Probability, Conditional Probability & Bayes Rule

UNIVERSITY of PENNSYLVANIA

A FAST REVIEW OF DISCRETE PROBABILITY (PART 2)

Discrete random variables

- **A *random variable* can take on one of a set of different values, each with an associated probability. Its value at a particular time is *subject to random variation*.**
  - *Discrete* random variables take on one of a discrete (often finite) range of values
  - Domain values must be *exhaustive* and *mutually exclusive*

- **For us, random variables will have a discrete, countable (usually finite) domain of *arbitrary values*.**
  - Mathematical statistics usually calls these *random elements*
  - **Example: Weather is a discrete random variable with domain {sunny, rain, cloudy, snow}.**
  - ***Example: A Boolean random variable* has the domain {true,false},**

Probability Distribution

- ***Probability distribution*** **gives values for all possible assignments:**
  - Vector notation: Weather is one of <0.72, 0.1, 0.08, 0.1>, where weather is one of <sunny,rain,cloudy,snow>.
  - ***P****(Weather) = <0.72,0.1,0.08,0.1>*
  - Sums to 1 over the domain

—***Practical advice: Easy to check***

—***Practical advice: Important to check***

Factored Representations: Propositions

- ***Elementary proposition*** **constructed by assignment of a value to a random variable:**
  - e.g. *Weather = sunny (*abbreviated as *sunny)*
  - e.g. *Cavity = false* (abbreviated as $\neg$*cavity*)
- ***Complex proposition*** **formed from elementary propositions & standard logical connectives**
  - e.g. *Weather = sunny* $\vee$ *Cavity = false*
- ***We will work with event spaces over such propositions***

A word on notation

Assume *Weather* is a discrete random variable with domain {sunny, rain, cloudy, snow}.

- *Weather = sunny* **abbreviated** *sunny*
- *P(Weather=sunny)=0.72* **abbreviated** *P(sunny)=0.72*

- *Cavity = true* **abbreviated** *cavity*
- *Cavity = false* **abbreviated** *¬cavity*

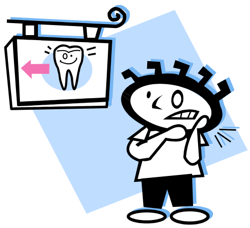
Vector notation:

- **Fix order of domain elements:**
  <sunny,rain,cloudy,snow>
- **Specify the probability mass function (pmf) by a vector:**
  P(Weather) = <0.72,0.1,0.08,0.1>

Joint probability distribution

- **Probability assignment to all combinations of values of random variables (i.e. all elementary events)**

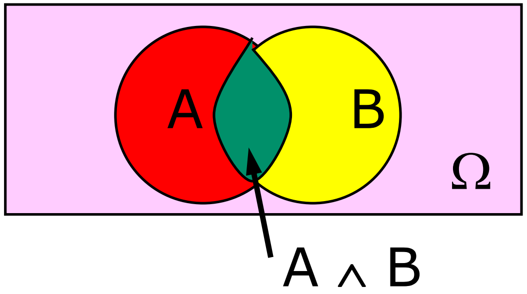
	toothache	¬ toothache
cavity	0.04	0.06
¬ cavity	0.01	0.89



- **The sum of the entries in this table has to be 1**
- ***Every question about a domain can be answered by the joint distribution***
- **Probability of a proposition is the sum of the probabilities of elementary events in which it holds**
  - P(cavity) = 0.1 [marginal of row 1]
  - P(toothache) = 0.05 [marginal of toothache column]

Conditional Probability

	toothache	¬ toothache
cavity	0.04	0.06
¬ cavity	0.01	0.89



- ***P(cavity)=0.1*** **and** ***P(cavity ∧ toothache)=0.04*** **are both** *prior* **(unconditional) probabilities**
- **Once the agent has new evidence concerning a** ***previously unknown*** **random variable, e.g. Toothache, we can specify** *a posterior* **(conditional) probability e.g.** ***P(cavity | Toothache=true)***

$$P(a \mid b) = P(a \wedge b)/P(b)$$

[Probability of a with the Universe Ω restricted to b]

→ The new information restricts the set of possible worlds ω_i consistent with it, so changes the probability.

- ***So P(cavity | toothache) = 0.04/0.05 = 0.8***

Conditional Probability (continued)

- **Definition of Conditional Probability:**
  $P(a \mid b) = P(a \wedge b)/P(b)$

- **Product rule gives an alternative formulation:**
  $$P(a \wedge b) = P(a \mid b) * P(b)$$
  $$= P(b \mid a) * P(a)$$

- **A general version holds for whole distributions:**
  $\mathbf{P}(Weather,Cavity) = \mathbf{P}(Weather \mid Cavity) * \mathbf{P}(Cavity)$

- **Chain rule is derived by successive application of product rule:**
  $$\mathbf{P}(A,B,C,D,E) = P(A|B,C,D,E)\, P(B,C,D,E)$$
  $$= P(A|B,C,D,E)\, P(B|C,D,E)\, P(C,D,E)$$
  $$= \ldots$$
  $$= P(A|B,C,D,E)\, P(B|C,D,E)\, P(C|D,E)\, P(D|E)\, P(E)$$

Probabilistic Inference

- ***Probabilistic inference:*** **the computation**
  - from *observed evidence*
  - of *posterior probabilities*
  - for *query propositions*.
- **We use the *full joint distribution* as the "knowledge base" from which answers to questions may be derived.**
- **Ex: three Boolean variables *Toothache (T), Cavity (C), ShowsOnXRay (X)***

	t		¬t	
	x	¬x	x	¬x
c	0.108	0.012	0.072	0.008
¬c	0.016	0.064	0.144	0.576

- **Probabilities in joint distribution sum to 1**

Probabilistic Inference II

	t		$\neg t$	
	x	$\neg x$	x	$\neg x$
c	0.108	0.012	0.072	0.008
$\neg c$	0.016	0.064	0.144	0.576

- **Probability of any proposition computed by finding atomic events where proposition is true and adding their probabilities**
  - ***P(cavity $\vee$ toothache)***
    = 0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064
    = 0.28
  - ***P(cavity)***
    = 0.108 + 0.012 + 0.072 + 0.008
    = 0.2
- ***P(cavity)*** **is called a** ***marginal probability*** **and the process of computing this is called** ***marginalization***

Probabilistic Inference III

	t		¬t	
	x	¬x	x	¬ x
c	0.108	0.012	0.072	0.008
¬c	0.016	0.064	0.144	0.576

- **Can also compute conditional probabilities.**
- ***P(¬ cavity | toothache)***
  = P(¬ cavity ∧ toothache)/P(toothache)
  = (0.016 + 0.064) / (0.108 + 0.012 + 0.016 + 0.064)
  = 0.4
- **Denominator is viewed as a** *normalization constant***:**
  - Stays constant no matter what the value of Cavity is.
    (Book uses α to denote normalization constant 1/P(X), for random variable X.)

Bayes Rule & Naïve Bayes

(some slides adapted from slides by Massimo Poesio, adapted from slides by Chris Manning)

Bayes' Rule & Diagnosis

$$\underbrace{P(a|b)}_{\text{Posterior}} = \frac{\overbrace{P(b|a)}^{\text{Likelihood}} * \overbrace{P(a)}^{\text{Prior}}}{\underbrace{P(b)}_{\text{Normalization}}}$$

- **Useful for assessing diagnostic probability from causal probability:**

$$P(Cause|Effect) = \frac{P(Effect|Cause) * P(Cause)}{P(Effect)}$$

Bayes' Rule For Diagnosis II

$$P(\text{Disease} \mid \text{Symptom}) = \frac{P(\text{Symptom} \mid \text{Disease}) * P(\text{Disease})}{P(\text{Symptom})}$$

Imagine:

- **disease = TB, symptom = coughing**
- *P(disease | symptom)* **is different in TB-indicated country vs. USA**
- *P(symptom | disease)* **should be the same**
  - It is more widely useful to learn *P(symptom | disease)*
- **What about P(symptom)?**
  - Use *conditioning* (next slide)
  - For determining, e.g., the *most likely* disease given the symptom, we can just ignore P(symptom)!!! (see slide 35)

Conditioning

- *Idea:* **Use *conditional probabilities* instead of joint probabilities**
- $P(a) = P(a \wedge b) + P(a \wedge \neg b)$
  $= P(a \mid b) * P(b) + P(a \mid \neg b) * P(\neg b)$

  Here:

  $P(symptom) = P(symptom \mid disease) * P(disease) + P(symptom \mid \neg disease) * P(\neg disease)$

- **More generally: $P(Y) = \sum_z P(Y|z) * P(z)$**
- **Marginalization and conditioning are useful rules for derivations involving probability expressions.**

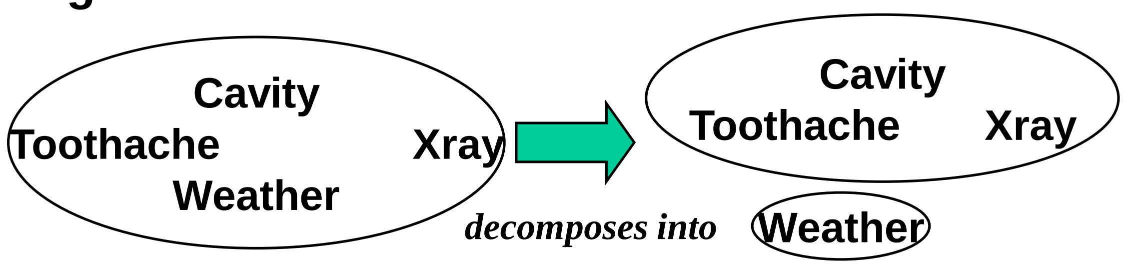
Exponentials rear their ugly head again…

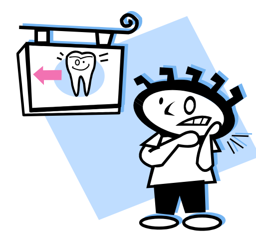
- **Estimating the necessary joint probability distribution for many symptoms is infeasible**
  - For $|D|$ diseases, $|S|$ symptoms where a person can have n of the diseases and m of the symptoms
    - $P(s|d_1, d_2, \ldots, d_n)$ requires $|S|\ |D|^n$ values
    - $P(s_1, s_2, \ldots, s_m)$ requires $|S|^m$ values

- **These numbers get big fast**
  - If $|S| = 1{,}000$, $|D| = 100$, $n=4$, $m=7$
    - $P(s|d_1, \ldots d_n)$ requires $1000 * 100^4 = 10^{11}$ values (-1)
    - $P(s_1 \ldots s_m)$ requires $1000^7 = 10^{21}$ values (-1)

The Solution: *Independence*

- **Random variables A and B are *independent* iff**
  - ***$P(A \wedge B) = P(A) * P(B)$***
  - ***equivalently: $P(A \mid B) = P(A)$ and $P(B \mid A) = P(B)$***
- ***A and B are independent if knowing whether A occurred gives no information about B (and vice versa)***
- **Independence assumptions are *essential* for efficient probabilistic reasoning**





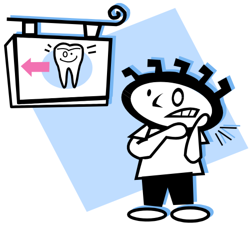
$$P(T, X, C, W) = P(T, X, C) * P(W)$$

- **15 entries (2^4-1) reduced to 8 (2^3-1 + 2-1)**
  For *n independent* biased coins, *$O(2^n)$ entries* $\rightarrow$*$O(n)$*

Conditional Independence

- **BUT *absolute* independence is rare**
- **Dentistry is a large field with hundreds of variables, none of which are independent. What to do?**

- **A and B are *conditionally independent given C* iff**
  - $P(A \mid B, C) = P(A \mid C)$
  - $P(B \mid A, C) = P(B \mid C)$
  - $P(A \wedge B \mid C) = P(A \mid C) * P(B \mid C)$



- **Toothache (T), Spot in Xray (X), Cavity (C)**
  - None of these are independent of the other two
  - But ***T and X are conditionally independent given C***

Conditional Independence II *WHY??*

- **If I have a cavity, the probability that the XRay shows a spot doesn't depend on whether I have a toothache (and vice versa):**

  $$P(X|T,C) = P(X|C)$$

- **From which follows:**

  $$P(T|X,C) = P(T|C) \text{ and } P(T,X|C) = P(T|C) * P(X|C)$$

- **By the chain rule), given conditional independence:**

  $$P(T,X,C) = P(T|X,C) * P(X,C) = P(T|X,C) * P(X|C) * P(C)$$

  $$= P(T|C) * P(X|C) * P(C)$$

- P(***Toothache, Cavity, Xray***) **has $2^3 - 1 = 7$ independent entries**
- **Given conditional independence, chain rule yields**
  $2 + 2 + 1 = 5$ independent numbers

Conditional Independence III

- **In most cases, the use of conditional independence reduces the size of the representation of the joint distribution from *exponential* in *n* to *linear* in *n*.**

- ***Conditional independence is our most basic and robust form of knowledge about uncertain environments.***

Another Example

- **Battery is dead (B)**
- **Radio plays (R)**
- **Starter turns over (S)**
- **None of these propositions are independent of one another**
- ***BUT: R and S are conditionally independent given B***

Naïve Bayes I

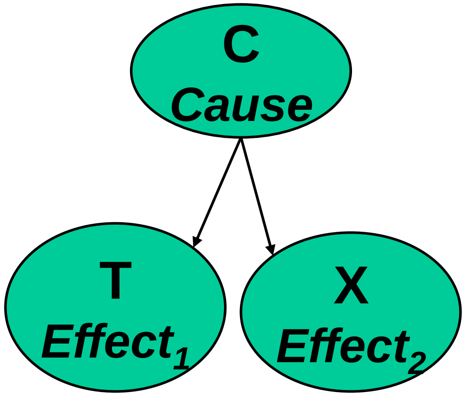
By Bayes Rule $$P(C|T,X) = \frac{P(T,X|C)P(C)}{P(T,X)}$$

If T and X are *conditionally independent given C*:

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

This is a *Naïve Bayes Model:*

All effects assumed conditionally independent given Cause

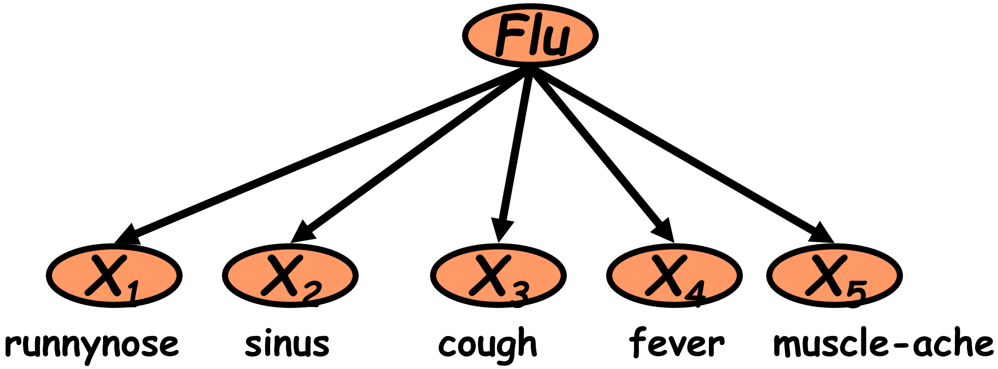


Bayes' Rule II

- **More generally**

$$P(Cause, Effect_1, \ldots, Effect_{n)} = P(Cause)\prod_i P(Effect_i \mid Cause)$$

- **Total number of parameters is *linear* in *n***



An Early Robust Statistical NLP Application

- A Statistical Model For Etymology (Church ’85)

- Determining etymology is crucial for text-to-speech

Italian	English
AldriGHetti	**lauGH, siGH**
IannuCCi	aCCept
ItaliAno	hAte

An Early Robust Statistical NLP Application

Angeletti	100%	Italian
Iannucci	100%	Italian
Italiano	100%	Italian
Lombardino	58%	Italian
Asahara	100%	Japanese
Fujimaki	100%	Japanese
Umeda	96%	Japanese
Anagnostopoulos	100%	Greek
Demetriadis	100%	Greek
Dukakis	99%	Russian
Annette	75%	French
Deneuve	54%	French
Baguenard	54%	Middle French

- A very simple statistical model (your next homework) solved the problem, despite a wild statistical assumption

Computing the Normalizing Constant *P(T,X)*

$$P(c|T,X) + P(\neg c|T,X) = 1$$

$$\frac{P(T|c)P(X|c)P(c)}{P(T,X)} + \frac{P(T|\neg c)P(X|\neg c)P(\neg c)}{P(T,X)} = 1$$

$$P(T|c)P(X|c)P(c) + P(T|\neg c)P(X|\neg c)P(\neg c) = P(T,X)$$

IF THERE'S TIME…..

BUILDING A SPAM FILTER USING NAÏVE BAYES

Spam or not Spam: that is the question.

From: "" <takworlld@hotmail.com>

Subject: real estate is the only way... gem oalvgkay

Anyone can buy real estate with no money down

Stop paying rent TODAY !

There is no need to spend hundreds or even thousands for similar courses

I am 22 years old and I have already purchased 6 properties using the methods outlined in this truly INCREDIBLE ebook.

Change your life NOW !

===

Click Below to order:

http://www.wholesaledaily.com/sales/nmd.htm

===

Categorization/Classification Problems

- **Given:**
  - A description of an instance, $x \in X$, where X is the *instance language* or *instance space*.
    - —*(Issue: how do we represent text documents?)*
  - *A fixed set of categories:*

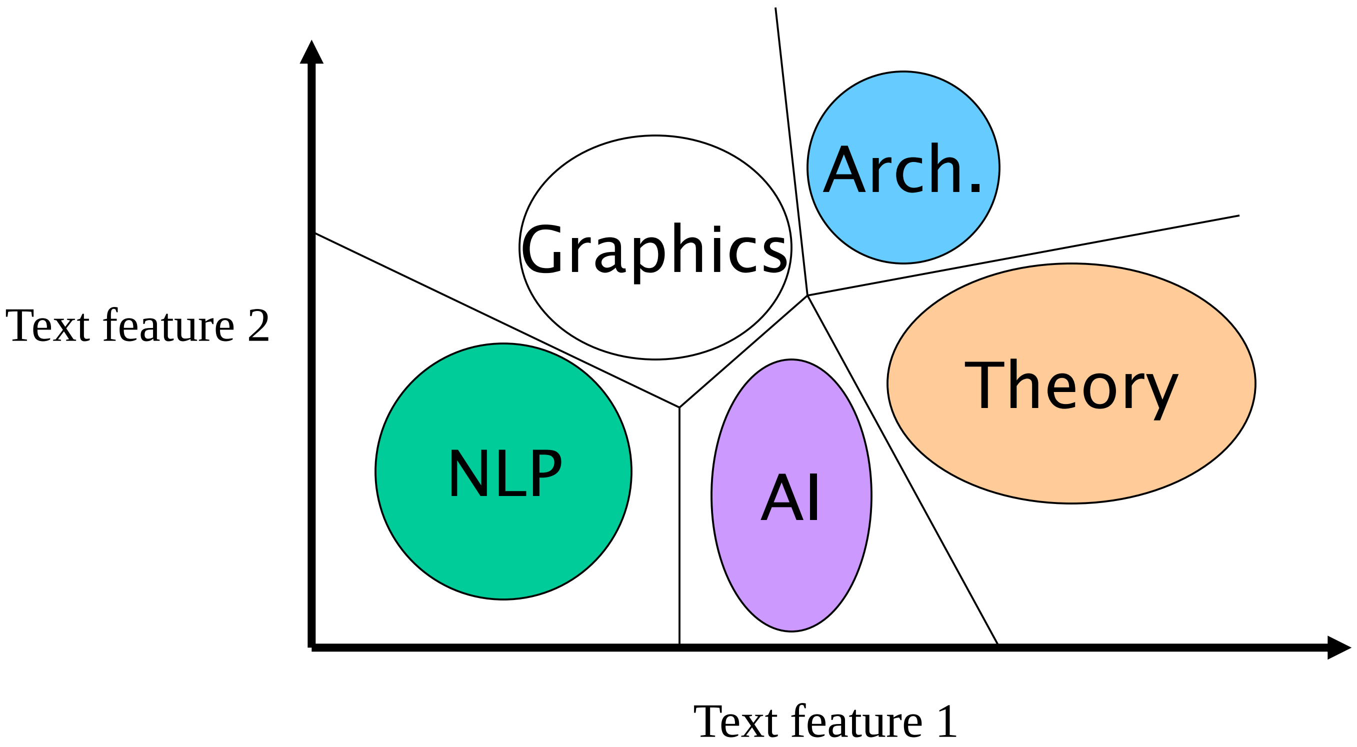
    $C = \{c_1, c_2, \ldots, c_n\}$

- **Determine:**
  - The category of x: $c(x) \in C$, where c(x) is a categorization function whose domain is X and whose range is C.
    - —*We want to automatically build categorization functions ("classifiers").*

EXAMPLES OF TEXT CATEGORIZATION

- **Categories = SPAM?**
  - “spam” / “not spam”
- **Categories = TOPICS**
  - “finance” / “sports” / “asia”
- **Categories = OPINION**
  - “like” / “hate” / “neutral”
- **Categories = AUTHOR**
  - “Shakespeare” / “Marlowe” / “Ben Jonson”
  - The Federalist papers

A Graphical View of Text Classification



Bayesian Methods for Text Classification

- **Uses *Bayes theorem* to build a *generative Naïve Bayes model* that approximates how data is produced**

$$P(C \mid D) = \frac{P(D \mid C)P(C)}{P(D)}$$

Where C: Categories, D: Documents

- **Uses *prior probability* of each category given *no* information about an item.**

- **Categorization produces a *posterior probability* distribution over the possible categories given a description of each document.**

Maximum a posteriori (MAP) Hypothesis

- **Goodbye to that nasty normalization constant!!**

$$c_{MAP} \equiv \underset{c \in C}{\operatorname{argmax}}\, P(c \mid D)$$

$$= \underset{c \in C}{\operatorname{argmax}}\, \frac{P(D \mid c)P(c)}{P(D)}$$

$$= \underset{c \in C}{\operatorname{argmax}}\, P(D \mid c)P(c)$$

No need to compute α, here *P(D)!!!!*

As *P(D)* is constant

Maximum likelihood Hypothesis

If all hypotheses are a priori equally likely, we only need to consider the $P(D|c)$ term:

$$c_{ML} \equiv \operatorname*{argmax}_{c \in C} P(D \mid c)$$

Maximum Likelihood Estimate ("MLE")

Naive Bayes Classifiers

Task: Classify a new instance *D* based on a tuple of attribute values $D = \langle x_1, x_2, \ldots, x_n \rangle$ into one of the classes $c_j \in C$

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}}\ P(c \mid x_1, x_2, \ldots, x_n)$$

$$= \underset{c \in C}{\operatorname{argmax}}\ \frac{P(x_1, x_2, \ldots, x_n \mid c) P(c)}{P(x_1, x_2, \ldots, x_n)}$$

$$= \underset{c \in C}{\operatorname{argmax}}\ P(x_1, x_2, \ldots, x_n \mid c) P(c)$$

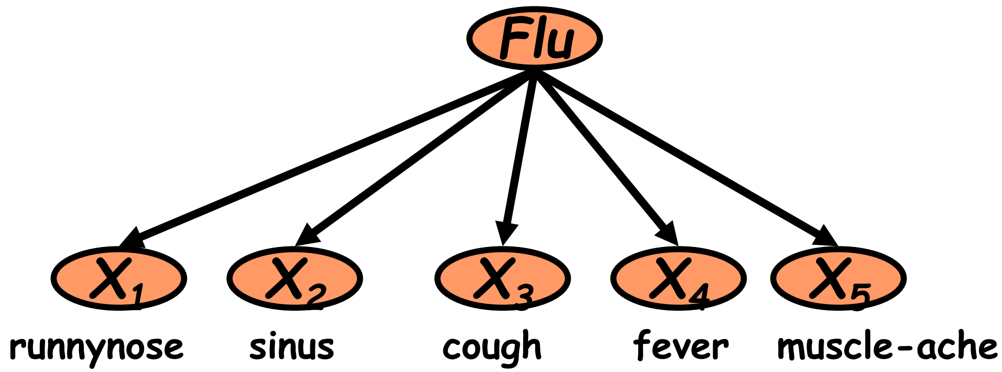
Naïve Bayes Classifier: Assumption

- $P(c_j)$
  - Can be estimated from the frequency of classes in the training examples.
- $P(x_1,x_2,\ldots,x_n|c_j)$
  - Again, O(|*X*|^n•|*C*|) parameters to estimate full joint prob. distribution
  - As we saw, can only be estimated if a ***vast*** number of training examples was available.

Naïve Bayes Conditional Independence Assumption:

$$P(x_i, x_2, \ldots, x_n \mid c_j) = \prod_i P(x_i \mid c_j)$$

The Naïve Bayes Classifier



- Conditional Independence Assumption: **features are independent of each other given the class:**

$$P(X_1,\ldots,X_5 \mid C) = P(X_1 \mid C) \bullet P(X_2 \mid C) \bullet \cdots \bullet P(X_5 \mid C)$$

- **This model is appropriate for binary variables**