Advanced probabilistic methods

Lecture 1

Pekka Marttinen

Aalto University

January, 2019

Lecture 1 overview[1]

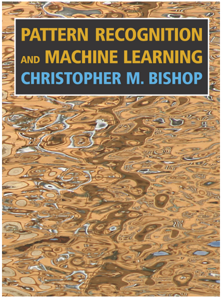
- Practical matters
  - Structure, workload, grading
  - Exercise format
  - Student feedback from 2018
- Course overview
- Basic probability calculus (Barber, Ch. 1)
- Basic graph concepts

[1] These slides build upon the book *Bayesian Reasoning and Machine Learning* and the associated teaching materials. The book and the demos can be downloaded from *www.cs.ucl.ac.uk/staff/D.Barber/brml*.

Course structure

- Structure
  - Lectures 10×2 hours
  - Exercise sessions $(1+9) \times 2$ hours
  - See *Timetable* in *myCourses/Materials*
- Grading based on
  - Exam: 70% of the total weight
  - Exercises: 30%
  - *Preliminary* boundaries: 1:50%, 2:60%, 3:70%, 4:80%, 5:90%
  - Minimum required: 1/2 of the exam points, 1/3 of exercise points

Course books



- Note: *Bayesian Reasoning and Machine Learning* is freely available for download at *www.cs.ucl.ac.uk/staff/D.Barber/brml*

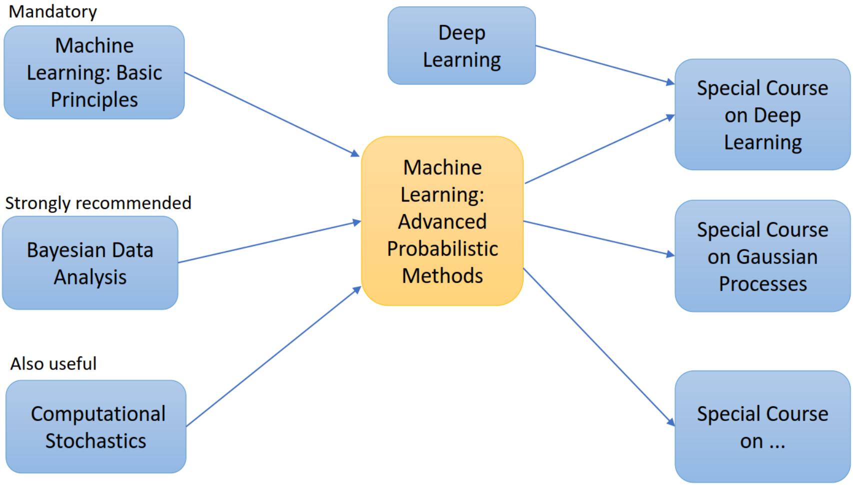
Estimated workload

- Lectures: 10×2h
- Preparation for lectures, reading the book (~ 200 pages): 9×4h
- Exercise sessions: 9×2h
- Doing the exercises: 9×4.5h
- Exercise 0, self-study, introduction to Python: 5.5h
- Preparing for the exam: 11h
- Exam: 4h
- Total 135h. As credits $135/27 = 5$cr.

Exercises

- Exercises must be returned to MyCourses by the deadline
  - A single PDF
  - Grading of the exercises (**2p**$\rightarrow$done, almost correct; **1p**$\rightarrow$done, but something clearly missing/incorrect; **0p**$\rightarrow$not done or completely incorrect)
  - Exercises are graded by the TAs, not corrected $\rightarrow$ Always make sure afterwards you know the correct answer, by attending the exercise sessions or going through 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

Relation to other courses



Student feedback from 2018

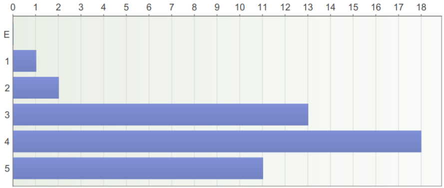
- About prerequisites
  - *'Bayesian Data Analysis should be listed as a prerequisite'*
    →See previous slide.
- About exam
  - *'Exam was difficult'*
    →Exam duration increased from 3 to 4 hours.
    →Overall level of performance was (and will be) taken into account in grading.
    →Questions in the exam will be similar to the exercise questions. Best way to prepare is to do the exercises.
    →Separate recap lecture added to the end of the course.

Student feedback from 2018

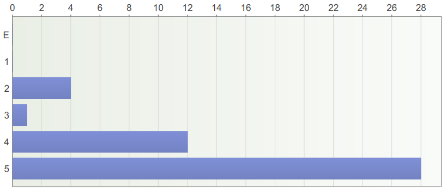
- About difficulty in general
  - *'At least to me there was steep curve in the topic difficulty starting from lecture 5. Compress lectures 1-4 and spend more time on the "new stuff"*
    →Take a full advantage of the Exercise sessions (also and especially on the 2nd half)
    →Ask clarifications on the lectures
- About exercises
  - *'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'*
  - *'TAs from the exercise session were very helpful and stayed often (a lot) longer!'*

Student feedback from 2018, overview

- Overall assessment

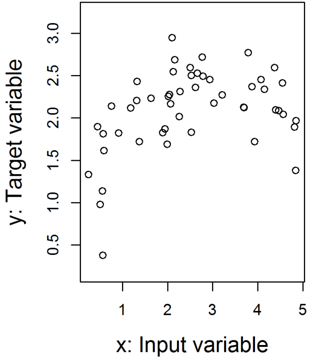


- I will benefit from things learnt on the course



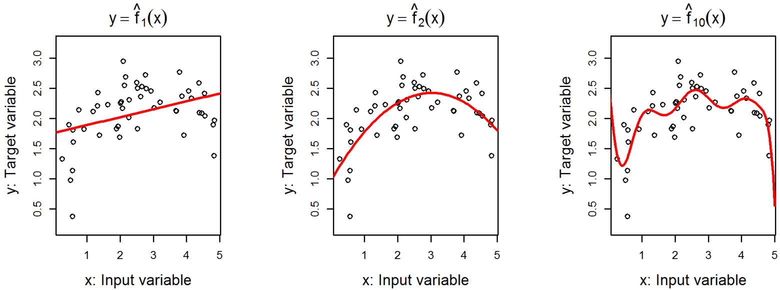
Probabilistic modeling overview (1/2)

- 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
  - ...



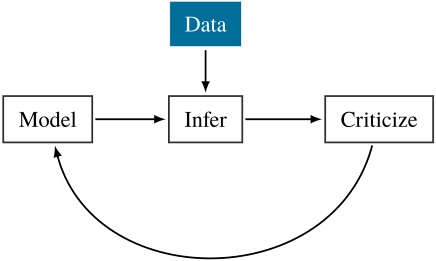
Probabilistic modeling overview (2/2)

- 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: $f_n(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n$



Course contents (1/2)

- 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)

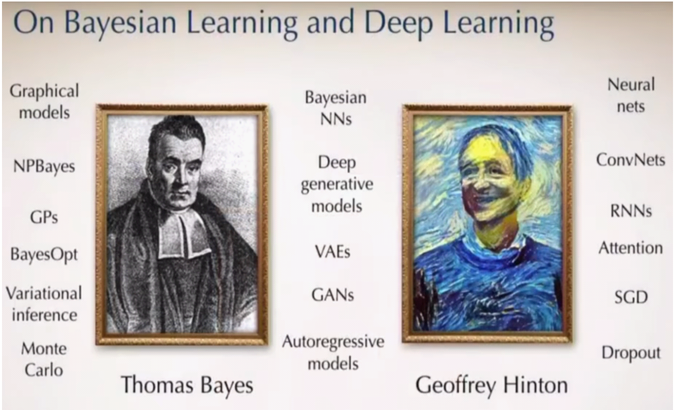
Course contents (2/2)

- A brief introduction to probabilistic programming using Tensorflow and Edward[2].

[2] Tensorflow Probability is another very recent tool that incorporates also the functionality of Edward, see: https://www.tensorflow.org/probability/.

Role of probabilistic machine learning today

- Keynote at NIPS in December 2017 by Yee Whye Teh



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

Basic probability calculus

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

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

Notation (1/2)

- Random variables: $X, Y, Z, \ldots$
- Values these random variables can take: $x, y, z, \ldots$
- 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
  - $dom(X)$ denotes all possible states for variable X.
- Distribution of a variable X consists of
  - its domain $dom(X)$
  - and full specification of probability values $p_X(x)$, for all possible $x \in dom(X)$
- Normalization
  - The summation over all the states

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

  - 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

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

- 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,

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

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: $\mathcal{A} = \{A_1, \ldots, A_k\}$ and $\mathcal{B} = \{B_1, \ldots, B_l\}$ are independent if

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

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,

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

- Similarly,

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

and

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

Example - Conditional independence (3/3)

B	M	K	$p(b,m \vert k)$	$p(b \vert k)$	$p(m \vert k)$	$p(b \vert k)p(m \vert k)$
1	1	1	**0.029**	**0.711**	**0.051**	**0.036**
0	1	1	$\dots$	$\dots$	$\dots$	$\dots$
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 $X_1, X_2, \ldots, X_n$ denote the cumulative sum of n dice throws, such that $dom(X_1) = \{1, \ldots, 6\}$, $dom(X_2) = \{2, \ldots, 12\}$, etc.
  - Is X_{n+1} independent of X_{n-1}?
  - Is X_{n+1} conditionally independent of X_{n-1} given X_n?
- X='Location of an airplane now', Y='Location of the plane 15s ago', Z='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)

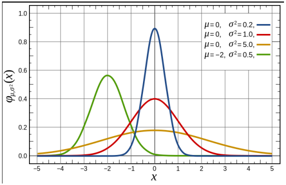
- Probability density function (pdf) for a continuous variable X, $f_X()$

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

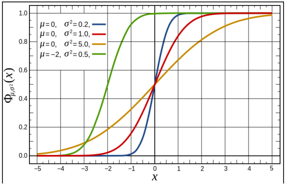
$$p(X \in [a, b]) = \int_{x=a}^{b} f_X(x) dx$$

- 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)

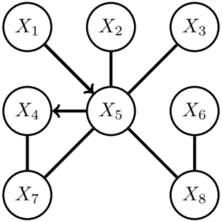
- Conditional distribution

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

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

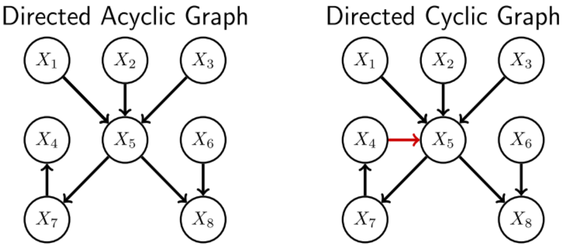
$$f_{X,Y|Z}(x,y|z) = f_{X|Z}(x|z)f_{Y|Z}(y|z)$$

Basic graph definitions



- A **graph** consists of **nodes** (vertices) and undirected of directed **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



- 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