Managing uncertainty

Part I: frederic.pennerath@supelec.fr

Part II: matthieu.geist@supelec.fr

What is this course about?

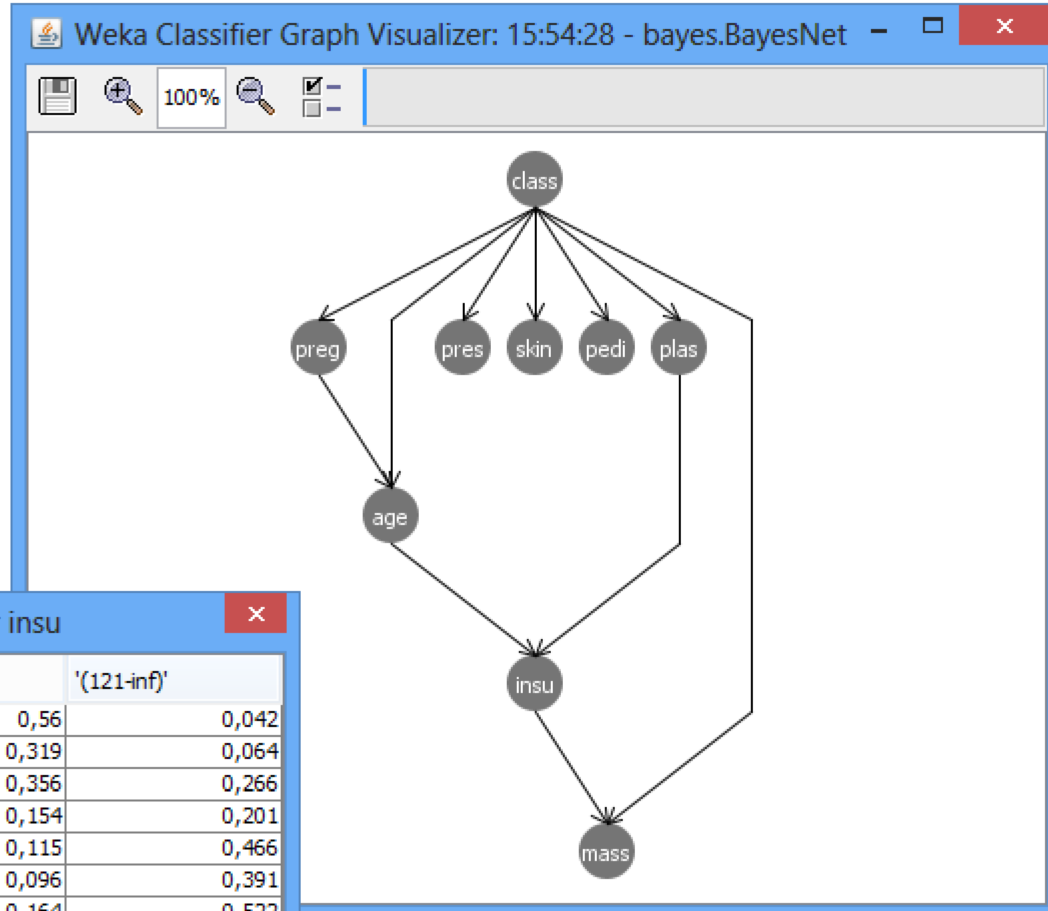
- A common issue:
  Making (good) **decision** when problem inputs are **uncertain**

- A common powerful paradigm based on **probability theory**:
  Bayesian Inference

- A theoretical course with a toolbox of reference methods
  - Graphical Models
  - Gaussian Processes
  - Bayesian Filtering (e.g. Kalman filter)
  - Hidden Markov Models (HMM)
  - MDP/POMDP,
  - Reinforcement learning …

- Underlying some of the most modern applications

Example of application: Data Analysis and Decision Theory



Probability Distribution Table For insu

plas	age	'(-inf-14.5]'	'(14.5-121]'	'(121-inf)'
'(-inf-99.5]'	'(-inf-28.5]'	0,398	0,56	0,042
'(-inf-99.5]'	'(28.5-inf)'	0,617	0,319	0,064
'(99.5-127.5]'	'(-inf-28.5]'	0,377	0,356	0,266
'(99.5-127.5]'	'(28.5-inf)'	0,645	0,154	0,201
'(127.5-154.5]'	'(-inf-28.5]'	0,42	0,115	0,466
'(127.5-154.5]'	'(28.5-inf)'	0,513	0,096	0,391
'(154.5-inf)'	'(-inf-28.5]'	0,313	0,164	0,522
'(154.5-inf)'	'(28.5-inf)'	0,508	0,038	0,454

Source: UCI Diabetes dataset

Example of application: Navigation Systems and Data Fusion

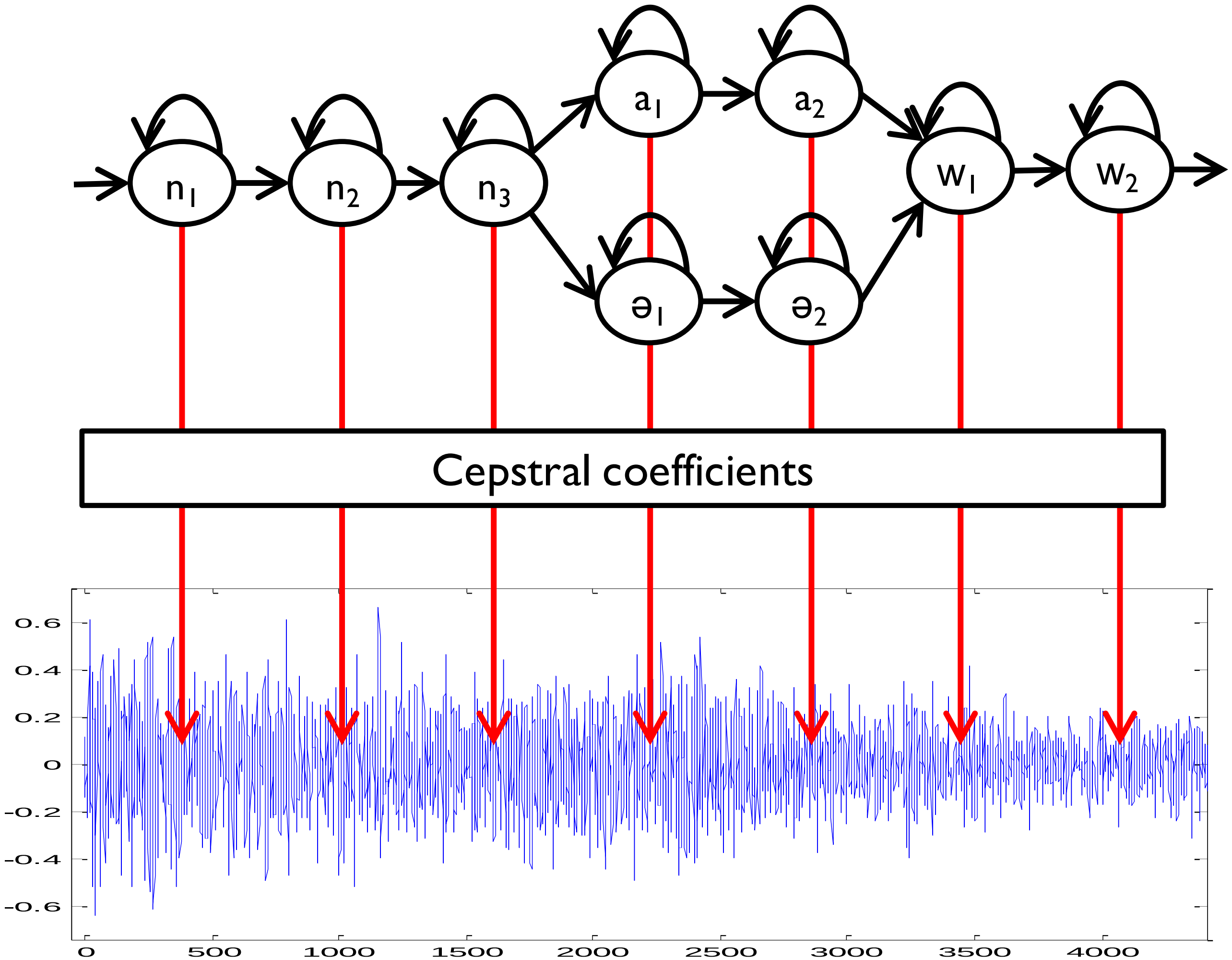
Navigation System:
Position (L, l, z)
Angles (yaw, pitch, roll)
Speed
Angular speed

Altimeter:
☺ No drift
☹ Altitude only
☹ Not accurate

Inertial System:
☺ Responsive
☹ Long term drift

GPS:
☺ No drift
😐 Medium accuracy
☹ Not responsive

Example of application: Speech recognition systems



Plan

Part I: Bayesian inference and graphical models

- Static models + bayesian filtering
- Given by myself
- Lessons from 1 to 4

Part II: Markov models and processes

- Dynamic models (Markov Chains, HMM, MDP, POMDP, etc)
- Given by Matthieu Geist
- Lessons from 5 to 8

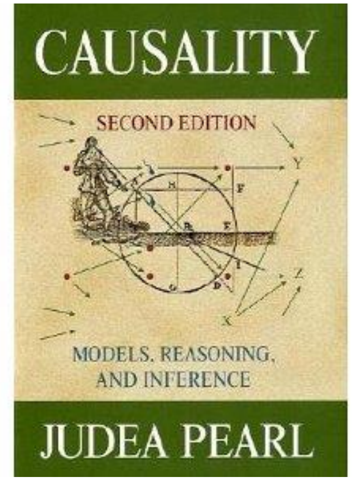
Detailed Plan of Part I

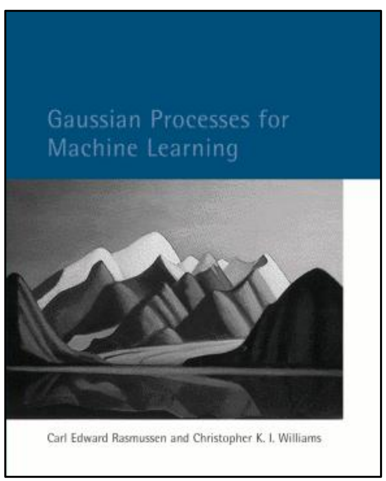
Part I: lessons 1 to 4 by myself

- Introduction: how to model uncertainty
  - Reminder of basic notions of probability theory
- Bayesian estimation: general principles
  - Bayes' rule
  - Bayes estimators
- Elementary bayesian methods
  - Classification: Naive Bayes
  - Regression: Linear models
  - Clustering: EM
- Gaussian Processes
- Bayesian Filtering and Kalman filters
- Graphical models:
  - Markov Random Fields
  - Bayesian Networks

References


BAYESIAN
REASONING
and
MACHINE
LEARNING
David Barber

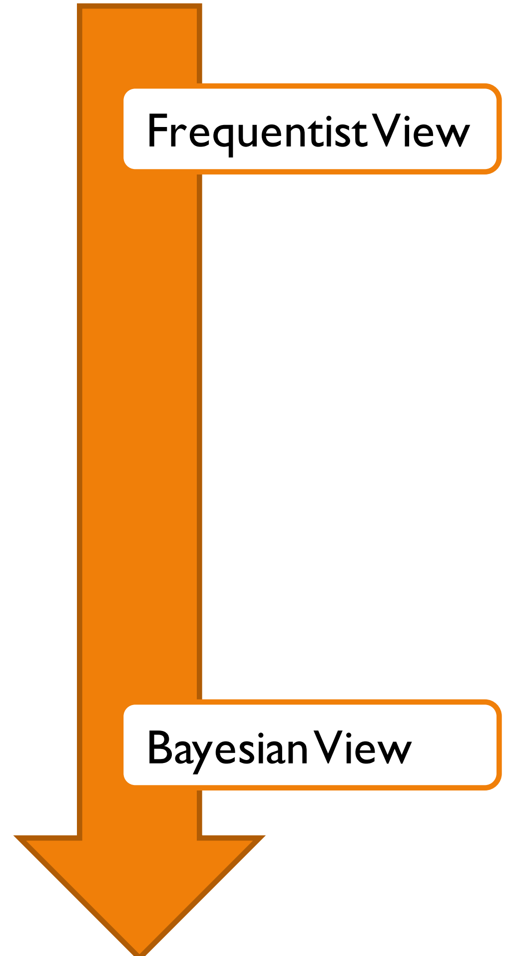

CAUSALITY
SECOND EDITION
MODELS, REASONING,
AND INFERENCE
JUDEA PEARL


Gaussian Processes for
Machine Learning
Carl Edward Rasmussen and Christopher K. I. Williams


Patrick Naïm, Pierre-Henri Wuillemin,
Philippe Leray, Olivier Pourret, Anna Becker
Réseaux
bayésiens
EYROLLES

Different types of uncertainty

- From hardly predictable future events ("randomness")
  - Chaotic events like throwing a dice
- From "myopic" views of reality because
  - Information is not easily observable.
    - E.g. kinetic energy of a given molecule in a gas?
  - Information is useless/too expensive to be stored
    - E.g. how many hair does one have on one's head?
    - E.g. data streams (logs etc)
- From lack of information/knowledge
  - Information: will I pass the exam?
  - Knowledge: HMM do not match brain perception of spoken languages



The Frequentist View

- Historical view of probabilities made by statisticians:

"How likely is an event to occur, given ***past observations of it****?"*

- Probability interpretation: *probabilities are limits of* ***frequencies***
- Fundamental principle: *law of large numbers* $\lim_{n \to +\infty} \frac{\sum_{i=1}^{n} X_i}{n} = E(X)$
- Restrictions: assume
  - A large number of observations are available
    - Either available from a large reservoir (population, etc)
    - Or outputs of repeatable experiences (throwing dice)
  - Stationarity
- At the origin of many useful notions:
  - Expected values (as a limit of average)
  - Independence and sampling (e.g. polls in a population)
  - Confidence intervals. Convergences and concentration inequalities

The Bayesian View

- Modern view of probabilities used in machine learning:

*"How likely is an event to occur, given **what I believe to know**?"*

- Probability interpretation:

*Probabilities are the amount of **confidence** that I grant to some events to occur, given what I know.*

- Fundamental principle: *Bayes' rule and inference*
- Advantages:
  - Encompasses frequentist interpretation of probability
  - Does **not** assumes events are observable or stationary:

    E.g. "Will I pass my exam?"
- Limitations:
  - Too ambitious to be scalable: every variable or parameter has to be described by its distribution

Before going further: Probability Reminder & Notation

Probability space: $(\Omega, \mathcal{E}, P)$

- A set Ω of possible **outcomes**
- A set $\mathcal{E}$ of events defined as **subsets** of outcomes closed under
  - Conjunction (and): $E_1 \cap E_2$
  - Disjunction (or): $E_1 \cup E_2$
  - Negation (not): $\overline{E} = \Omega \setminus E$
- A function $P: \mathcal{E} \to [0,1]$ mapping events to probabilities s.t.
  - $P(\Omega) = 1$
  - $P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$

Random variable:

- A function $X: \Omega \to D_X$ mapping every outcome to a value in D_X
- Such that the fact X takes its value into some "reasonable" subset $\mathrm{V} \in \Sigma$ is mapped to some probability:

$$\forall \mathrm{V} \in \Sigma, X^{-1}(V) \in \mathcal{E}$$

(where $\Sigma \in 2^{D_X}$ is closed under $\cap, \cup, \setminus$)

- **Distribution** of X: $\mathrm{P_X(x)} = P(X = x) = P(X^{-1}(\{x\}))$

Joint distribution and the curse of dimensionality

Joint distribution :

- Given a set of random variables $X_1, \dots, X_n$,
- **Joint variable** is $(X_1, \dots, X_n): \Omega \to D_{X_1} \times \cdots \times D_{X_n}$ whose
- **Joint distribution** is:

$$\mathrm{P}_{(X_1,\dots,X_n)}(\mathrm{x}_1, \dots, \mathrm{x}_\mathrm{n}) = P(X_1 = x_1 \cap \cdots \cap X_n = x_n)$$
$$= P(X_1^{-1}(\{x_1\}) \cap \cdots \cap X_n^{-1}(\{x_n\}))$$

Curse of dimentionality:

- Joint distribution contains all information we need but …
- But if $X_1, \dots, X_n$ can take each m values, need a table of size m^n

→ Probabilistic models do not scale

→ Unless further hypothesis (independence, Markov property, etc)

Probability Theory: the two main operations to know

- **Marginalization** ≡ reducing joint distribution to a subset of variables
- Information loss
- Obtained by the **sum rule**:

$$P_{V_1,\dots,V_n}(v_1,\dots,v_n) = \sum_{h_1,\dots,h_m} P_{V_1,\dots,V_n,H_1,\dots,H_m}(v_1,\dots,v_n,h_1,\dots,h_m)$$

- **Conditioning** ≡ restricting joint distribution by a subset of values
- Information gain
- Obtained by the **product rule**:

$$P_{V_1,\dots,V_n}(v_1,\dots,v_n|K_1 = k_1,\dots,K_m = k_m) = \frac{P_{V_1,\dots,V_n,K_1,\dots,K_m}(v_1,\dots,v_n,k_1,\dots,k_m)}{P_{K_1,\dots,K_m}(k_1,\dots,k_m)}$$

- Requires marginalization

Probability Independence

Definition:

- Two events A and B are **independent** iff:

$$P(A \cap B) = P(A) \times P(B) \text{ or equiv. } P(A|B) = P(A)$$

- Extension to random variables:

$$\forall x, \forall y, P(X = x \cap Y = y) = P(X = x) \times P(Y = y)$$

- Extension to a set of events/variables $(A_i)_{1 \leq i \leq n}$:

$$P(\bigcap_{1 \leq i \leq n} A_i) = \prod_{1 \leq i \leq n} P(A) \text{ or equiv. } \forall I, P\left(\bigcap_I A_i \,\middle|\, \bigcap_{[1,n] \setminus I} A_i\right) = P(\bigcap_I A_i)$$

Examples:

$$P(dice\ 1\ \&\ 2\ show\ 1) = P(dice\ 1\ shows\ 1) \times P(dice\ 2\ shows\ 1)$$
$$P(Hurricane\ x | Butterfly\ y\ flaps\ its\ wings) = P(Hurricane\ x)$$

Remarks:

- Independence is very common (to a first approximation)
- Independence is scalable (factorizes joint distribution).