

# Advanced probabilistic methods

## Lecture 1

Pekka Marttinen

Aalto University

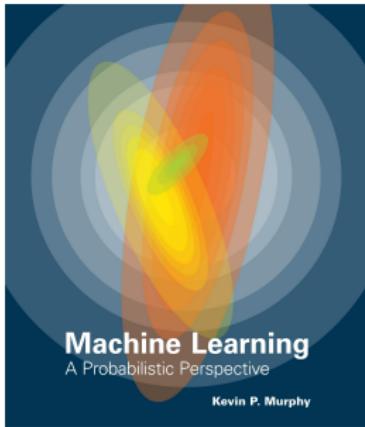
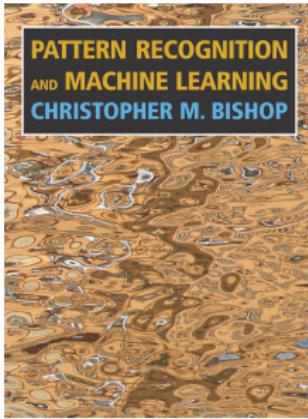
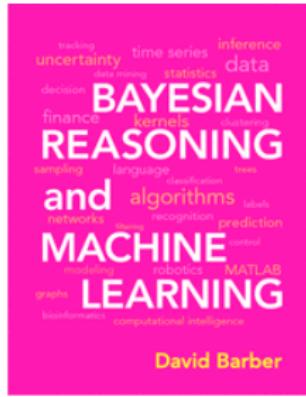
January, 2021

# Lecture 1 overview

- Practical matters
  - Structure, grading, workload, online implementation
  - Exercise format
  - Student feedback from previous years
- Course overview
- Basic probability calculus (Barber, Ch. 1)
- Basic graph concepts

- Structure
  - Lectures  $9 \times 2$  hours
  - Exercise sessions  $9 \times 2$  hours
  - See *Timetable* in *myCourses/Materials*
- Grading based on
  - Exam: 55% of the total weight
  - Each assignment: 5% of the weight (sum 45%)
  - *Preliminary boundaries:* 1:50%, 2:60%, 3:70%, 4:80%, 5:90%

# Course books



- *Bayesian Reasoning and Machine Learning* available at [www.cs.ucl.ac.uk/staff/D.Barber/brml](http://www.cs.ucl.ac.uk/staff/D.Barber/brml)
- *Pattern Recognition and Machine Learning* available at [www.microsoft.com/en-us/research/people/cmbishop/](http://www.microsoft.com/en-us/research/people/cmbishop/)

# Estimated workload

- Lectures:  $9 \times 2\text{h}$
- Preparation for lectures, reading the book ( $\sim 200$  pages):  $9 \times 4\text{h}$
- Exercise sessions:  $9 \times 2\text{h}$
- Doing the exercises:  $9 \times 5\text{h}$
- Preparing for the exam:  $14\text{h}$
- Exam:  $4\text{h}$
- Total  $135\text{h}$ . As credits  $135/27 = 5\text{cr.}$

# Online implementation in 2021

- The lectures are recorded and released in advance
- Students can post questions about lectures in **Slack**.
- Each lecture is followed by an online Q&A session in Zoom. The lecturer will go through questions related to the lectures posted in the Slack and the students can also ask additional questions.
- Links to Slack and Zoom will be posted in MyCourses.

# Exercises

- Exercises must be returned by the deadline using JupyterHub.
  - See instructions in MyCourses.
  - Grading principle: **2p**→done, almost correct; **1p**→done, but something clearly missing/incorrect; **0p**→not done or completely incorrect (may be modified case-by-case).
  - Exercises are graded by the TAs, not corrected → Make sure you know the correct answer afterwards by attending the exercise sessions or checking the model solutions.
- Exercise session format
  - Help for getting started with next week's exercises
  - Possibility to ask about next week's exercises or previous week's solutions
  - Two assistants present

# Exercises

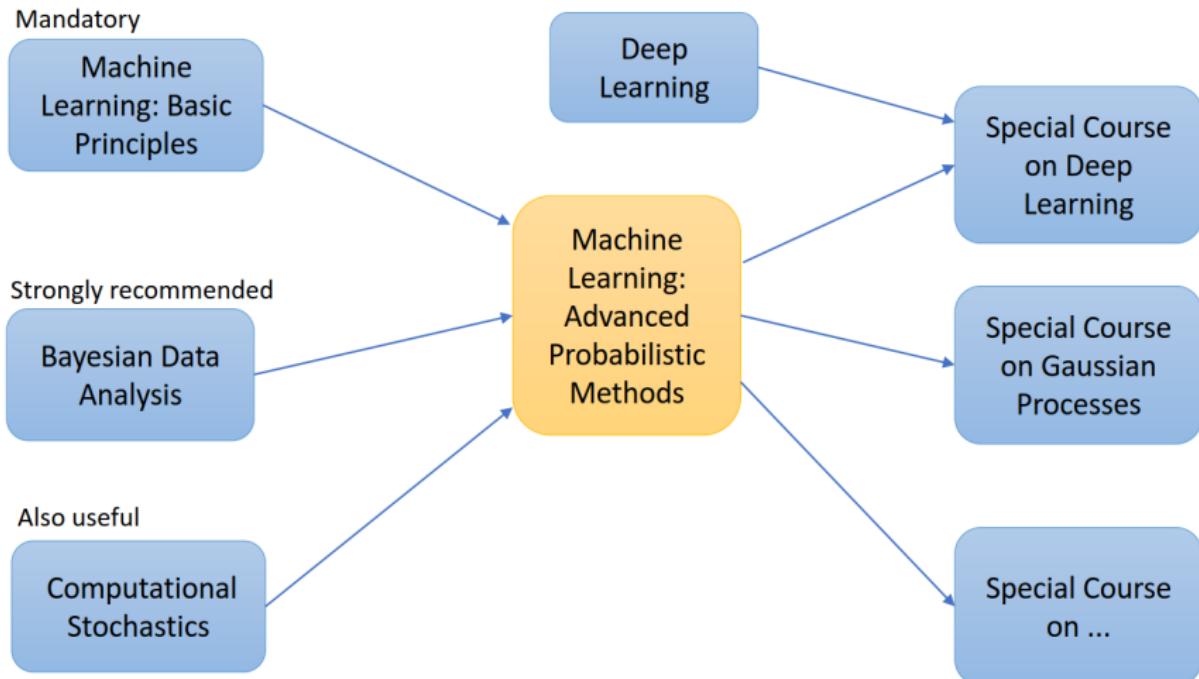
- Slack

- Write a question in a dedicated Slack channel. The TAs will answer them at the times of the exercise sessions (possibly at other times, see details in MyCourses)
- Students are welcome to comment and give hints to each other's questions in the Slack; however, do not reveal the full answer.

- Other policies

- Collaborating with your colleagues to solve the exercises is allowed and encouraged, but in the end you are expected to write your own solution (copy-paste from someone else is not allowed and will be checked for).
- Sharing solutions online or use of any such material is prohibited. Suspected violations will be reported for further investigation.

# Relation to other courses



# Student feedback from previous years

- About prerequisites

- '*Bayesian Data Analysis should be listed as a prerequisite*'  
→ See previous slide.

- About lectures

- '*One of the courses where lecture attendance is really beneficial, I would definitely point this out in the intro lecture as students are trying to optimize their time between other lectures and exercise sessions*'
  - '*All lectures could be recorded*'  
→ This will be done.

# Student feedback from previous years

- About exercises (positive)

- *'Participating in the exercise sessions was extremely useful and the course assistants were great'*
- *'The exercises really helped me to understand the concepts related to this course, and I think that they played a huge role in my learning'*
- *'The exercise sessions were very good. Without them I would never have been able to achieve that far.'*

- About exercises (things to improve)

- *'A slack group would have been useful'*  
→ This will be implemented.

# Student feedback from previous years

- About difficulty in general

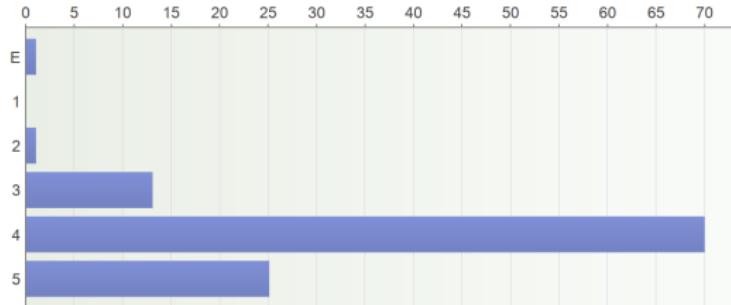
- *'At least to me there was steep curve in the topic difficulty starting from lecture 5. Compress lectures 1-4 and spend more time on the "new stuff"*
  - Some changes in the material to this direction.
  - Take a full advantage of the Exercise sessions (also and especially on the 2nd half).
  - Ask clarifications in Slack and the Q&A session.

- About contents

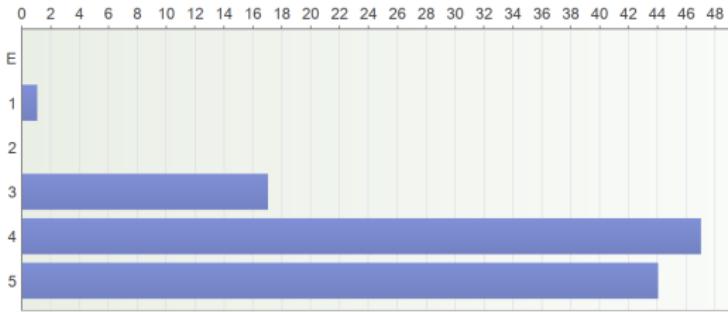
- *'Add a small project'*
  - We used to have one, but dropped that to keep the workload manageable.
- *'How were these topics chosen? Does the course aim to practical work life skills, or prepare for research career or is it just general background knowledge?'*
  - See slides *Course contents*, *Course outline* and *What to expect below*.

# Student feedback from 2020, overview

- Overall assessment

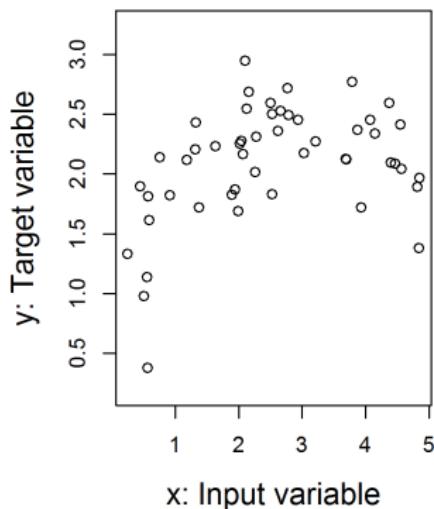


- I will benefit from things learnt on the course



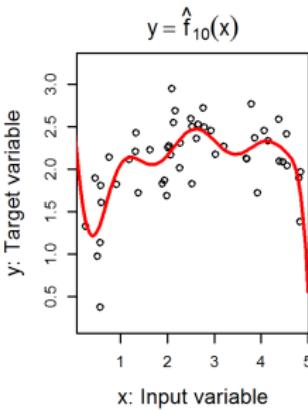
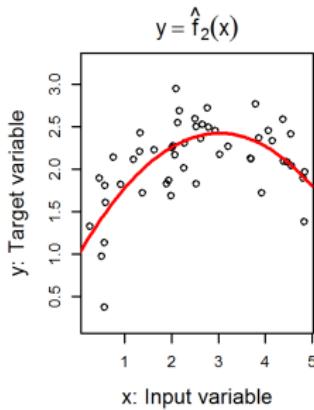
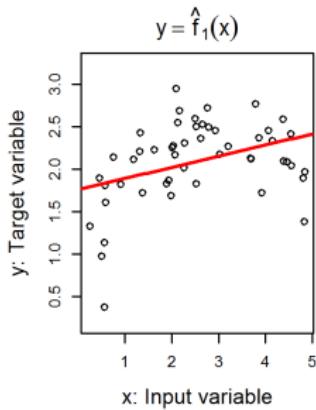
# Probabilistic modeling overview (1/2)

- The goal of **probabilistic modeling** is to answer a question about the data:
  - Classify the samples into groups
  - Create prediction for future observations
  - Select between competing hypotheses
  - Estimate a parameter, such as the mean, of the population
  - ...
- and **quantify uncertainty** using probabilities.



# Probabilistic modeling overview (2/2)

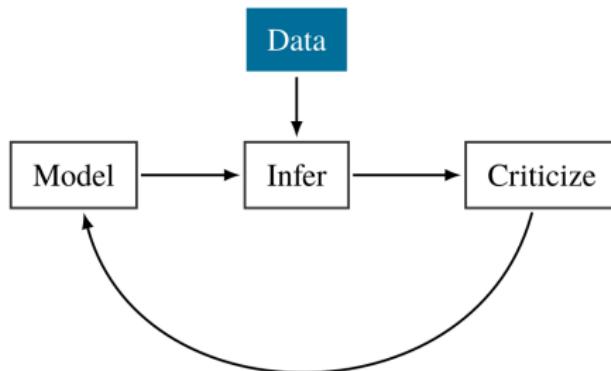
- Probabilistic modeling in a nutshell
  - ① Select a **model**
  - ② Infer the parameters of the model (train/fit the model)
  - ③ Use the fitted model to answer the question of interest
- Usually several models are considered, requiring **model selection**.
- For example:  $f_n(x) = a_0 + a_1x + a_2x^2 + \dots + a_nx^n$



# Course contents

- Ingredients of probabilistic modeling

- **Models:** Bayesian networks, Sparse Bayesian linear regression, Gaussian mixture models, latent linear models
- **Methods for inference:** maximum likelihood, maximum a posteriori (MAP), Laplace approximation, expectation maximization (EM), Variational Bayes (VB), Stochastic variational inference (SVI)
- **Ways to select between models**



Box's loop (Blei, 2014)

# Course outline

- “Probabilistic Modelling Course” (the last column) from the preface of Barber’s book is used as the backbone.

Deterministic approximate inference techniques (28) have been added to this, and correspondingly less weight is given to some other topics.

Part I:  
Inference in Probabilistic Models

- 1: Probabilistic Reasoning
- 2: Basic Graph Concepts
- 3: Belief Networks
- 4: Graphical Models
- 5: Efficient Inference in Trees
- 6: The Junction Tree Algorithm
- 7: Making Decisions

Part II:  
Learning in Probabilistic Models

- 8: Statistics for Machine Learning
- 9: Learning as Inference
- 10: Naive Bayes
- 11: Learning with Hidden Variables
- 12: Bayesian Model Selection

Part III:  
Machine Learning

- 13: Machine Learning Concepts
- 14: Nearest Neighbour Classification
- 15: Unsupervised Linear Dimension Reduction
- 16: Supervised Linear Dimension Reduction
- 17: Linear Models
- 18: Bayesian Linear Models
- 19: Gaussian Processes
- 20: Mixture Models
- 21: Latent Linear Models
- 22: Latent Ability Models

Part IV:  
Dynamical Models

- 23: Discrete-State Markov Models
- 24: Continuous-State Markov Models
- 25: Switching Linear Dynamical Systems
- 26: Distributed Computation

Part V:  
Approximate Inference

- 27: Sampling
- 28: Deterministic Approximate Inference

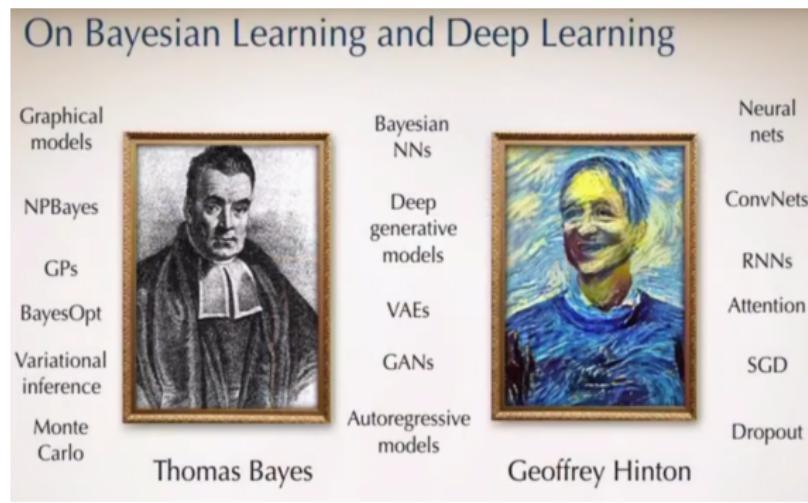


# What to expect on the course

- More emphasis on
  - fundamental principles of probabilistic modeling and machine learning
  - introducing simple models that serve as building blocks for more complex (or realistic) models
  - building a detailed understanding of the algorithms through simple examples, which allows us to focus on the principles
  - enhancing mathematical skills to operate with probabilities
- Less emphasis on
  - hands-on learning on how to use a specific software package to fit 'advanced' models to a real-world data set (though we'll use *PyTorch* in some exercises to show how things are *nowadays* done 'in real life')
- Goal
  - knowledge needed to start understanding and applying probabilistic modeling using existing software packages, and those to come

# Role of probabilistic machine learning today

- Keynote at *NeurIPS 2017* by Yee Whye Teh



<https://www.youtube.com/watch?v=9saauSBgmcQ>

- *NeurIPS 2020*: articles with a keyword in the title: *Bayes* (n=63), *Uncertainty* (n=18), *Variational* (n=36). Many of these coming from the major IT companies.

# Probability theory, basics

- Marginalization
- Independence
- Conditional distribution
- Conditional independence
- Continuous random variables

(To recap these, see *Additional Reading* in *myCourses/Materials*)

## Notation (1/2)

- Random variables:  $X, Y, Z, \dots$
- Values these random variables can take:  $x, y, z, \dots$
- Probability
  - The following notations are used interchangeably

$$p(X = x) = p_X(x) = p(x)$$

- All are interpreted as the probability that variable  $X$  is in state  $x$

## Notation (2/2)

- Domain
  - $\text{dom}(X)$  denotes all possible states for variable  $X$ .
- Distribution of a variable  $X$  consists of
  - its domain  $\text{dom}(X)$
  - and full specification of probability values  $p_X(x)$ , for all possible  $x \in \text{dom}(X)$
- Normalization
  - The summation over all the states

$$\sum_{x \in \text{dom}(X)} p(X = x) = 1$$

- The sum can be written as:  $\sum_x p(x) = 1$

## Example - probability table

$B$	$M$	$K$	$p(b, m, k)$
1	1	1	0.012
1	1	0	0.108
1	0	1	0.288
1	0	0	0.192
0	1	1	0.016
0	1	0	0.064
0	0	1	0.096
0	0	0	0.224

- The probability table lists the probabilities of all possible combinations of the random variables.

- The *joint* distribution of  $B$ ,  $M$  and  $K$
- For example

$$p_{B,M,K}(1, 1, 0) = p(B = 1, M = 1, K = 0) \\ = 0.108$$

- Modified from Example 1.3 "Inspector Clouseau"  
 $M$  = 'Maid is the murderer'  
 $B$  = 'Butler is the murderer'  
 $K$  = 'Knife is the murder weapon'

# Marginalization

- Given a joint dist  $p_{X,Y}(x,y)$ , the marginal dist of  $X$  is defined by

$$p_X(x) = \sum_y p_{X,Y}(x,y)$$

- More generally,

$$p(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) = \sum_{x_i} p(x_1, \dots, x_n)$$

## Example - marginalization (1/2)

$B$	$M$	$K$	$p(b, m, k)$
1	1	1	0.012
1	1	0	0.108
1	0	1	0.288
1	0	0	0.192
0	1	1	0.016
0	1	0	0.064
0	0	1	0.096
0	0	0	0.224

- What is the marginal distribution of  $B$  and  $M$ ?
- We need to compute  $p_{B,M}(b, m)$ , for all possible  $b$  and  $m$ .

## Example - marginalization (2/2)

- Use

$$p_{B,M}(b, m) = \sum_{k=0}^1 p_{B,M,K}(b, m, k)$$

- For example:

$$\begin{aligned} p_{B,M}(0, 0) &= p_{B,M,K}(0, 0, 0) + p_{B,M,K}(0, 0, 1) \\ &= 0.096 + 0.224 = 0.32 \end{aligned}$$

- Doing this for all  $B, M$  combinations, we get the marginal probability table

$B$	$M$	$p(b, m)$
1	1	0.12
1	0	0.48
0	1	0.08
0	0	0.32

# Independence

- Random variables  $X$  and  $Y$  are independent if

$$p_{X,Y}(x,y) = p_X(x)p_Y(y)$$

for all  $x$  and  $y$ .

- Intuitively, this means that knowing the value of  $X$  does not provide any information about the value of  $Y$ .
- Notation:  $X \perp\!\!\!\perp Y$
- More generally:  $\mathcal{A} = \{A_1, \dots, A_k\}$  and  $\mathcal{B} = \{B_1, \dots, B_l\}$  are independent if

$$\begin{aligned} & p_{A_1, \dots, A_k, B_1, \dots, B_l}(a_1, \dots, a_k, b_1, \dots, b_l) \\ &= p_{A_1, \dots, A_k}(a_1, \dots, a_k)p_{B_1, \dots, B_l}(b_1, \dots, b_l) \end{aligned}$$

## Example - Independence (1/2)

$B$	$M$	$p(b, m)$
1	1	0.12
1	0	0.48
0	1	0.08
0	0	0.32

- Are  $B$  and  $M$  independent?

## Example - Independence (2/2)

- Marginal distributions

$B$	$p(b)$	$M$	$p(m)$
1	0.6	and	1 0.2
0	0.4		0 0.8

- Direct computation gives

$B$	$M$	$p(b)p(m)$	$p(b, m)$
1	1	0.12	0.12
1	0	0.48	0.48
0	1	0.08	0.08
0	0	0.32	0.32

- Hence,  $B$  and  $M$  are (marginally) independent

# Statistical vs. causal independence

- $D$ ='number of people drowned',  $A$ ='amount of ice-cream sold'
  - Are  $D$  and  $A$  independent?
  - Are  $D$  and  $A$  causally dependent?

# Conditional distribution

- Conditional distribution

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

specifies the probability of each possible value  $x$  of  $X$  given that we have observed variable  $Y$  in state  $y$ .

## Example - Conditional distribution

- For example:

$$p(K = 1 | B = 1, M = 1) = \frac{p(B = 1, M = 1, K = 1)}{p(B = 1, M = 1)} = 0.1$$
$$p(K = 0 | B = 1, M = 1) = 0.9$$

- All conditional probabilities in the last column

B	M	K	$p(b, m, k)$	$p(k b, m)$
1	1	1	0.012	0.1
1	1	0	0.108	0.9
1	0	1	0.288	0.6
1	0	0	0.192	0.4
0	1	1	0.016	0.2
0	1	0	0.064	0.8
0	0	1	0.096	0.3
0	0	0	0.224	0.7

# Conditional independence

- $X \perp\!\!\!\perp Y|Z$  denotes that variables  $X$  and  $Y$  are conditionally independent of each other, given the state of variable  $Z$ . This is formally defined by condition

$$p_{X,Y|Z}(x,y|z) = p_{X|Z}(x|z)p_{Y|Z}(y|z)$$

for all states  $x, y, z$  of variables  $X, Y, Z$ .

- Intuitively, this means that if we know the value of  $Z$ , knowing in addition the value of  $Y$  does not provide any information about the value of  $X$ . Indeed, provided  $p(y,z) > 0$ , we have

$$X \perp\!\!\!\perp Y|Z \implies p_{X|Y,Z}(x|y,z) = p_{X|Z}(x|z)$$

# Conditional independence

$$X \perp\!\!\!\perp Y|Z \implies p_{X|Y,Z}(x|y,z) = p_{X|Z}(x|z)$$

- Proof

$$\begin{aligned} p(x|y,z) &= \frac{p(x,y,z)}{p(y,z)} = \frac{p(x,y|z)p(z)}{p(z)p(y|z)} \\ &= \frac{p(x|z)p(y|z)p(z)}{p(z)p(y|z)} = p(x|z) \end{aligned}$$

- The general *chain rule of probability*

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

follows from iterative use of the definition of conditional probability.

## Example - Conditional independence (1/3)

$B$	$M$	$K$	$p(b, m, k)$
1	1	1	0.012
1	1	0	0.108
1	0	1	0.288
1	0	0	0.192
0	1	1	0.016
0	1	0	0.064
0	0	1	0.096
0	0	0	0.224

- Are  $M$  and  $B$  conditionally independent, given  $K$ ?
  - We need to compare
    - $p_{M|K}(m|k)p_{B|K}(b|k)$
    - $p_{B,M|K}(b, m|k)$
- for all  $m, b, k$ .

## Example - Conditional independence (2/3)

- For example,

$$\begin{aligned} p(B = 1, M = 1 | K = 1) &= \frac{p(B = 1, M = 1, K = 1)}{p(K = 1)} \\ &= \frac{0.012}{0.012 + 0.288 + 0.016 + 0.096} \approx 0.0291 \end{aligned}$$

- Similarly,

$$\begin{aligned} p(M = 1 | K = 1) &= \frac{p(M = 1, K = 1)}{p(K = 1)} \\ &= \frac{0.012 + 0.016}{0.012 + 0.288 + 0.016 + 0.096} \approx 0.0508 \end{aligned}$$

and

$$p(B = 1 | K = 1) = \dots \approx 0.7110$$

## Example - Conditional independence (3/3)

$B$	$M$	$K$	$p(b, m k)$	$p(b k)$	$p(m k)$	$p(b k)p(m k)$
1	1	1	<b>0.029</b>	<b>0.711</b>	<b>0.051</b>	<b>0.036</b>
0	1	1	...	...	...	...
1	0	1				
0	0	1				
1	1	0				
0	1	0				
1	0	0				
0	0	0				

- Because  $0.029 \neq 0.036$ , it follows that  $B$  and  $M$  are not conditionally independent given  $K$ .

# Intuition for independence and conditional independence (1/2)

- Let  $X_1, X_2, \dots, X_n$  denote the cumulative sum of  $n$  dice throws, such that  $\text{dom}(X_1) = \{1, \dots, 6\}$ ,  $\text{dom}(X_2) = \{2, \dots, 12\}$ , etc.
  - Is  $X_{n+1}$  independent of  $X_{n-1}$ ?
  - Is  $X_{n+1}$  conditionally independent of  $X_{n-1}$  given  $X_n$ ?
- $X = \text{'Location of an airplane now'}$ ,  $Y = \text{'Location of the plane 15s ago'}$ ,  
 $Z = \text{'Location 15s from now'}$ 
  - Is  $Y$  independent of  $Z$ ?
  - Is  $Y$  conditionally independent of  $Z$  given  $X$ ?

# Intuition for independence and conditional independence (2/2)

- $S$ ='sunshine',  $D$ ='number of people drowned',  $A$ ='amount of ice-cream sold'
  - Are  $D$  and  $A$  independent?
  - Are  $D$  and  $A$  conditionally independent given  $S$ ?
- $A$ ='The alarm is on',  $B$ =There is a burglar in the house",  $T$ ='A truck passes the house'
  - Suppose that the alarm can be triggered either by a burglar or by a passing truck
  - Are  $B$  and  $T$  independent?
  - Are  $B$  and  $T$  conditionally independent given  $A$

# Continuous random variables (1/3)

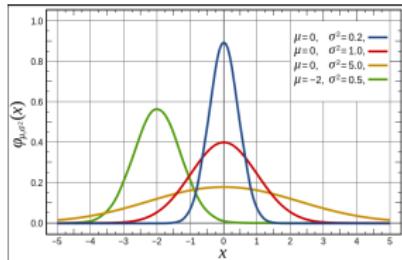
- Probability density function (pdf) for a continuous variable  $X$ ,  $f_X()$

$$\int_{x \in \mathcal{R}} f_X(x) dx = 1$$

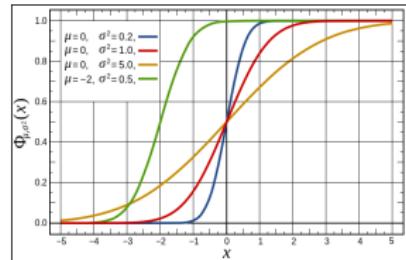
$$p(X \in [a, b]) = \int_{x=a}^b f_X(x) dx$$

- Cumulative distribution function (cdf)

$$F_X(x) = p(X \leq x) = \int_{t=-\infty}^x f_X(t) dt$$



$N(\mu, \sigma^2)$  pdf (Wikip.)



$N(\mu, \sigma^2)$  cdf (Wikip.)

## Continuous random variables (2/3)

- Concepts presented can be generalized to continuous random variables
- Marginalization
  - Discrete:  $p_X(x) = \sum_y p_{X,Y}(x,y)$
  - Continuous:  $f_X(x) = \int_y f_{X,Y}(x,y) dy$
- Expected value
  - Discrete:  $E(X) = \sum_x x p_X(x)$
  - Continuous:  $E(X) = \int_x x f_X(x) dx$

# Continuous random variables (3/3)

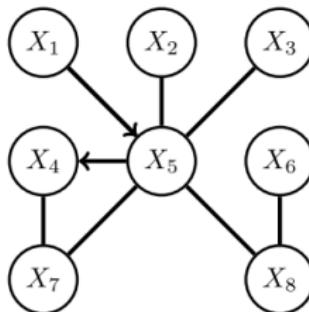
- Conditional distribution

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

- (conditional) independence:  $X \perp\!\!\!\perp Y|Z$ , if

$$f_{X,Y|Z}(x,y|z) = f_{X|Z}(x|z)f_{Y|Z}(y|z)$$

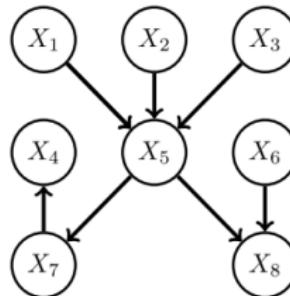
# Basic graph definitions



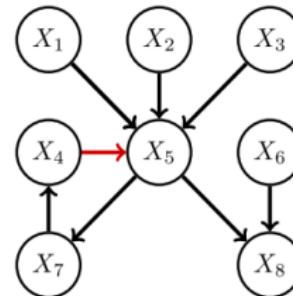
- A **graph** consists of **nodes** (vertices) and **edges** (links) between nodes.
- A path from  $X_i$  to  $X_j$  is a sequence of connected nodes starting at  $X_i$  and ending at  $X_j$ .

# Directed graphs

Directed Acyclic Graph



Directed Cyclic Graph



- A Directed Acyclic Graph (**DAG**) is a directed graph without cycles
- **Parents, Children, Ancestors, Descendants,...** (see Ch. 2)

# Important points

- marginalization
- conditional distribution
- conditional/marginal independence
- probability density function, cumulative distribution function
- Basic graph concepts