Week 10 - Probability, Probabilistic Models (PGM), Beyesian Models and Probabilistic Classification

Section 15~16 of Lecture Notes

Statistical machine learning

Probability

- Random Variables
  - A random variable is some aspect of the world about which we (may) have uncertainty
  - We denote random variables with capital letters
  - Like variables in a **Constraint Satisfaction Problems (CSP)**, random variables have domains
  - Example:
    - R = Is it raining? R in {true, false} (often write as {+r, -r})
- Probability Distributions
  - **A distribution is a TABLE of probabilities of values**
  - Features

$$\forall x \; P(X=x) \geq 0 \;\; and \;\; \sum_{x} P(X=x) = 1$$

- Joint Distributions
  - Example: P(T, W) / Distribution over T, W

T	W	P
hot	sun	0.4
hot	rain	0,1
cold	sun	0.2
cold	rain	0.3

- Probabilistic Models
  - A probabilistic model is **a joint distribution over a set of random variables**
  - Constraint Satisfaction Problems (CSP)
    - Constraints over T, W

T	W	P
hot	sun	T
hot	rain	F
cold	sun	F
cold	rain	T

- Events
  - An event is a set E of outcomes

$$P(E) = \sum_{(x_1 \cdots x_n) \in E} P(x_1 \cdots x_n)$$

- Conditional Probabilities
  - **the probability of A given B**, which is the probability of event A happen after event B happened.
    In this case, if the temperature is cold, then the probability of the weather is sunny is 0.2/0.5= 0.4

T	W	P
hot	sun	0.4
hot	rain	0,1
cold	sun	0.2
cold	rain	0.3

$$P(a|b) = \frac{P(a,b)}{P(b)}$$

  - Conditional distributions are probability distributions over some variables **given fixed values of others**
- Normalization Trick

$$P(T,W)$$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$$\begin{aligned}P(W=s|T=c) &= \frac{P(W=s,T=c)}{P(T=c)}\\ &= \frac{P(W=s,T=c)}{P(W=s,T=c)+P(W=r,T=c)}\\ &= \frac{0.2}{0.2+0.3} = 0.4\end{aligned}$$

$$\begin{aligned}P(W=r|T=c) &= \frac{P(W=r,T=c)}{P(T=c)}\\ &= \frac{P(W=r,T=c)}{P(W=s,T=c)+P(W=r,T=c)}\\ &= \frac{0.3}{0.2+0.3} = 0.6\end{aligned}$$

$$P(W|T=c)$$

W	P
sun	0.4
rain	0.6

- The Product Rule

$$P(y)P(x|y) = P(x,y)$$

$$P(y)P(x|y) = P(x,y)$$

P(y)*P(x,y)\P(y)=P(x,y)

- Example:

$$P(W)$$

R	P
sun	0.8
rain	0.2

$$P(D|W)$$

D	W	P	
wet	sun	0.1	1
dry	sun	0.9	
wet	rain	0.7	1
dry	rain	0.3	

⇔

$$P(D,W)$$

D	W	P		
wet	sun	0.1 * 0.8	0.08	0.8
dry	sun	0.9 * 0.8	0.72	
wet	rain	0.7 * 0.2	0.14	0.2
dry	rain	0.3 * 0.2	0.06	

- The Chain Rule
  - More generally, can always write any joint distribution as an **incremental product of conditional distributions**

$$P(x_1,x_2,x_3) = P(x_1)P(x_2|x_1)P(x_3|x_2,x_1)$$

$$P(x_1,x_2,\cdots,x_i) = \prod_i P(x_1,x_2,\cdots,x_{i-1})$$

- **Bayes' Rule**

$$P(x,y) = P(x|y)P(y) = P(y|x)P(x)$$

Divided, we get

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

- Why is this at all helpful?
  - Lets us build one conditional from its reverse
  - Often one conditional is tricky but the other one is simple
  - Foundation of many systems we'll see later (e.g. ASR, MT)
- In the running for most important AI equation!

```
# random variable, head, tail
O = {H, T}
```

- **Rules (axioms) of probability**
  - P(A) ≥ 0 Probability is always non-negative value

$$P(A \cup B) = P(A) + P(B)$$
$$even\ P(A \cap B) = \emptyset$$

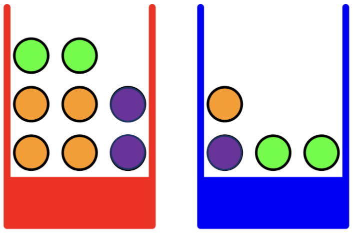
$$if\ \Omega = A, B, ..., N,\ then\ P(\Omega) = 1$$

- **Probability distribution functions** informs how are the "chances" of obtaining outcomes for a random variable (denoted in capital) can be **discrete** or **continuous**.

Summation for discrete, integration for continuous 离散型求和，连续型积分

- ***Discrete*** Probability mass function (PMF) gives the probability that the RV *X* takes on the value x (i.e., $P(X = x)$)
  - Each possibility lies in the range [0,1] and must have $\Sigma_x P(X = x) = 1$ (Normalized: the sum of all posibility is 1)
  - Categorical distribution
  - Uniform distribution: each probability is equal
- ***Continuous*** probability distribution functions (PDF) gives the amount of **probability per unit** (probability density)

- The probability of every event x is greater or equal to 0 $\forall x \in \Omega_x,\ P(x) \geq 0$
- Volume under this function gives the probability of the event represented by that volume, $\int_A p(x)dx = P(A)$ and it must be normalized,$\int_{\Omega x} p(x)dx = 1.$
  - In mathematics, an ***integral*** $\int$ is the **continuous analog of a sum**, which is used to calculate areas, volumes, and their generalizations.
- **Bernoulli Distribution**
  - Binary distribution, example, coin flip
  - **Sample space**: Ω = {r,b}
  - **Mass function (PMF)**: $p \in [0, 1]$



Identity of the box is a random variable: B

Two possible values (outcomes): red ($B = r$) or blue ($B = b$)

$$P(B = r) = 0.4$$

$$P(B = b) = 0.6$$

    - $P(B = red) = p$
    - $P(B = blue) = 1 - p$
  - **Normalization**

$$\sum_{x \in \Omega} P(X = x) = P(X = red) + P(X = blue)$$

$$= p + 1 - p = 1$$

- **Categorical Distribution (Like the color of a ball)**
  - Discrete distribution, example,
  - **Sample space**: Ω = {1, 2,..., N}
  - **Mass function (PMF)**: each outcome can have a different probability, so $P(X = x) = p_x$ (i.e., parameter vector with each $p_x \in [0, 1]$)
  - **Normalization**: the sum of the probability of each color category

$$\sum_{x \in \{1,2,\cdots,N\}} p_x = 1$$

- **Uniform Distribution (**a special case of the **categorical distribution)**
  - Special case of the categorical distribution where each outcome is equally likely, example, an obvious example is the fair dice with N = 6
  - **Sample space**: Ω = {1, 2,..., N}
  - **Mass function (PMF)**: each outcome can have a different probability, so

$$P(X = x) = p = \frac{1}{N} \quad p \in [0, 1]$$

  - **Normalization**

$$\sum_{x \in \{1,2,\cdots,N\}} \frac{1}{N} = N * \frac{1}{N} = 1$$

- **Binomial distribution** which can be considered a model of a sequence of independent, but biased
  - It can be shown that the probability of any sequence of flips with $x \leq N$ heads is $(1 - p)^{N-x} p^x$.
  - The number of ways of obtaining x heads in this sequence is given by the **binomial coefficient**

$$\binom{N}{x} = \frac{N!}{x!(N-x)!}$$

  - This is the number of combinations of k elements among N, from which we obtain the binomial mass function

$$P(X = x) = \binom{N}{x} (1-p)^{N-x} p^x = \frac{N!}{x!(N-x)!} (1-p)^{N-x} p^x$$

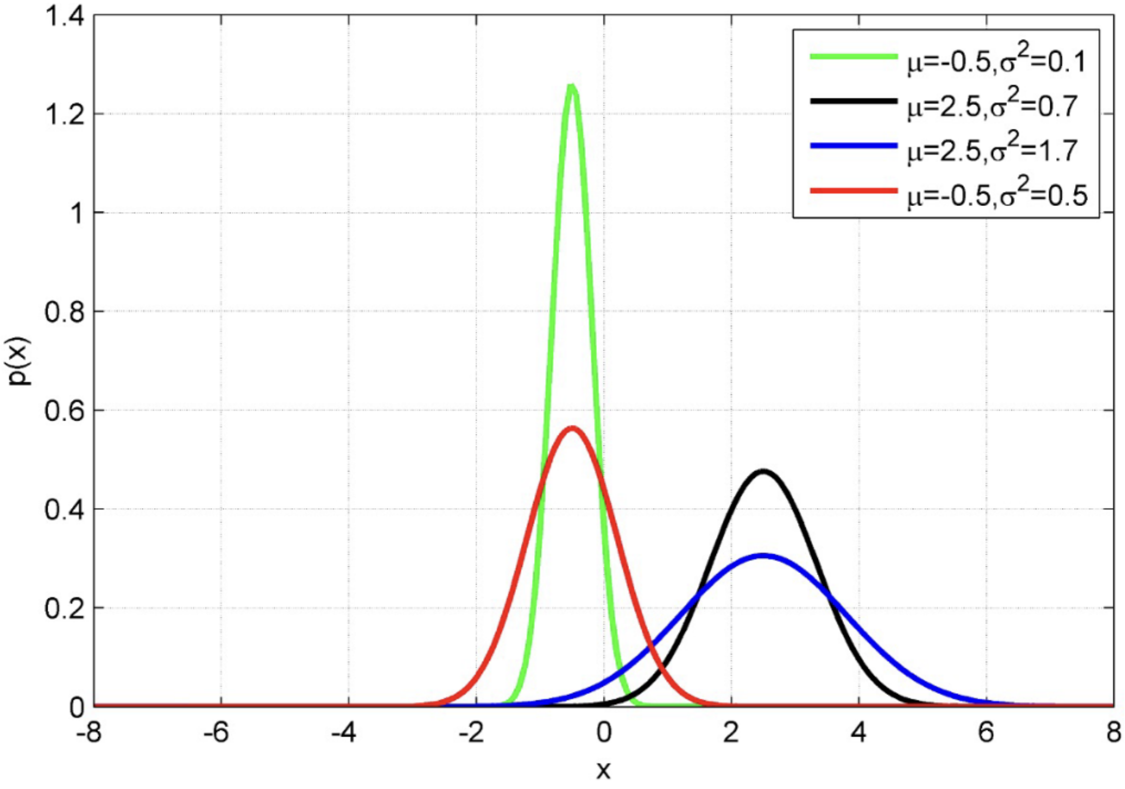
- Gaussian (Normal) distribution
  - Most ubiquitous continuous distribution
  - **Sample space**: Ω = ℝ
  - **Density function (PDF)**:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{(-\frac{(x-\mu)^2}{2\sigma^2})}$$

    with **mean** $\mu \in \mathbb{R}$ and **variance** $\sigma^2 > 0$.

  - **Notes:** μ is both **mean**, **median** and **mode** of the density

        - The Gaussian has many remarkable properties. Among these: the **mean**, **median** (value which has as much probability on one side as the other) and **mode** (the highest probability value) are all equal to μ, and the distribution of a sum of an infinite number of random variables (with finite mean and variance), is Gaussian (via the famous **central limit theorem**). For this reason (and many others), the Gaussian is widely used in probabilstic ML and AI.



- **Other probabilities intuition**
    - Joint Probability $Pr(X = x \ and \ Y = y) = P(X = x, Y = y)$
        - defined on the joint sample space $\Omega_X * \Omega_Y$. As an example, consider the case of a coin flip paired with a dice throw, the joint sample space has 2 × 6 = 12 possible outcomes, $\Omega_X * \Omega_Y = \{(H,1),(H,2),(H,3),(H,4),(H,5),(H,6),(T,1),(T,2),(T,3),(T,4),(T,5),(T,6)\}$
        - As a proper distribution function, the joint distribution must be normalized, so $\sum_{(x,y)\in\Omega_X\times\Omega_Y} P(X = x, Y = y) = 1.$
        - Joint density functions should be normalized

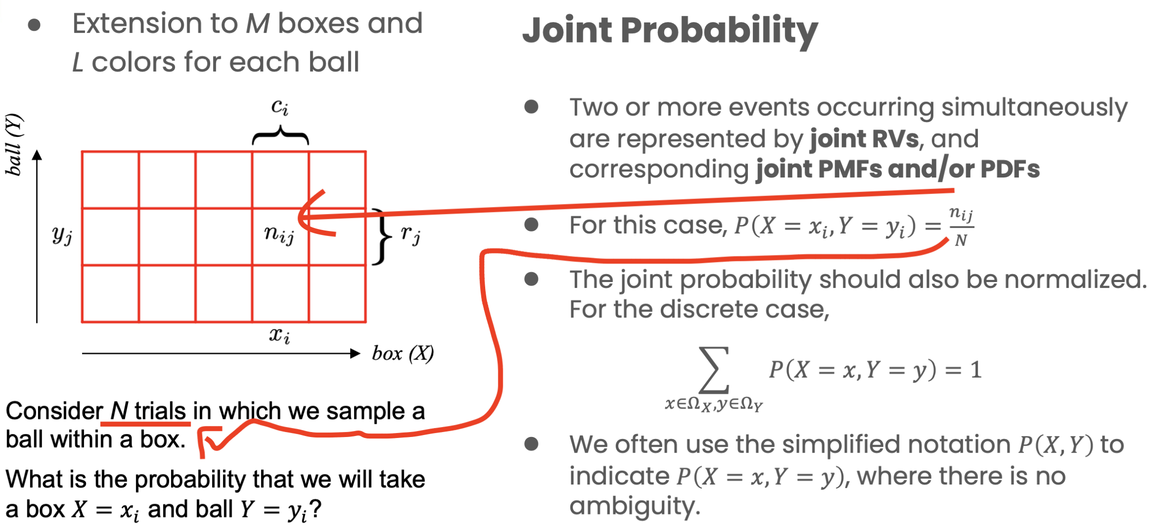
$$\int_{\Omega_X * \Omega_Y} p(x,y)dxdy = 1$$

1. **Integration over the Sample Space:** The integral $\int_{(}\Omega_x * \Omega_y)p(x,y)dxdy$ integrates the joint PDF $p(x,y)$ over the entire joint sample space (Ω_x × Ω_y) defined for the continuous variables.
2. **Why it equals 1:** Similar to the discrete case, this integration ensures that we account for the probability distribution across all possible combinations of values for X and Y within the continuous sample space. By integrating the PDF over the entire space, we capture the total probability under the curve, which must equal 1 to represent the certainty of encountering one outcome from the continuous distribution.

In essence, both the summation and integration ensure that the total probability assigned to all possible outcomes within a sample space adds up to 1. This reflects the fundamental principle of probability that exactly one outcome will occur from a given event.

> ## Other probabilities intuition

#Code



- Extension to M boxes and L colors for each ball

Consider N trials in which we sample a ball within a box.

What is the probability that we will take a box $X = x_i$ and ball $Y = y_i$?

Joint Probability

- Two or more events occurring simultaneously are represented by **joint RVs**, and corresponding **joint PMFs and/or PDFs**
- For this case, $P(X = x_i, Y = y_i) = \frac{n_{ij}}{N}$
- The joint probability should also be normalized. For the discrete case,

$$\sum_{x \in \Omega_X, y \in \Omega_Y} P(X = x, Y = y) = 1$$

- We often use the simplified notation $P(X, Y)$ to indicate $P(X = x, Y = y)$, where there is no ambiguity.

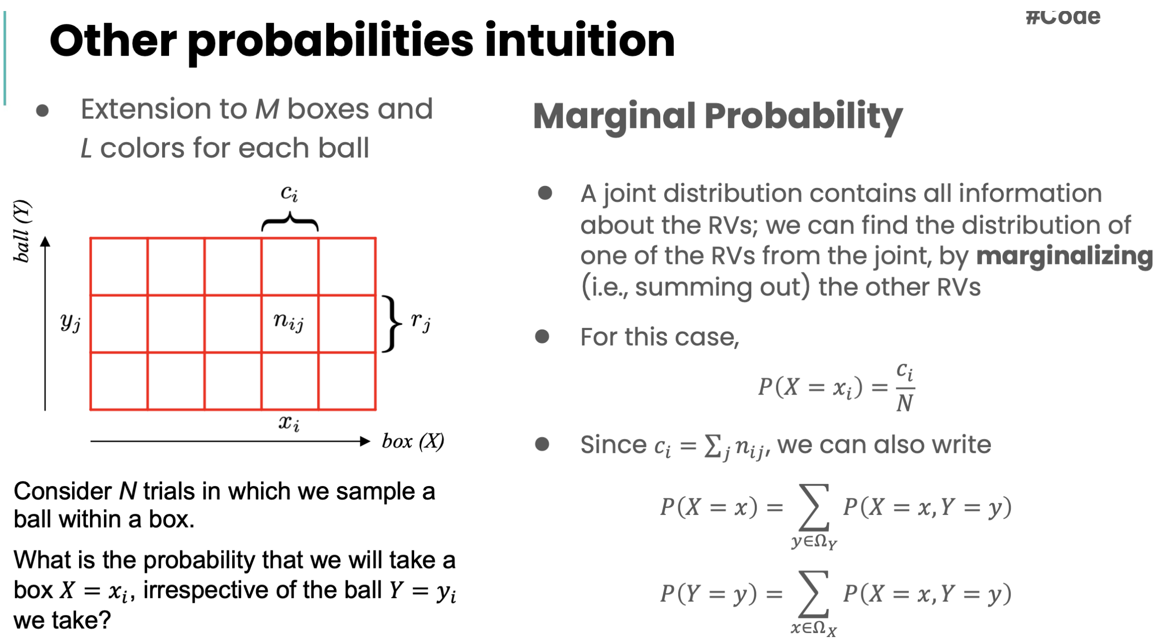
- Marginal Probability
  - A joint distribution contains all information about the joint RVs, so for instance, we can find the distribution function of one of the RVs from the joint, by **marginalization, and the same result applies to continuous distributions using integration (积分) rather than summation (求和)**

$$P(Y = y) = \sum_{x \in \Omega_X} P(X = x, Y = y)$$

$$P(X = x) = \sum_{y \in \Omega_Y} P(X = x, Y = y)$$

- If the probability of a conditional distribution $P(X|Y)$ is unaffected by the conditioning variable Y , we say that X is **independent** of Y .
- In other words, if $P(X|Y) = P(X)$, then we can rearrange this to write $P(X,Y) = P(X)P(Y)$

  The fact that $P(X,Y) \neq P(X)P(Y)$ means that X and Y are not independent.



#Code

Other probabilities intuition

- Extension to *M* boxes and *L* colors for each ball

Consider *N* trials in which we sample a ball within a box.

What is the probability that we will take a box $X = x_i$, irrespective of the ball $Y = y_i$ we take?

Marginal Probability

- A joint distribution contains all information about the RVs; we can find the distribution of one of the RVs from the joint, by **marginalizing** (i.e., summing out) the other RVs
- For this case,

$$P(X = x_i) = \frac{c_i}{N}$$

- Since $c_i = \sum_j n_{ij}$, we can also write

$$P(X = x) = \sum_{y \in \Omega_Y} P(X = x, Y = y)$$

$$P(Y = y) = \sum_{x \in \Omega_X} P(X = x, Y = y)$$

- Conditional Probability
  - Properly normalized, fixing one variable allows the joint to act as a new distribution called the **conditional**
    - Condition X=x, given Y=y, the shorthand P(X|Y)

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

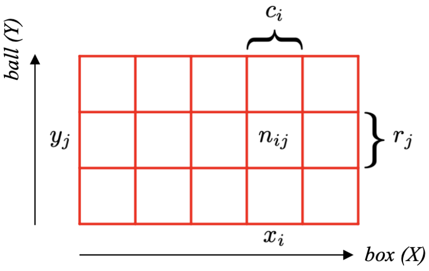
    - Condition Y=y, given X=x, the shorthand P(Y|X)

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

Other probabilities intuition

#Code

- Extension to *M* boxes and *L* colors for each ball



Consider *N* trials in which we sample a ball within a box.

What is the probability that we will take a ball $Y = y_i$, given we took box $X = x_i$ already?

Conditional Probability

- Properly normalized, fixing one variable allows the joint to behave as a new distribution called the **conditional**.
- For this case,

$$P(Y = y_i | X = x_i) = \frac{n_{ij}}{c_i}$$

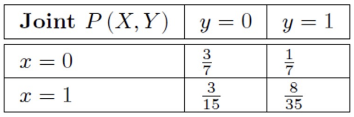
- Since $p(X = x, Y = y) = \frac{n_{ij}}{N}$,

$$P(X = x, Y = y) = p(Y = y|X = x)p(X = x)$$

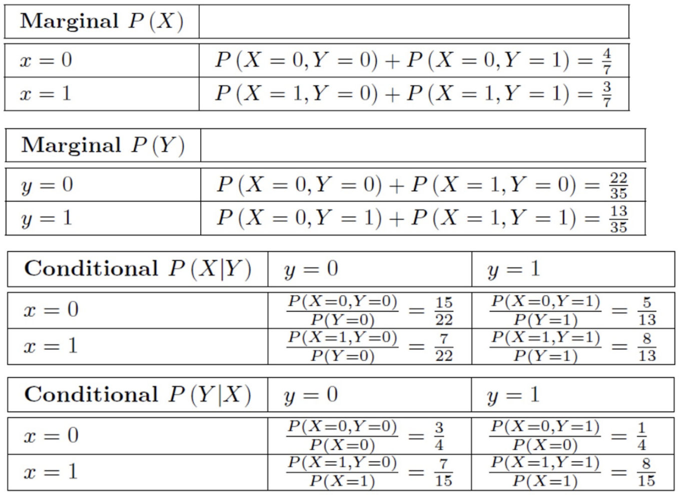
$P(X, Y) = p(Y|X)p(X)$ or $P(X, Y) = p(X|Y)p(Y)$

Marginals and conditionals: another example (Table 15.1)

#Code



Joint $P(X,Y)$	$y=0$	$y=1$
$x=0$	$\frac{3}{7}$	$\frac{1}{7}$
$x=1$	$\frac{3}{15}$	$\frac{8}{35}$



Marginal $P(X)$	
$x=0$	$P(X=0,Y=0)+P(X=0,Y=1)=\frac{4}{7}$
$x=1$	$P(X=1,Y=0)+P(X=1,Y=1)=\frac{3}{7}$

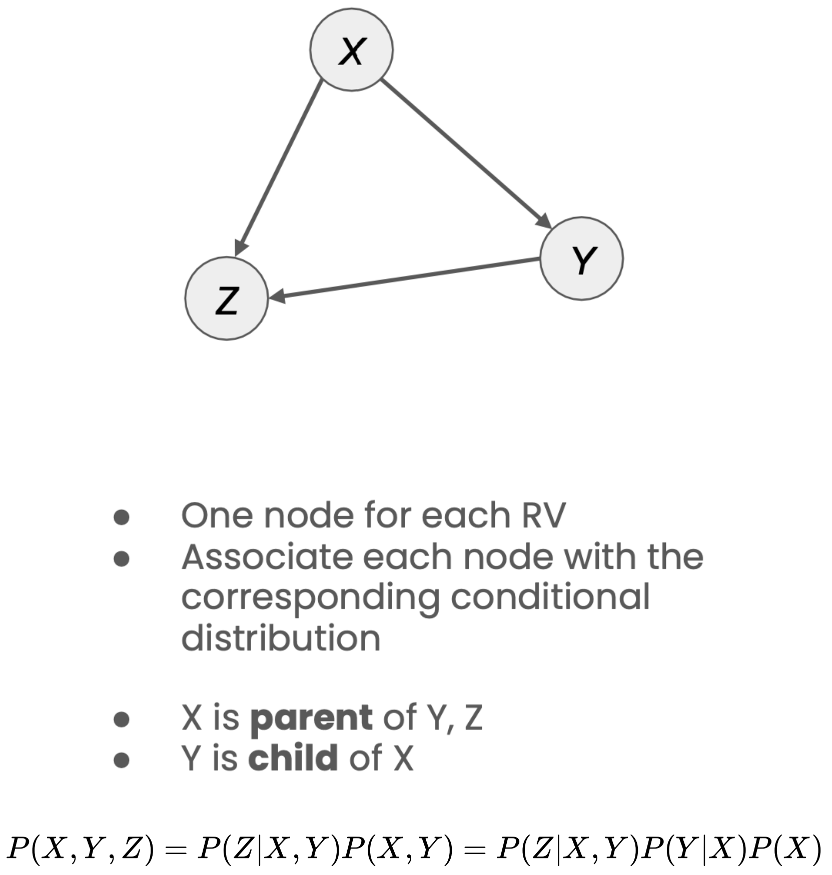
Marginal $P(Y)$	
$y=0$	$P(X=0,Y=0)+P(X=1,Y=0)=\frac{22}{35}$
$y=1$	$P(X=0,Y=1)+P(X=1,Y=1)=\frac{13}{35}$

Conditional $P(X\|Y)$	$y=0$	$y=1$
$x=0$	$\frac{P(X=0,Y=0)}{P(Y=0)}=\frac{15}{22}$	$\frac{P(X=0,Y=1)}{P(Y=1)}=\frac{5}{13}$
$x=1$	$\frac{P(X=1,Y=0)}{P(Y=0)}=\frac{7}{22}$	$\frac{P(X=1,Y=1)}{P(Y=1)}=\frac{8}{13}$

Conditional $P(Y\|X)$	$y=0$	$y=1$
$x=0$	$\frac{P(X=0,Y=0)}{P(X=0)}=\frac{3}{4}$	$\frac{P(X=0,Y=1)}{P(X=0)}=\frac{1}{4}$
$x=1$	$\frac{P(X=1,Y=0)}{P(X=1)}=\frac{7}{15}$	$\frac{P(X=1,Y=1)}{P(X=1)}=\frac{8}{15}$

  - Independence/conditional independence such as this, is of critical important in probabilistic AI and ML, since it allows joint RVs to be efficiently modelled by subsets of the data. This is the basis of **probabilistic graphical models (PGMs)**

- Probabilistic graphical models: must **be directed acyclic graph (DAG)**
  - Given a probability distribution with multiple RVs, we can represent their mutual conditional dependence graphically using a **probabilistic graphical model** (PGM)

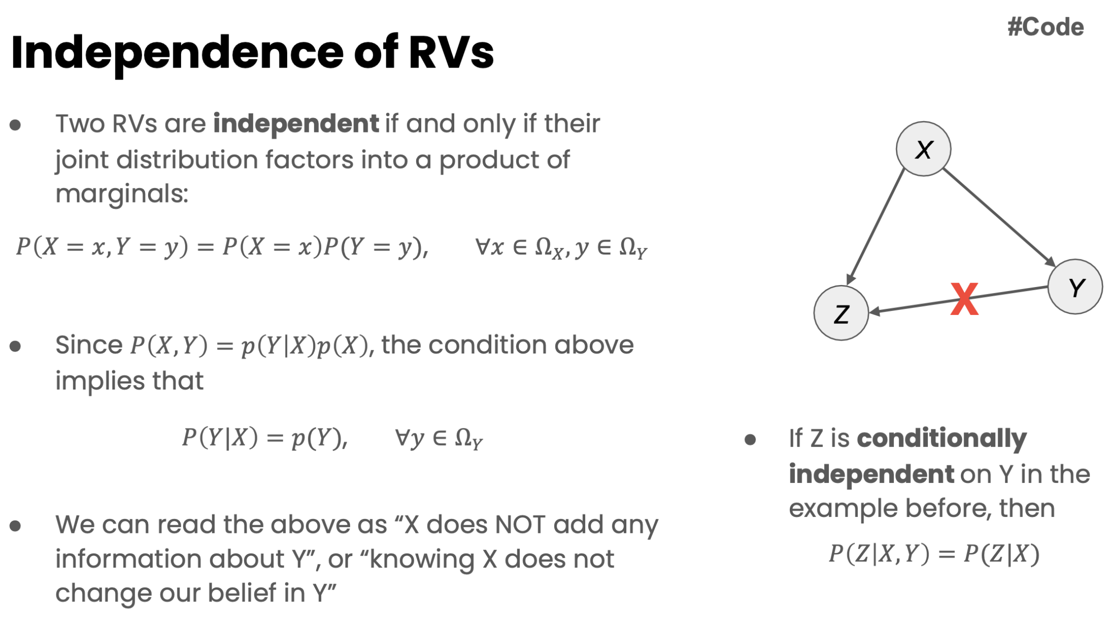


- One node for each RV
- Associate each node with the corresponding conditional distribution

- X is **parent** of Y, Z
- Y is **child** of X

$$P(X, Y, Z) = P(Z|X, Y)P(X, Y) = P(Z|X, Y)P(Y|X)P(X)$$

- Independence of RVs



Independence of RVs

#Code

- Two RVs are **independent** if and only if their joint distribution factors into a product of marginals:

$$P(X = x, Y = y) = P(X = x)P(Y = y), \qquad \forall x \in \Omega_X, y \in \Omega_Y$$

- Since $P(X,Y) = p(Y|X)p(X)$, the condition above implies that

$$P(Y|X) = p(Y), \qquad \forall y \in \Omega_Y$$

- We can read the above as “X does NOT add any information about Y”, or “knowing X does not change our belief in Y”

- If Z is **conditionally independent** on Y in the example before, then

$$P(Z|X, Y) = P(Z|X)$$

- Independence/conditional independence such as this, is critical important in probabilistic AI and ML, since it allows joint RVs to be modelled by subsets of the data; basis of **probabilistic**

graphical models

  - Special graph structures for applications such as time or spatially ordered data
- The graph must have **no cycles** (loops), a **directed acyclic graph** (DAG)
- Every PGM has a corresponding **Markov factorization**, which is a form of the chain rule compatible with a **topological ordering (拓扑序)** of the DAG, taking into account the conditional independencies due to the absence of edges
  - In a directed acyclic graph (DAG), a topological ordering is a sequence that visits each node exactly once, such that for any directed edge from node A to node B, node A appears before node B in the sequence.
  - **In essence, topological ordering provides a structured approach to navigating the conditional dependencies within probabilistic graphical models, enabling efficient and accurate probabilistic inference in AI and ML tasks.**

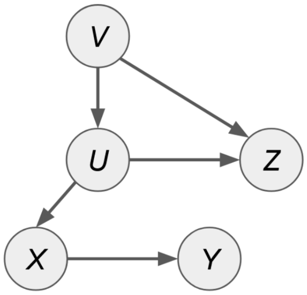
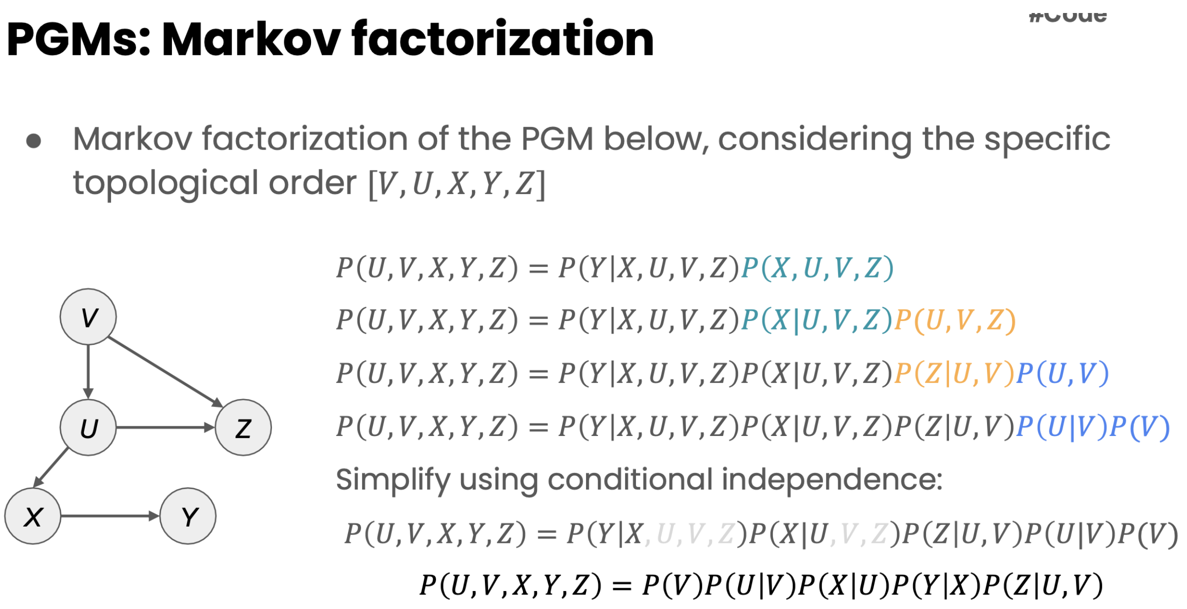


Figure 15.2.: Simple example probabilistic graphical model over the five RVs V, U, X, Y and Z. We can immediately see the child/parent relationships between variables; some of them: V is a parent of both U and Z, Y is a child of X and X is a child of U in turn. We can therefore read off the conditional independence relationships, for example, $P(Z|V,U) = P(Z|V,U,X,Y)$ (in words, Z is conditionally independent of X and Y given U and V), $P(Y|X) = P(Y|X,Z)$ (Y is conditionally independent of Z given X).

- **Topological ordering**
  - We can find a sequence of variables e.g. [X, U, V] or [U, V, X], such that, every variable in the sequence occurs **after its parents** (or before its children), in the DAG. Such a sequence (there can be more than one of them) is called a **topological ordering** of the DAG.
- Markov factorization
  - The simplest conditionals of factorization (which is unique), is known as the **Markov factorization** of the PGM. Given the PGM, we can simply 'read off' this factorization directly.
  - Simplify: only write down its parent node



#Code

PGMs: Markov factorization

- Markov factorization of the PGM below, considering the specific topological order $[V,U,X,Y,Z]$

$$P(U,V,X,Y,Z) = P(Y|X,U,V,Z)P(X,U,V,Z)$$
$$P(U,V,X,Y,Z) = P(Y|X,U,V,Z)P(X|U,V,Z)P(U,V,Z)$$
$$P(U,V,X,Y,Z) = P(Y|X,U,V,Z)P(X|U,V,Z)P(Z|U,V)P(U,V)$$
$$P(U,V,X,Y,Z) = P(Y|X,U,V,Z)P(X|U,V,Z)P(Z|U,V)P(U|V)P(V)$$

Simplify using conditional independence:

$$P(U,V,X,Y,Z) = P(Y|X,U,V,Z)P(X|U,V,Z)P(Z|U,V)P(U|V)P(V)$$
$$P(U,V,X,Y,Z) = P(V)P(U|V)P(X|U)P(Y|X)P(Z|U,V)$$

Beyesian Models, and Probabilistic Classification

How to use probability in classification

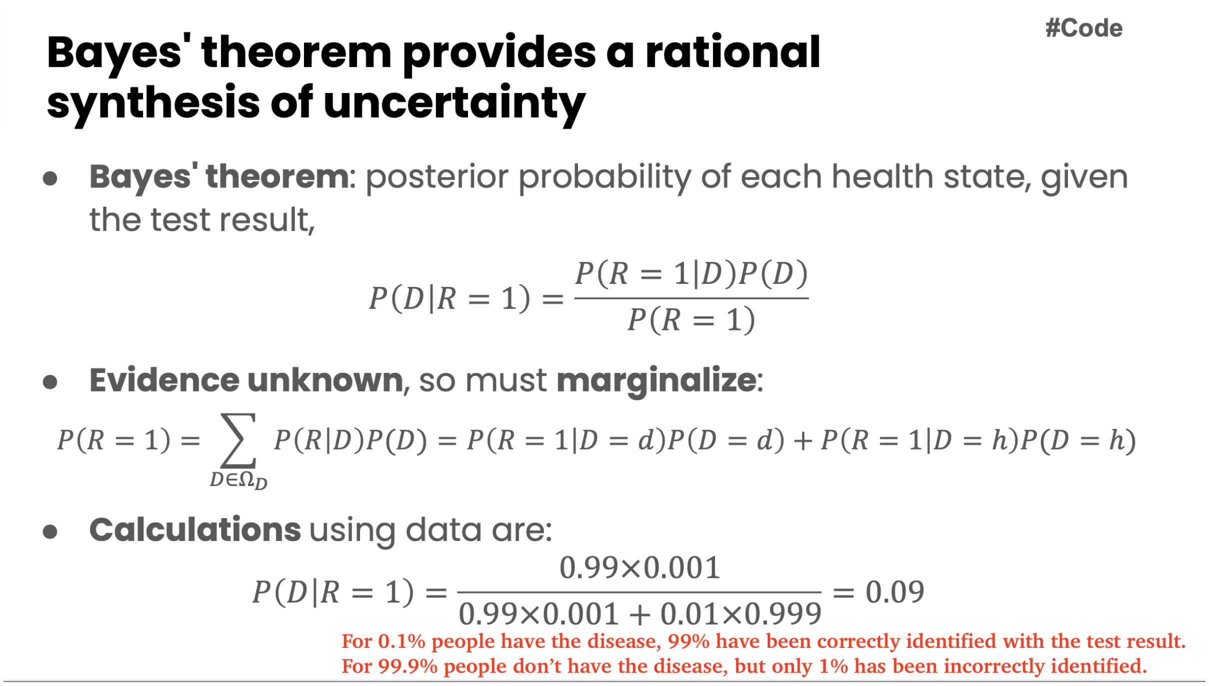
- Bayes’ theorem: **precise synthesis** of disparate sources of **uncertain information**
  - Bayes' theorem helps us determine the **positive predictive value (PPV)**, which is the probability of actually having the disease given a positive test result.

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

  - If P(Y), the evidence is unknown, it equals to $\sum_{x\in\Omega_X} P(Y|X=x)P(X=x)$
  - When multiplying, it is always the later $P(former|later)P(later)$
  - P(X) Prior
  - P(X|Y) Posterior
  - P(Y) Evidence
    - If evidence P(Y) is unknown, can use $P(Y) = \sum_{x\in\Omega X} P(Y|X=x)P(X=x)$
  - P(Y|X) Likelihood distributions
  - ***Naive Bayes classifier***

$$P(Y|X) = P(X^1|Y)P(X^2|Y)\cdots P(X^D|Y)$$

$$y = \underset{y\in\Omega Y}{\arg\max}\, P(X^1|Y)P(X^2|Y)\cdots P(X^D|Y)P(Y=y)$$



#Code

Bayes' theorem provides a rational synthesis of uncertainty

- **Bayes' theorem**: posterior probability of each health state, given the test result,

$$P(D|R=1) = \frac{P(R=1|D)P(D)}{P(R=1)}$$

- **Evidence unknown**, so must **marginalize**:

$$P(R=1) = \sum_{D\in\Omega_D} P(R|D)P(D) = P(R=1|D=d)P(D=d) + P(R=1|D=h)P(D=h)$$

- **Calculations** using data are:

$$P(D|R=1) = \frac{0.99\times 0.001}{0.99\times 0.001 + 0.01\times 0.999} = 0.09$$

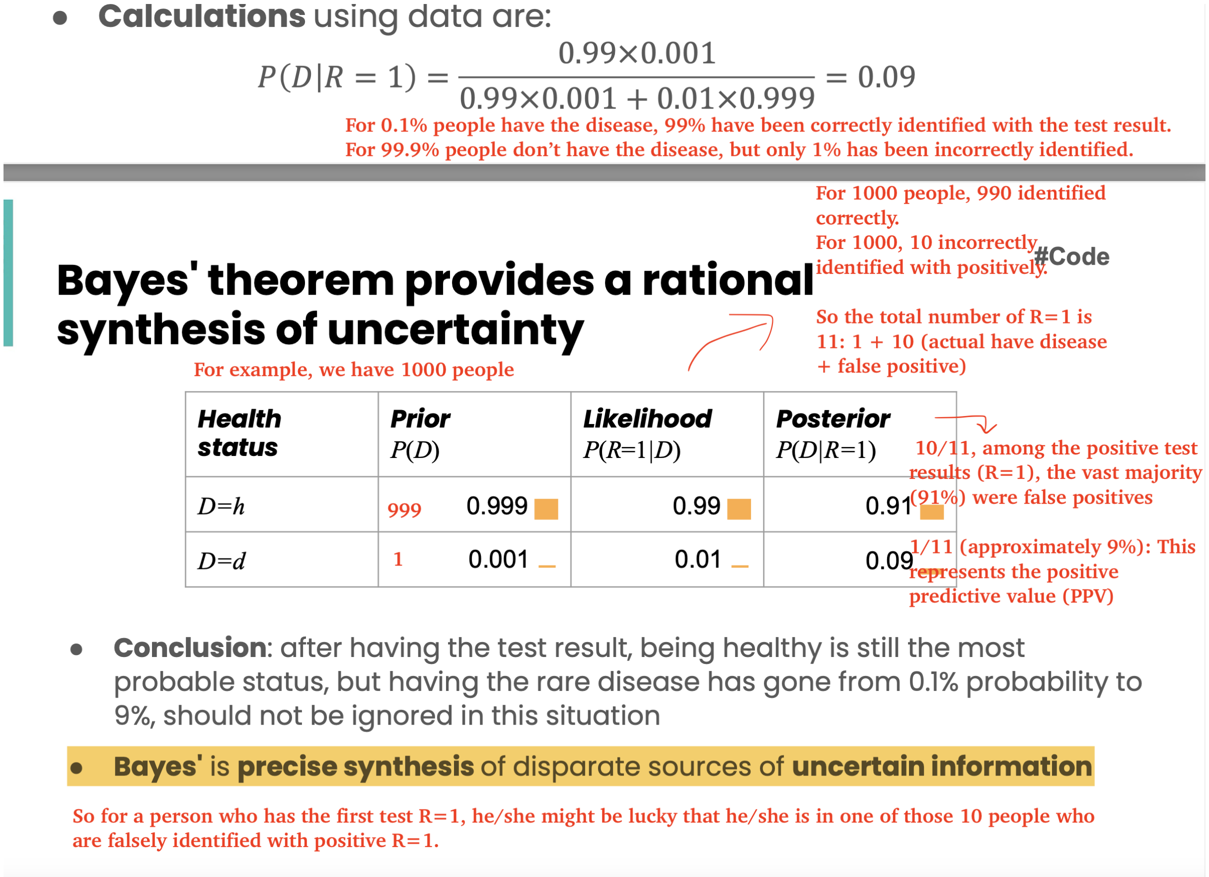
For 0.1% people have the disease, 99% have been correctly identified with the test result.
For 99.9% people don't have the disease, but only 1% has been incorrectly identified.

- **Intuitive Explanation:**

  Imagine 1000 people:

  - 1 person (0.1%) has the disease.
  - 999 people (99.9%) don't have the disease.
  - **Correct Test Results:**
    - Ideally, the test correctly identifies the 1 person with the disease (true positive).
    - Ideally, it also correctly identifies all 999 healthy people (true negative).
  - **Reality with False Positives:**
    - In reality, the test might miss identifying a small fraction (1%) of the person with the disease (false negative). Let's say it misses 0 people here for simplicity.
    - More importantly, the test might incorrectly identify 1% (around 10 people) of the healthy individuals as having the disease (**false positive**).
- **Impact on Bayes' Theorem:**
  - When you calculate P(D|R=1) using Bayes' theorem, you consider both true and false positives in the equation. The **large** number of false positives (10 people from the healthy group) compared to the true positives (1 person with the disease) significantly reduces the final probability of actually having the disease given a positive test.
- **Summary:**
  - While it's important to consider the accuracy for people with the disease (0.1% with the disease, 99% correctly identified), the much larger population of healthy people (99.9%) and the possibility of false positives (1% incorrectly identified) play a crucial role in Bayes'

theorem. This is because it significantly impacts the final probability of having the disease given a positive test result, highlighting the importance of considering both factors.



- **Calculations** using data are:

$$P(D|R=1)=\frac{0.99\times 0.001}{0.99\times 0.001+0.01\times 0.999}=0.09$$

For 0.1% people have the disease, 99% have been correctly identified with the test result.
For 99.9% people don't have the disease, but only 1% has been incorrectly identified.

Bayes' theorem provides a rational synthesis of uncertainty

For 1000 people, 990 identified correctly.
For 1000, 10 incorrectly identified with positively.

#Code

So the total number of R=1 is 11: 1 + 10 (actual have disease + false positive)

For example, we have 1000 people

Health status	**Prior** $P(D)$	**Likelihood** $P(R=1\|D)$	**Posterior** $P(D\|R=1)$
$D=h$	999 0.999 🟧	0.99 🟧	0.91 🟧
$D=d$	1 0.001 ▁	0.01 ▁	0.09 ▁

10/11, among the positive test results (R=1), the vast majority (91%) were false positives

1/11 (approximately 9%): This represents the positive predictive value (PPV)

- **Conclusion**: after having the test result, being healthy is still the most probable status, but having the rare disease has gone from 0.1% probability to 9%, should not be ignored in this situation
- **Bayes'** is **precise synthesis** of disparate sources of **uncertain information**

So for a person who has the first test R=1, he/she might be lucky that he/she is in one of those 10 people who are falsely identified with positive R=1.

- Probabilistic classification using Bayes' theorem
  - A rational decision is to select the value of Y which maximizes the posterior $P(Y|X)$, given condition X the probability of Y happens, which is called the **maximum a-posteriori (MAP)** decision rule

Probabilistic classification using Bayes' theorem

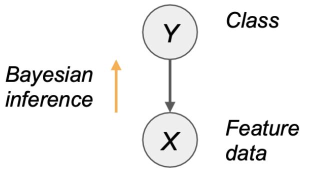
#Code

- **Probabilistic classification** can be expressed as an application of Bayes' rule:
  - given some **input** (feature) data X, determine the **probability** $P(Y|X)$ of the **class Y to which X belongs** (posterior), taking into account $P(Y)$ (prior) and how probable that class is before having seen the data, $P(X|Y)$
- A good decision is to select the value of Y which maximizes $P(Y|X)$, called the **maximum a-posteriori** (MAP) decision:

$$y^\star = \arg\max_{y \in \Omega_Y} P(Y = y|X = x)$$

and can avoid the need to have the evidence $P(X = x)$, since it does not depend upon Y:

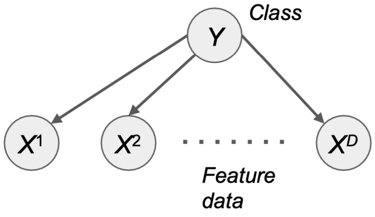
$$y^\star = \arg\max_{y \in \Omega_Y} P(X = x|Y = y)P(Y = y)$$



Naive Bayes classifier: MAP solution

#Code

- In general, the input features X will be **multidimensional** (a vector of values) and will not be independent of each other making it difficult to estimate the likelihood $P(X|Y)$ from the data
- The so-called **naive Bayes' classifier** simplifies the classification model by assuming that each feature is conditionally independent of the others, given the class.



Markov Factorization:

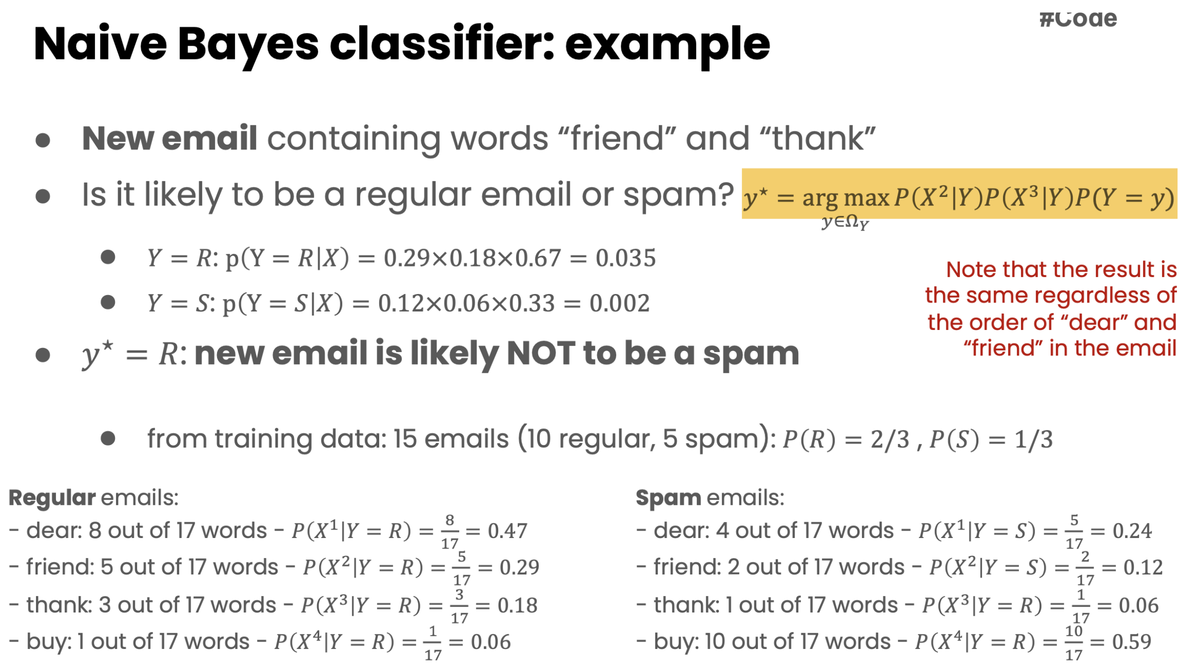
$$P(X|Y) = P(X^1|Y)P(X^2|Y) \cdots P(X^D|Y)$$

Using Bayes' theorem:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} = \frac{P(X^1|Y)P(X^2|Y) \cdots P(X^D|Y)P(Y)}{P(X)}$$

Since $P(X)$ is independent of Y,

$$y^\star = \arg\max_{y \in \Omega_Y} P(X^1|Y)P(X^2|Y) \cdots P(X^D|Y)P(Y = y)$$



#Code

Naive Bayes classifier: example

- **New email** containing words “friend” and “thank”
- Is it likely to be a regular email or spam? $y^\star = \arg\max_{y \in \Omega_Y} P(X^2|Y)P(X^3|Y)P(Y=y)$
  - $Y = R$: $\mathrm{p}(\mathrm{Y} = R|X) = 0.29 \times 0.18 \times 0.67 = 0.035$
  - $Y = S$: $\mathrm{p}(\mathrm{Y} = S|X) = 0.12 \times 0.06 \times 0.33 = 0.002$
- $y^\star = R$: **new email is likely NOT to be a spam**

Note that the result is the same regardless of the order of “dear” and “friend” in the email

  - from training data: 15 emails (10 regular, 5 spam): $P(R) = 2/3$, $P(S) = 1/3$

Regular emails:
- dear: 8 out of 17 words - $P(X^1|Y = R) = \frac{8}{17} = 0.47$
- friend: 5 out of 17 words - $P(X^2|Y = R) = \frac{5}{17} = 0.29$
- thank: 3 out of 17 words - $P(X^3|Y = R) = \frac{3}{17} = 0.18$
- buy: 1 out of 17 words - $P(X^4|Y = R) = \frac{1}{17} = 0.06$

Spam emails:
- dear: 4 out of 17 words - $P(X^1|Y = S) = \frac{5}{17} = 0.24$
- friend: 2 out of 17 words - $P(X^2|Y = S) = \frac{2}{17} = 0.12$
- thank: 1 out of 17 words - $P(X^3|Y = R) = \frac{1}{17} = 0.06$
- buy: 10 out of 17 words - $P(X^4|Y = R) = \frac{10}{17} = 0.59$

- **Naive Bayes classifier: analysis**
  - Naive Bayes surprisingly good for high-dimensional problems (*D* large), since **does not require a large amount of training data**
  - **Estimating feature distribution** parameters is **very quick**: linear in *D*, the number of features
  - **Making a prediction** requires **evaluating *D* times** $|\Omega_Y|$(the number of classes), which is usually easy to carry out in practice
  - Nonetheless, assumption of **feature independence is unrealistic** for many practical ML problems
- Explain the rational behind, probabilistic structure of, and method of, naive Bayes classification: apply naive Bayes’ to small classification problems.

Naive Bayes Classification Explained

Naive Bayes is a family of supervised learning algorithms for classification tasks. It works based on Bayes' theorem and a key assumption of **feature independence**. Let's break down the core aspects:

Rationale:

- **Probabilistic approach:** Classifies data by calculating the probability of an instance belonging to a particular class.
- **Simplicity and efficiency:** Relatively easy to implement and computationally inexpensive, making it suitable for large datasets.

Probabilistic Structure:

- **Bayes' Theorem:** Calculates the posterior probability (probability of a class given an instance) using prior probabilities, class likelihoods, and feature probabilities.
- **Naive Assumption:** The key assumption is that features (attributes used for classification) are conditionally independent given the class label. In simpler terms, the presence of one feature doesn't influence the presence of another feature, considering the class is already known. While this assumption might not always hold true in reality, it often works well in practice.

Method:

1. **Training:** The algorithm learns from a labeled dataset where each instance has features and a corresponding class label. It calculates the probabilities of each feature value occurring for each class and the prior probability of each class existing in the data.
2. **Classification:** For a new, unlabeled instance, the algorithm calculates the posterior probability of each class using Bayes' theorem and the learned probabilities. The class with the highest posterior probability is assigned to the new instance.

Applying Naive Bayes to Small Classification Problems

Naive Bayes can be a good choice for small classification problems due to its simplicity and efficiency. Here's an example:

Imagine classifying emails as spam or not spam based on features like наличие слова "бесплатно" (presence of the word "free" in Russian) and наличие знаков препинания в начале письма (presence of punctuation marks at the beginning of the email).

1. **Training:** Train the model on a dataset of labeled emails (spam/not spam) with these features. Calculate probabilities:
   - P(наличие слова "бесплатно" | spam) - Probability of "free" appearing in spam emails.
   - P(наличие знаков препинания в начале письма | not spam) - Probability of punctuation at the beginning in non-spam emails.
   - Prior probabilities of spam and not spam emails (P(spam) and P(not spam)).
2. **Classification:** For a new email, calculate the posterior probability of being spam and not spam using Bayes' theorem and the learned probabilities. The class with the higher probability is the predicted class (spam or not spam).

Note: Naive Bayes might not perform as well for small datasets with very few examples of each class compared to the number of features. This is because the estimation of probabilities might be less accurate with limited data.