Statistical Learning on graphs

Probabilistic Graphical Modelling

Presented by Levente Torok

General Electric Digital / Peter Pazmany Catholic University

15th September, 2016

Content

- Basic probability calculus
  - Understanding probability
  - pdf, pmf, cdf, join distributions
  - Product rule, sum rule, Bayes rule
  - Marginalization, Inference
- Directed Acyclic Graphs of probablity variables
  - Inference, evidental reasoning
  - Effects: D-separation, explaining away
  - Markov blanket

Probability defined by truth table

Flu	Cough	Probability
1	0	0.01
1	1	0.04
0	0	0.855
0	1	0.095

Probability defined by truth table

Flu	Cough	Probability
f	$\backslash c$	0.01
f	c	0.04
$\backslash f$	$\backslash c$	0.855
$\backslash f$	c	0.095

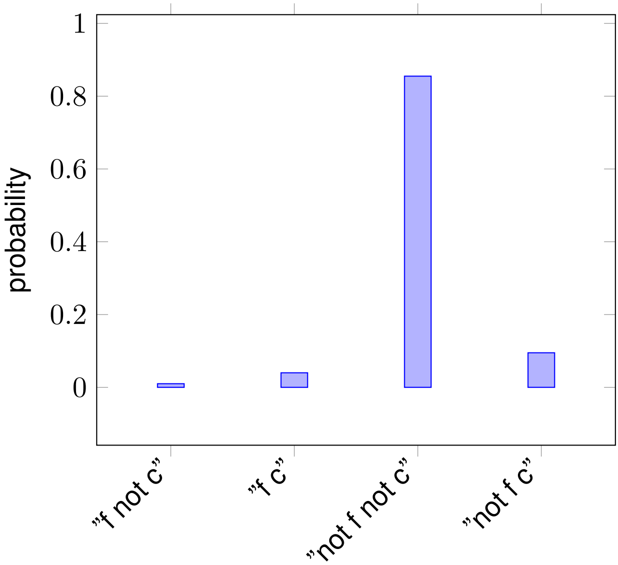
Note that:

$$P(f \wedge c) \neq P(F, C)$$

$$P(F = f, C = c) = P(f, c) \neq P(F, C)$$

This $P(F, C)$ is called **joint probablity distribution.**

Probability Mass Function



Sum Rule / Marginalization

Question

What is the probability that I am coughing?

$$P(c) = P(c, f) + P(c, \backslash f)$$

Or in general

$$P(x) = \sum_{y' in \mathbb{Y}} P(x, y') = \int_{y'} P(x, y')$$

Sum Rule / Marginalization

Flu	Cough	Probability
f	$\backslash c$	0.01
f	c	0.04
$\backslash f$	$\backslash c$	0.855
$\backslash f$	c	0.095

$\Longrightarrow$

Cough	Probability
$\backslash c$	0.865
c	0.135

Or in general

Probability	c	$\backslash c$	Marginal
f	0.04	0.01	$P(f) = 0.05$
$\backslash f$	0.095	0.855	$P(\backslash f) = 0.95$
Marginal	$P(c) = 0.135$	$P(\backslash c) = 0.865$	

Bayes rule

Question

Knowing that I am coughing, how likely is that I have flew?

$$P(f|c) = \frac{P(f,c)}{P(c)},$$

where

$$P(c) = P(\backslash f, c) + P(f, c)$$

This is called conditional probability.

Maximum Aposteriori (MAP) decisions

Why are doctors so expensive?

- Let's denote symptoms with x, sickness with y
- Student books contain $P(x|y)$
- But what we really need is $y^* = \arg\max_y P(y|x)$
- How to get there? $P(y|x) = \frac{P(x|y)P(y)}{P(x)}$
- Since we need only the $\arg\max$ of y, $P(x)$ are just constant in the denumerator $P(y|x) \propto P(x|y)P(y)$
- So $\arg\max_y P(y|x) = \arg\max_y P(x|y)P(y)$
- So we are paying for $P(y)$ what we call "experience".

Product rule

From Bayes rule

$$P(c|f) = \frac{P(c,f)}{P(f)}$$

we can get

$$P(c,f) = P(c|f)P(f),$$

where

$$P(f) = P(f,c) + P(f,\backslash c)$$

Chain rule

From Bayes rule

$$P(c|f) = \frac{P(c,f)}{P(f)}$$

we can get

$$P(c,f) = P(c|f)P(f)$$

Let's assume that we have d, e as other probability variables, then one can come up with the joint distribution as

$$P(c,d,e,f) = P(e|c,d,f)P(d|c,f)P(c|f)P(f)$$

The complete model

$$P(a, b, c, d, ..., x, y)$$

From the **joint distribution** every marginal and conditional can be derived. Which doesn't hold vice versa. So the most valuable distirbution we can have is always the joint distribution.
However, it is requires **exponentially large number of samples** to represent the complete joint. So we are looking for simplifications.

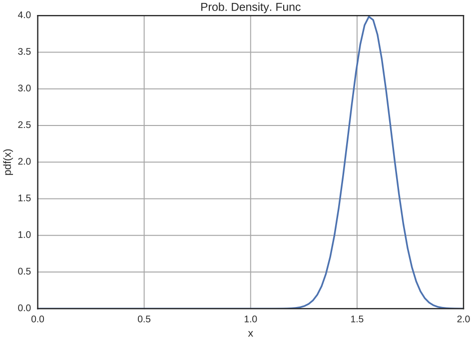
Excercise

Please build a sampling machine that

- will generate samples of tuples, describing if our client has a cough and/or flu proportional to the joint distribution
- given the **joint distribution**

Continuous variables

Lets have a class of students. Their height can be modelled by Gaussian distribution for example. $P(X) = \mathcal{N}(\mu, \sigma)$



Question

What is the probability that a sample member of the class has height of $1.7m$?

Continuous variables

Answer:

$$P(X = 1.7) = 0$$

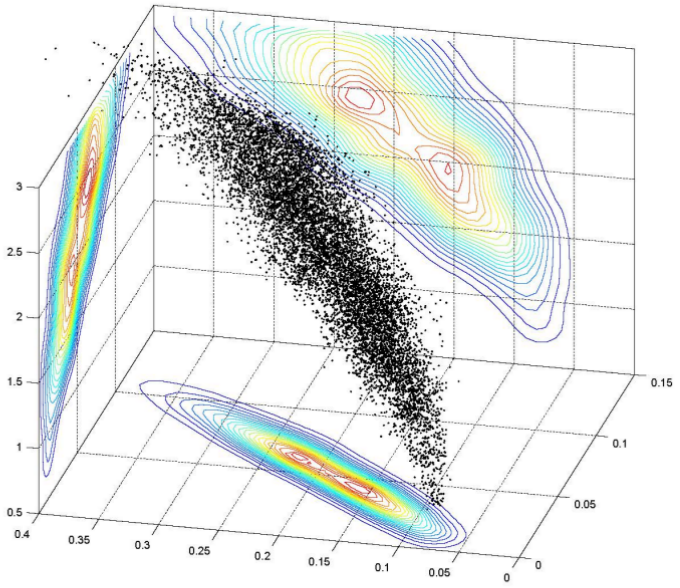
Since the pmf of a continuous probability variable is $= 0$. But pdf is not.

$$\int_a^b pdf(x)\,dx = P(a < x < b)$$

Conclusion:

$$pmf \neq pdf$$

Joint trivariate



Independence

Criteria

$$p(x, y) = p(x)p(y)$$

which also means

$$p(x|y) = p(x)$$

since we have no information gain by knowing y
For example playing dice. Each drawing is independent. So that, knowing the result of previous draws, doesn't help knowing the result of the next draw,

$$p(d_1|d_2)p(d_2) = p(d_1)p(d_2)$$

This is usually denoted as $x \perp y$, if x and y are independent.

Question

If x, y are independent, $p(x, y) = p(x)p(y)$ holds.
But is it true, vice versa?

Question 1
Are the length of 2 arbitrary pencils (l_1, l_2) dependent or independent?

Since they are roughly of the same length, observing the length of one pencil (l_1) will provide a information about the length (l_2) of the other pencil. So they seem to be dependent.

$$p(l_1, l_2) = p(l_2|l_1)p(l_1) \neq p(l_2)p(l_1)$$

Question 2
However if I know, that these pencils come from the same factory and should be produced with the same length μ (with Gaussian inaccuracies), by knowing l_1 will not effect my estimate on l_2.

$$p(l_i|L) = \mathcal{N}(\mu, \sigma)$$

so

$$p(l_1, l_2|L) = p(l_1|L)p(l_2|L)$$

So they became independent. This is called **conditional independence**. This is usually denoted by

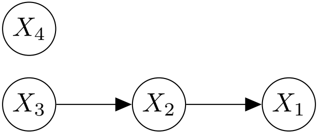
$$x \perp y | z$$

Directed graphical models

Given data samples one can figure out that not every possilbe dependence holds in between variables, so one can remove from the complete factorization the unneccessary elements and get a result as this

$$P(X_1, X_2, X_3, X_4, X_5) = p(X_1|X_2)p(X_2|x_3)p(X_3)p(X_4),$$

for example. This can be represented in directed acylic graphs (DAG) as



D-separation

u and v are D-separated, if and only if all paths in between them are "blocked" by observing "m" in terms of conditional independence.

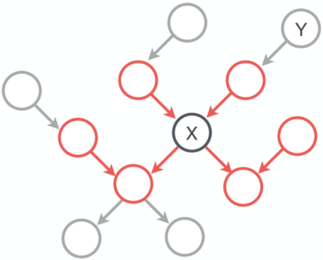
$$u \perp v | m$$

Markov blanket

Markov Blanket is a collection of such m's that, for a specific node X, no effect can reach (ie. m is d-separating) from any Y nodes.

$$\forall Y : X \perp Y \mid MB(X)$$

- parents
- children
- parent of childrens



Probabilistic graphical models (PGM)

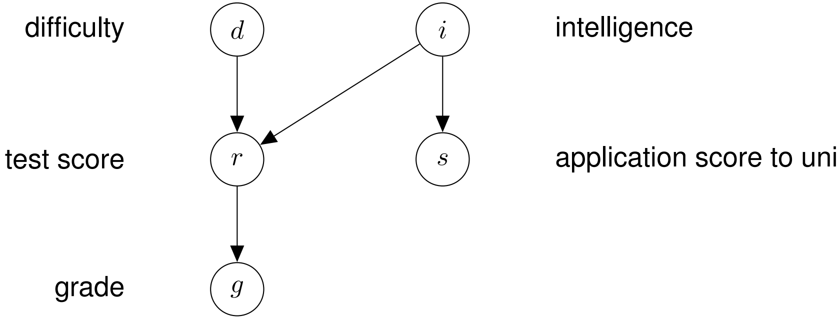
Joint distributions of DAGs can always be written as

$$P(X_1, X_2, X_3, X_4, X_5) = \sum_i P(X_i | parent(X_i))$$

Note that

- The distribution can be factorized into simpler distributions according to the DAG
- The graph encodes all the independencies
- Independent relations can be read from the graph

Examples


difficulty
d
i
intelligence
test score
r
s
application score to uni
grade
g

Excercises

Question
How can we define a distribution for describing a PGM on discrete probablity variables? (CPD)

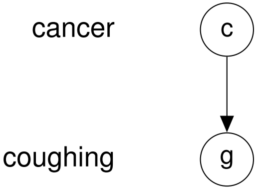
Excercise Generate data samples from a model described by the model above.

Excercise 2 Estimate the CPD tables from these samples.
Excercise 3 Write down, how to estimate parent distribution from samples.

Relevant questions in graphical models

- Fit model to data: Find the best $P(x_i|Parent(x_i))$
- Marginalization $P(x)$ (visualization)
- Inference $P(x|y)$ (supervised usage)

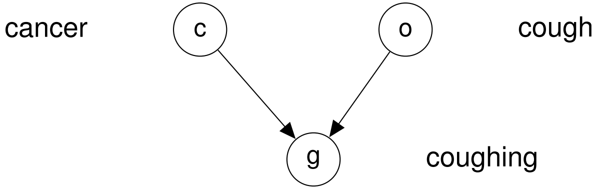
Reasonings



- Causal reasoning: $P(coughing|cancer) = ?$
- Evidental reasoning: $P(cancer|coughing) = ?$

Explaining Away

Some calls it ”Inter-causal reasoning”



$$P(cancer|coughing) \neq P(cancer|coughing, cough)$$

1. I am coughing, may I have cancer?
2. I am coughing and have cough, I am almost sure that I don’t have cancer.