4. Probability and Bayes Network

Comment about the dataset and its notation.
In classification problem we have to learn a function $f: X \to Y$ and Y is FINITE (set of possible classes).

- In supervised learning the dataset is a SUBSET of the cartesian product between X and Y:

$$D \subset \{X \times Y\} \text{ in particular:}$$

$$D = \{(x_i, y_i) \mid x_i \in X, y_i \in Y\}$$

D is smaller than X, to denote the subset of X that is in D we usually use X_D:

$$X_D = \{x_i \in X \mid (x_i, y_i) \in D\}$$

Sometimes we make a simplification of notation by using:

short not. $x \in X_D \to$ belong to the projection of X on D

$\hookrightarrow x \in D \to$ not very exactly precise, but it is useful in order to make everything more compact.

HOW WE SPLIT D?

D: | S | T |

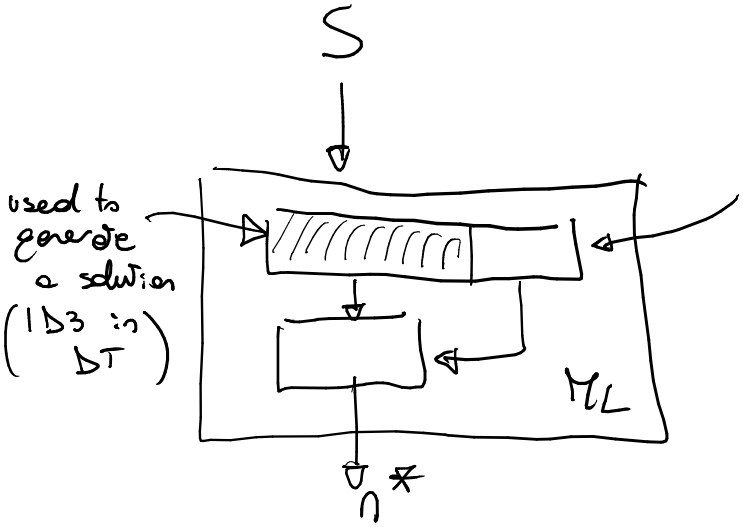
(*) ML → n^* → Evaluation → Accuracy

↳ In this part we can make hyperparameter TUNING

Very often we have to split D in two parts:
- S
- T

Techniques based on partitioning the dataset

Sometimes the machine learning algorithm inside can make some operations; let's expande the ML-box:



used in a subsequent phase, to find tuning of parameters (pruning in DT)

Therefore, if you two have two random seed sources you can change them to modify:

1. The partition between S and T
2. partition randomly the data into the ML-algorithm

SMALL DATASET?

All techniques used to improve performances may be not effective

- **Uncertainty**

A_t = leave for the airport t minutes before flights

New machine learning approach is based on probab. estimation and no explict representation of H (hypothesis space)

Will A_t get me there on time?

I don't know the traffic conditions, there are uncertainties, IN GENERAL IS NOT POSSIBLE TO GIVE AN ANSWER TO THIS QUESTIONS.

The purely logical approach allow me to say T or F given the predicate A_t, but I have some problems.

In a purely logical approach is difficult to represent uncertainty, we don't have a way to express the degree of uncertainty.

With probability I can represent information about uncertainty by associating a number, i.e. i.e. A_{25} will take me ... with probability 0.04.

PROBABILITY

- Ω SAMPLE SPACE (set of possibilities)
- Any point is denoted with ω that is a particular ATOMIC EVENT (a particular outcome of a random process).

PROBABILITY SPACE: $P: \Omega \to \mathbb{R}$ s.t. $0 \leq P(\omega) \leq 1$ and $\sum_{\omega \in \Omega} P(\omega) = 1$

An event any subset of the sample space Ω.
In particular an EVENT is defined as:

We cannot use these definitions to compute probab.; in some cases we cannot even represent Ω

Set of situations referred to this event and its probability is:

$$P(A) = \sum_{\omega \in A} P(\omega)$$

The sum of the prop. of the situations that make a true

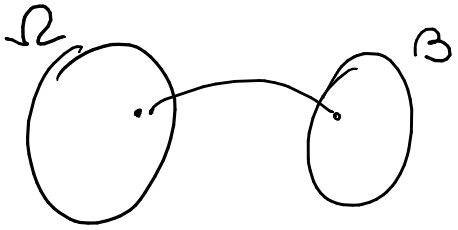
Example: "dice roll < 4", $A_1 = \{1, 2, 3\} \subset \Omega$

Δ Random variables ...

Random variables

A random variable (outcome of a random phenomena) is a function from the sample Ω to some range B:

$X : \Omega \to B$ is a function to any other set



A random variable perform this mapping. If I cannot represent Ω, the mapping happens but I cannot represent it!

Sometimes I cannot model Ω, so if I forget about the sample space and I look only at what happens (at the output), sometimes the random variable takes a value, sometimes another value, i.e. some takes T or sometimes F.

$$P(X = x_i) = \sum_{\{\omega \in \Omega \mid X(\omega) = x_i\}} P(\omega)$$ ← not possible to compute sometimes.

Propositions

A proposition is the event (subset of Ω) where the proposition is true. We can combine propositions

Prior probability

Prior or UNCONDITIONAL PROBABILITY of propositions corresponds to belief prior to arrival of any (new) evidence.

Is the probability of an event that we have without any knowledge.

The probability of rolling a dice without any experiment

Probability distribution

Is the set of all possible prob. values for all the possible values that the random variable can take.

A probability distribution is the set of prob. values for all possible assignments of a random variable.

The sum must be 1!

Joint probability distribution

Is the probability associated to a set of random variables. To consider all possible combinations we can consider a matrix that is N-dimensional in general and each dimension has the size of all the possible values which the the random variable can take.

This is still true for continuous random variables.

PROBLEM: Exponential number of parameters depending on the number of random variables.

Conditional / Posterior Probability

Belief after the arrival of some evidence.

I know the outcome of a random variable, how this affect the probability of other random variables?

We have information I know the value of some other random variable:

$$P(\text{Cavity} = T \mid \text{weather} = \text{Sunny})$$

In general:

$P(Cavity = T \mid weather = Sunny) \neq$
$P(cavity = true, weather = sunny) \neq$
$P(Cavity = true)$

Definition of conditional probability:

$$P(a|b) = \frac{P(a \wedge b)}{P(b)} \quad \text{if} \quad P(b) \neq 0$$

$$P(a \wedge b) = P(a|b)\,P(b) = P(b|a)\,P(a)$$

For a boolean random variable B:

$P(a) = P(a|b)P(b) + P(a|\neg b)P(\neg b)$, in general

for a random variable Y accepting mutually exclusive values y_i:

$$P(x) = \sum_{y_i \in D(y)} P(x|Y=y_i)\,P(Y=y_i)$$

$D(Y)$ is the set of values for Y

The CHAIN RULE is derived by successive application of the product rule:

$P(X_1, X_2) = P(X_1)P(X_2|X_1)$ (product rule)

CHAIN RULE:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid X_1, \ldots, X_{i-1})$$

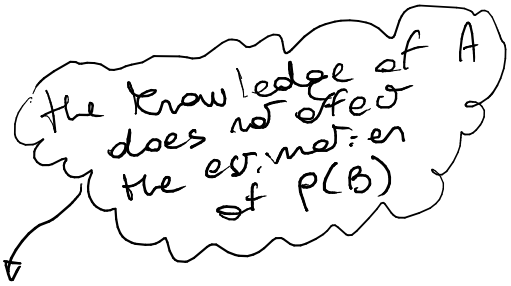
In the conditional probability the denominator can be seen as a NORMALIZATION CONSTANT α, it can be computed at the end.
In some cases we are not interested in the exact number, sometimes we want just the ARGMAX, not affected by α.

Note: when we have a complete joint probability is easy to compute the other probabilities (a priori, conditioned, ...)

Example of toothache and cavity table

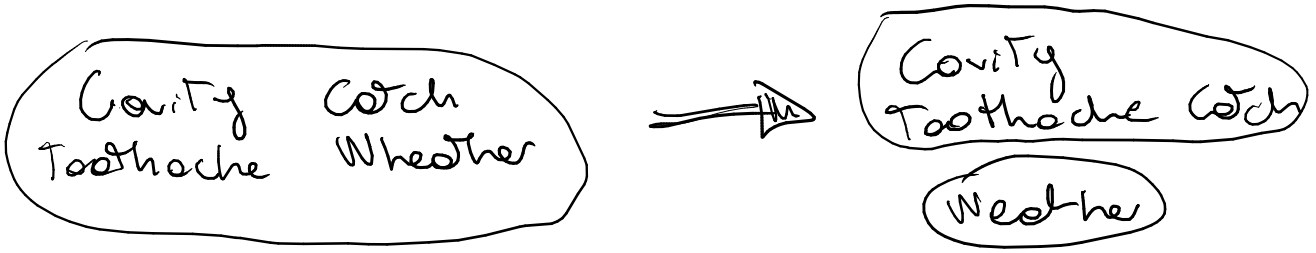
Independence

A and B are independent IFF:



$$P(A) = P(A|B) \text{ or } P(B) = P(B|A) \text{ or } P(A,B) = P(A)P(B)$$

Example:



$$P(\text{Toothache}, \text{Cavity}, \text{Catch}, \text{Weather}) = P(\text{Weather})\,P(\text{Toothache}, \text{Cavity}, \text{Catch})$$

(absolute independence is powerful but rare)

Complex systems have hundreds of variables, none of which are independent.

Independence is a big advantage in terms of number of parameters, since you reduce the number of these parameters:

> In general if you have n random variables and you assume that the joint prob. has 2^n entries, if you assume independence, you reduce the number of parameters to n, from exponential to linear

Conditional independence

This ensure independence between random variables when something happens, is not an absolute independence like the previous one.

$$P(X, Y \mid Z) = P(X \mid Y, Z) = P(X \mid Z)\, P(Y \mid Z)$$

Two random variables are conditionally independent given Z iff $P(X \mid Y, Z) = P(X \mid Z)$.

When Z is known Y does not contribute anymore to the estimation of $P(X)$.

When the conditional probability is true, we can simplify also the joint probability:

$$P(X, Y, Z) = P(X \mid Y, Z)\, P(Y, Z) = P(X \mid Y, Z)\, P(Y, Z) P(Z)$$
$$= P(X \mid Z)\, P(Y \mid Z)\, P(Z)$$

What is interesting in ML is to apply this property to n random variables, assumed to be conditional independent given some Z

This is a big simplification, otherwise in the chain rule, if we don't exploit the conditional independence, each term would have a joint effect!

$$\Rightarrow P(x_1)\, P(x_2 \mid x_1)\, P(x_3 \mid x_1, x_2)\, P(x_4 \mid x_1, x_2, x_3) \dots$$

The use of conditional independence reduces the size of the representation of the joint distribution from EXP in n to LINEAR in n.

Bayes' Rule

Product rule $P(a,b) = P(a|b)\,P(b) = P(b|a)\,P(a)$

Bayes Rule: $P(a|b) = \dfrac{P(b|a)\,P(a)}{P(b)}$

Bayes' rule is important because it allows to revert the cond. and prob. distribution

↓

One way may be easier than the other

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

→ Easy to model, it is easy to generate our experiment in which I know the cause and measure the effect

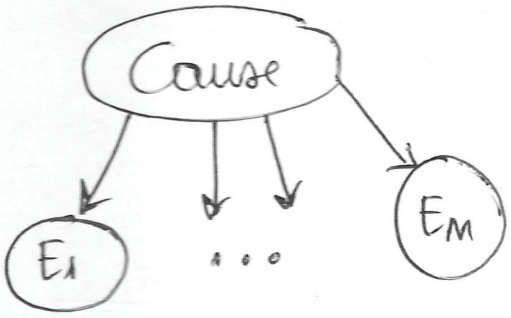
Effects of CONDITIONALLY INDEPENDENCE:

$$P(Cause|E_1,\dots,E_n) = \alpha\, P(Cause) \prod_i P(E_i|Cause)$$

The total number of parameter is n, this eq. is NOT TRUE in general

Bayes Network

Sometimes is useful to have a graphical representation of random variables.



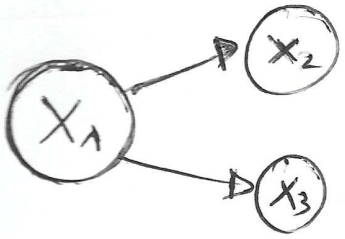
EDGES = There is an effect on the prob. distributions

NODES = Random variables

IS INDEPENDENT BECAUSE IS DISJOINT.

A bayes network is a simple, graphical notation for conditional independence assertions and hence FOR COMPACT SPECIFICATION OF FULL JOINT DISTRIBUTIONS. The topology of the network encodes conditional indepence assumptions.

Another structure that we have can be the following:



All the variables are dependent each other, X_2 depends on X_1, X_3 depends on X_1 and indirectly depends on X_2.

When X_1 is KNOWN the X_2 and X_3 ARE CONDITIONALLY INDEPENDENT!

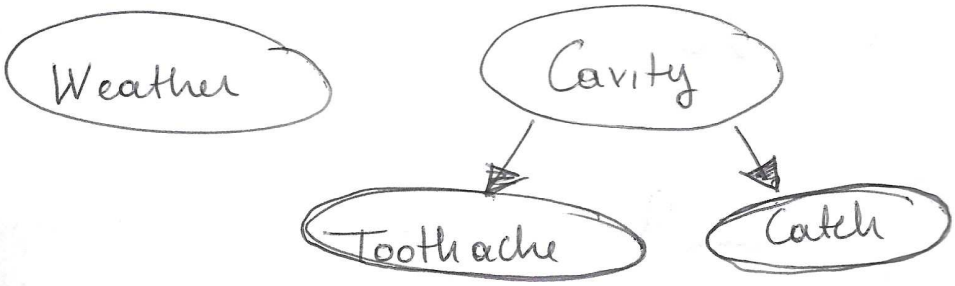
THIS IS NOT TRUE in general.

In general what we need to express with a prob. model given a BN, for any random variable I have to compute:

$$P(X_i \mid \text{DIRECT-PARENTS}(X_i))$$

when there is no parents we have just a priori probability.

Let's see the example!



$P(\text{weather}) =$ Sunny, Cloudy, Rainy, Snow

4 parameter, but I have to compute 3 independent parameter since the sum should be one!

The 4TH : 1 − sum of other parameters.

$P(\text{cavity}) = T/F$ 1 indep. parameter.

$P(\text{catch} \mid \text{cavity}) =$ 2×2 matrix (both are binary random variables)

2 independent parameters

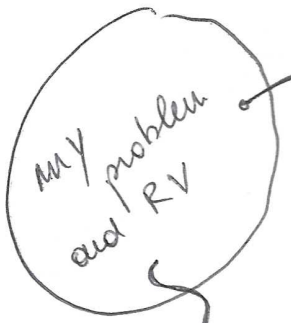
the same for $P(\text{Toothache} \mid \text{Cavity})$

SIZE OF THE MODEL: # independent parameters!

$3 + 1 + 2 + 2 = 8$

Summary!

We have a problem and we can define some random that describe the problem. In general, if I need to compute properties I have to ask questions about these random variables; TWO WAYS:



THEORETICAL WAY

TOO complex!
2^n parameters (EXPONENTIONAL)

compute the joint probability of the variables

$P(X_1, \ldots, X_n)$ then you can answer any query (remember what written before)

§§ ← WITH CONDITIONAL INDEP. WE APPROXIMATE THE JOINT DISTR.

approximate (hopefully to be true)

MAKE ASSUMPTIONS

We are not sure of our assumptions, such as conditional independence!

Sometimes we make this assump. of cond. indepen. even when it is not true! WE COMPUTE A FEW CONDITIONAL DEPENDENCIES

At the end we are not interested in computing the real prob. distribution but just the argmax or max, we ARE NOT INTERESTED INTO NUMBER. Even if probabilities are very different, THE ARGMAX CAN BE SIMILAR.

→ Given a problem:

- identify random variables
- identify reasonable assumptions
- simplify the model
- find a solution

CLASSIFICATION as PROBABILISTIC ESTIMATION

Given a target function $f : X \to V$ and a dataset D, I want to compute an approximation function $\hat{f}$ that gives the best prediction of all instances, especially for instances not in the dataset D.

$$\hat{f}(x') = v^* \quad , \quad v^* = \underset{v \in V}{\arg\max} \; P(v \mid x', D)$$

The value of the new instance x' and the dataset D are known! The argmax gives the best approximation. In general we may want to compute the probability distribution over V:

$$\boxed{P(V \mid x', D)}$$

Given a Dataset D and hypothesis space H, we can compute:

$P(H \mid D)$: the probability of each single hypothesis has generated this dataset.

By applying the Bayes rule:

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

$P(D)$ ← normalization factor.