayes' theorem with independence assumptions between predictors
- A Naïve Bayes model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful to deal with very large datasets
- The Naïve Bayes classifier often does surprisingly well, outperforming more sophisticated classification methods



15

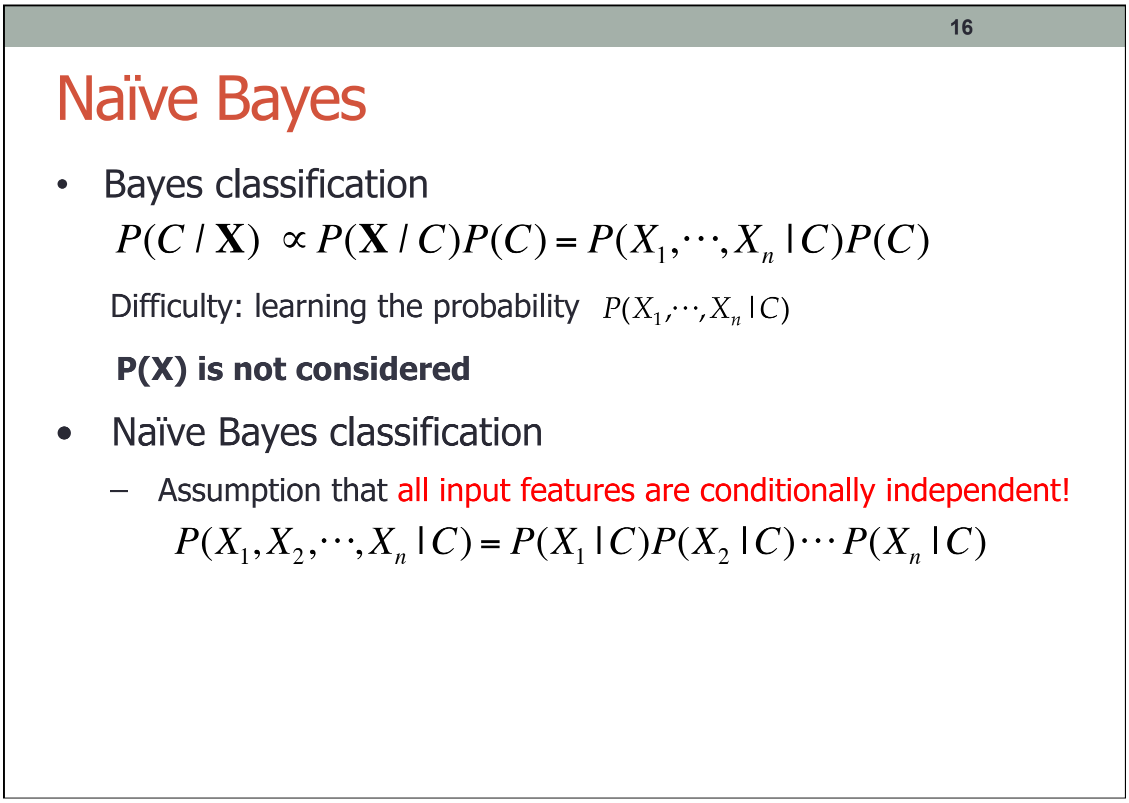
The Naïve Bayes Classifier

- The Naïve Bayes method assumes that the probability P(B∧C∧D|A), which is difficult to compute, can instead be substituted by a "naïve" approximation that assumes, for a given class, the values of the attributes to be independent.

- This means that P(B∧C∧D|A) is replaced by

  P(B|A) x P(C|A) x P(D|A)

- Which is easy to compute, since each of these factors can be easily estimated from the table of instances.



16

Naïve Bayes

- Bayes classification

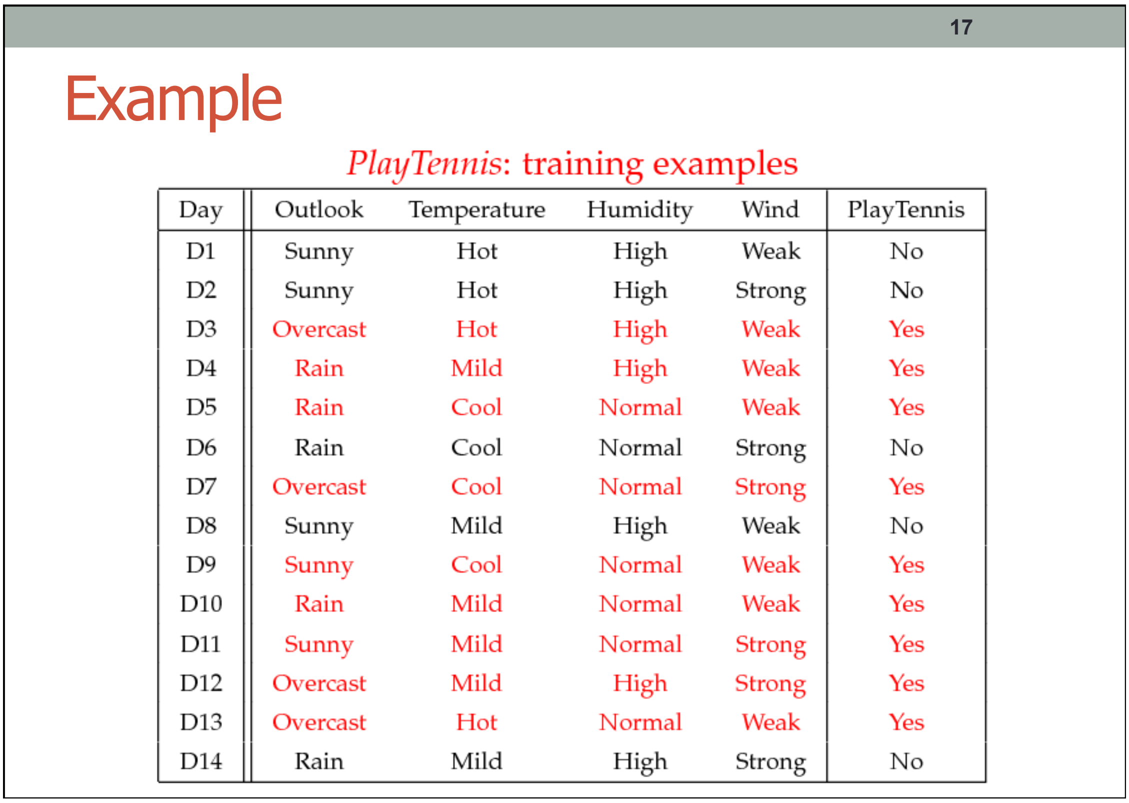
  $$P(C\,/\,\mathbf{X}) \propto P(\mathbf{X}\,/\,C)P(C) = P(X_1,\cdots,X_n\,|\,C)P(C)$$

  Difficulty: learning the probability $P(X_1,\cdots,X_n\,|\,C)$

  P(X) is not considered

- Naïve Bayes classification
  - Assumption that all input features are conditionally independent!

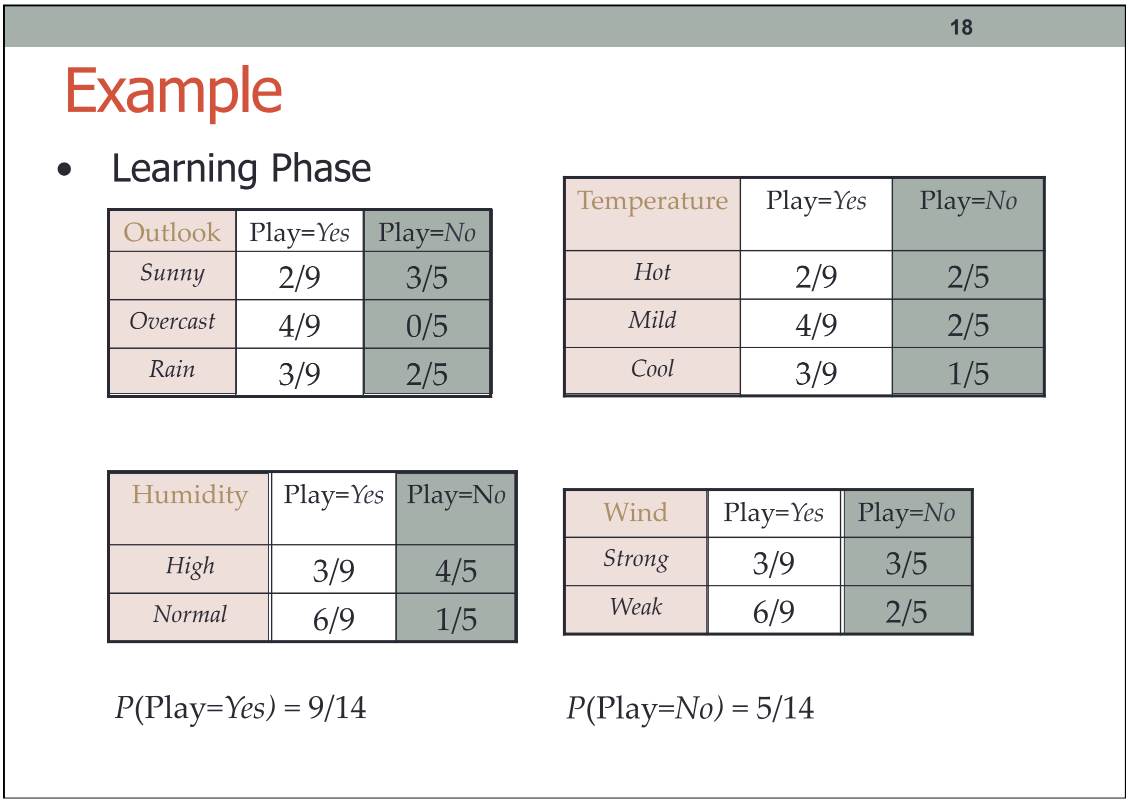
    $$P(X_1,X_2,\cdots,X_n\,|\,C) = P(X_1\,|\,C)P(X_2\,|\,C)\cdots P(X_n\,|\,C)$$



Example

PlayTennis: training examples

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

- Learning Phase

Outlook	Play=*Yes*	Play=*No*
Sunny	2/9	3/5
Overcast	4/9	0/5
Rain	3/9	2/5

Temperature	Play=*Yes*	Play=*No*
Hot	2/9	2/5
Mild	4/9	2/5
Cool	3/9	1/5

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

Wind	Play=*Yes*	Play=*No*
Strong	3/9	3/5
Weak	6/9	2/5

P(Play=*Yes*) = 9/14

P(Play=*No*) = 5/14

Example

- Test Phase
  - Given a new instance, predict its label

    $\mathbf{x}'=(\text{Outlook}=\textit{Sunny}, \text{Temperature}=\textit{Cool}, \text{Humidity}=\textit{High}, \text{Wind}=\textit{Strong})$
  - Look up tables achieved in the learning phrase

$P(\text{Outlook}=\textit{Sunny} \mid \text{Play}=\textit{Yes}) = 2/9$	$P(\text{Outlook}=\textit{Sunny} \mid \text{Play}=\textit{No}) = 3/5$
$P(\text{Temperature}=\textit{Cool} \mid \text{Play}=\textit{Yes}) = 3/9$	$P(\text{Temperature}=\textit{Cool} \mid \text{Play}==\textit{No}) = 1/5$
$P(\text{Huminity}=\textit{High} \mid \text{Play}=\textit{Yes}) = 3/9$	$P(\text{Huminity}=\textit{High} \mid \text{Play}=\textit{No}) = 4/5$
$P(\text{Wind}=\textit{Strong} \mid \text{Play}=\textit{Yes}) = 3/9$	$P(\text{Wind}=\textit{Strong} \mid \text{Play}=\textit{No}) = 3/5$
$P(\text{Play}=\textit{Yes}) = 9/14$	$P(\text{Play}=\textit{No}) = 5/14$

  - Decision making

$$P(\textit{Yes} \mid \mathbf{x}') \approx [P(\textit{Sunny} \mid \text{Yes})P(\textit{Cool} \mid \textit{Yes})P(\textit{High} \mid \text{Yes})P(\textit{Strong} \mid \textit{Yes})]P(\text{Play}=\textit{Yes}) = 0.0053$$

$$P(\textit{No} \mid \mathbf{x}') \approx [P(\textit{Sunny} \mid \text{No})\ P(\textit{Cool} \mid \text{No})P(\textit{High} \mid \textit{No})P(\textit{Strong} \mid \textit{No})]P(\text{Play}=\textit{No}) = 0.0206$$

Given the fact $P(\textit{Yes} \mid \mathbf{x}') < P(\textit{No} \mid \mathbf{x}')$, we label $\mathbf{x}'$ to be "*No*".

Naïve Bayes – HIV example

HIV global prevalence = 0,008

Test with 95% Specificity and Sensitivity

P(T|HIV) = 95%

P(~T|~HIV) = 95%

Perform a first and the result is positive.

Perform a second different and independent test with the same Sensitivity and Specificity. The result is positive.

What is the probability of having HIV?

Naïve Bayes – HIV example

P(HIV|T1,T2) ≈ P(T1|HIV) x P(T2|HIV) x P(HIV)
= 0,95 x 0,95 x 0,008 = 0,00722 (2,9x)

P(~HIV|T1,T2) ≈ P(T1|~HIV) x P(T2|~HIV) x P(~HIV)
= 0,05 x 0,05 x 0,992 = 0,00248

Naïve Bayes

- Algorithm: Continuous-valued Features
  - Numberless values for a feature
  - Conditional probability often modeled with the normal distribution

$$\hat{P}(X_j \mid C = c_i) = \frac{1}{\sqrt{2\pi}\sigma_{ji}} \exp\left(-\frac{(X_j - \mu_{ji})^2}{2\sigma_{ji}^2} \right)$$

μ_{ji} : mean (avearage) of feature values X_j of examples for which $\mathrm{C} = c_i$

σ_{ji} : standard deviation of feature values X_j of examples for which $C = c_i$

  - Learning Phase: for $\mathbf{X} = (X_1, \cdots, X_n),\ C = c_1, \cdots, c_L$
    Output: $n \times L$ normal distributions and $P(C = c_i)\ i = 1, \cdots, L$
  - Test Phase: Given an unknown instance $\mathbf{X}' = (a'_1, \cdots, a'_n)$
    - Instead of looking-up tables, calculate conditional probabilities with all the normal distributions achieved in the learning phrase

Naïve Bayes

- Example: Continuous-valued Features
  - Temperature is naturally of continuous value.

    Yes: 25.2, 19.3, 18.5, 21.7, 20.1, 24.3, 22.8, 23.1, 19.8

    No: 27.3, 30.1, 17.4, 29.5, 15.1
  - Estimate mean and variance for each class

$$\mu = \frac{1}{N}\sum_{n=1}^{N} x_n, \quad \sigma^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)^2$$

$$\mu_{Yes} = 21.64, \ \sigma_{Yes} = 2.35$$
$$\mu_{No} = 23.88, \ \sigma_{No} = 7.09$$

  - **Learning Phase**: output two Gaussian models for P(temp|C)

$$\hat{P}(x \mid Yes) = \frac{1}{2.35\sqrt{2\pi}}\exp\left(-\frac{(x-21.64)^2}{2\times 2.35^2}\right) = \frac{1}{2.35\sqrt{2\pi}}\exp\left(-\frac{(x-21.64)^2}{11.09}\right)$$

$$\hat{P}(x \mid No) = \frac{1}{7.09\sqrt{2\pi}}\exp\left(-\frac{(x-23.88)^2}{2\times 7.09^2}\right) = \frac{1}{7.09\sqrt{2\pi}}\exp\left(-\frac{(x-23.88)^2}{50.25}\right)$$

Relevant Issues

- Violation of Independence Assumption
  - For many real world tasks, $P(X_1,\cdots,X_n|C) \neq P(X_1|C)\cdots P(X_n|C)$
  - Nevertheless, naïve Bayes works surprisingly well anyway!
- Zero conditional probability Problem
  - If no example contains the feature value $X_j = a_{jk}, \hat{P}(X_j = a_{jk}|C = c_i) = 0$
  - In this circumstance, $\hat{P}(x_1|c_i)\cdots\hat{P}(a_{jk}|c_i)\cdots\hat{P}(x_n|c_i) = 0$ during test
  - For a remedy, conditional probabilities re-estimated with

$$\hat{P}(X_j = a_{jk}|C = c_i) = \frac{n_c + mp}{n + m}$$

n_c : number of training examples for which $X_j = a_{jk}$ and $C = c_i$

n : number of training examples for which $C = c_i$

p : prior estimate (usually, $p = 1/t$ for t possible values of X_j)

m : weight to prior (number of "virtual" examples, $m \geq 1$)