CentraleSupélec

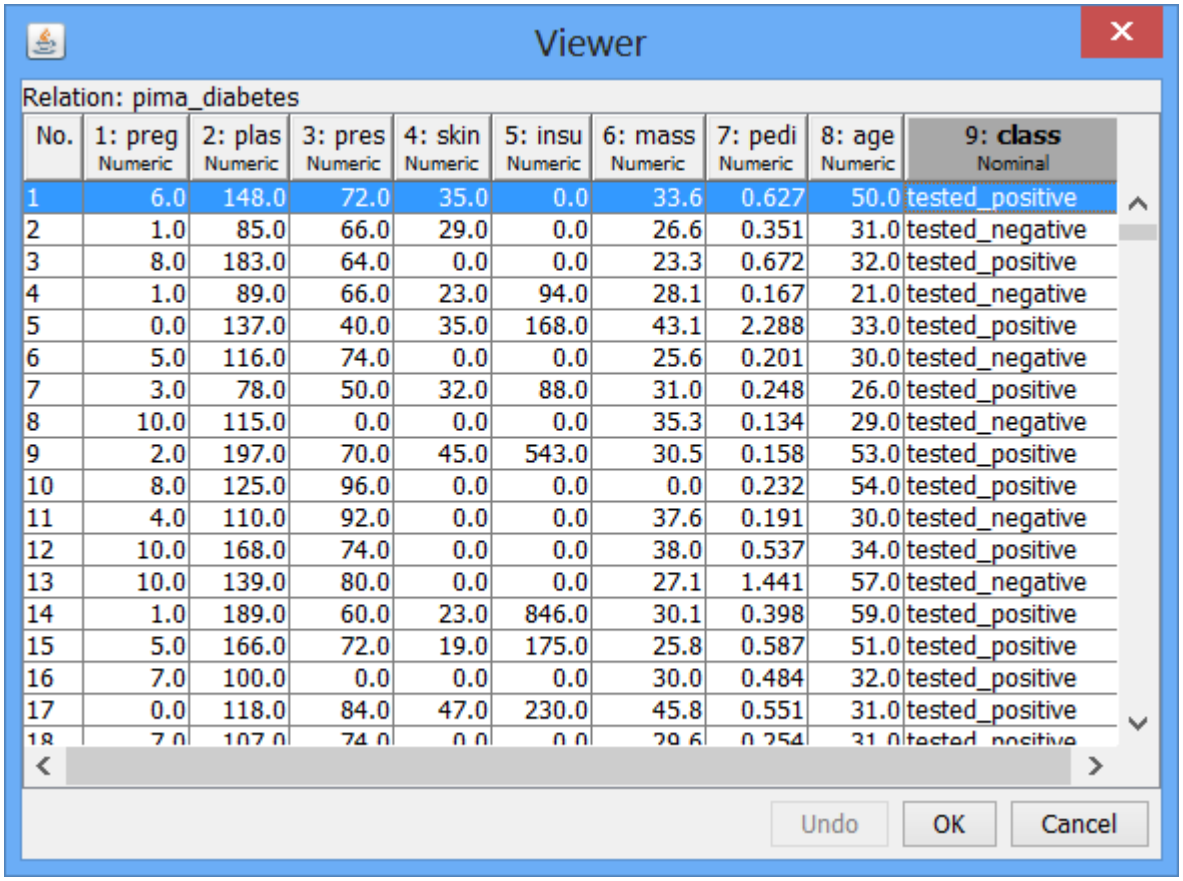
BAYESIAN MACHINE LEARNING

frederic.pennerath@centralesupelec.fr

What is this "Bayesian Machine Learning" course about?

- A course emphasizing the few essential theoretical ingredients
  - Probabilistic generative models and Bayesian inference
  - MLE & MAP estimators
  - Graphical Models (Bayesian Network, Markov Random Fields)
  - Latent variables & EM
  - Sampling Techniques
- An introduction to some reference methods
  - Basic Bayesian classification (Naïve Bayes) and clustering (GMM)
  - Markov Chains and Hidden Markov Models
  - State space models (Kalman filter, particle filters)
  - Gaussian Processes
  - Topic Model extraction (LDA)
  + Reinterpretation/extension of classical methods (linear & logistic regression, etc)
- Underlying some of the most modern applications

Example of application: Data Analysis and Decision Theory



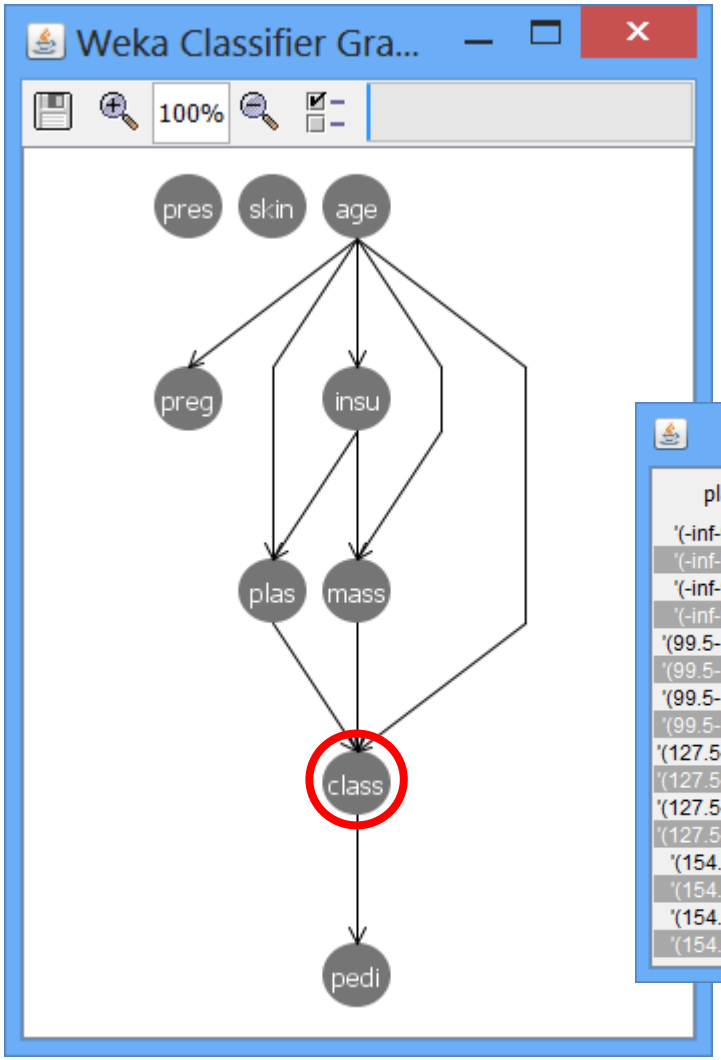
No.	1: preg Numeric	2: plas Numeric	3: pres Numeric	4: skin Numeric	5: insu Numeric	6: mass Numeric	7: pedi Numeric	8: age Numeric	9: class Nominal
1	6.0	148.0	72.0	35.0	0.0	33.6	0.627	50.0	tested_positive
2	1.0	85.0	66.0	29.0	0.0	26.6	0.351	31.0	tested_negative
3	8.0	183.0	64.0	0.0	0.0	23.3	0.672	32.0	tested_positive
4	1.0	89.0	66.0	23.0	94.0	28.1	0.167	21.0	tested_negative
5	0.0	137.0	40.0	35.0	168.0	43.1	2.288	33.0	tested_positive
6	5.0	116.0	74.0	0.0	0.0	25.6	0.201	30.0	tested_negative
7	3.0	78.0	50.0	32.0	88.0	31.0	0.248	26.0	tested_positive
8	10.0	115.0	0.0	0.0	0.0	35.3	0.134	29.0	tested_negative
9	2.0	197.0	70.0	45.0	543.0	30.5	0.158	53.0	tested_positive
10	8.0	125.0	96.0	0.0	0.0	0.0	0.232	54.0	tested_positive
11	4.0	110.0	92.0	0.0	0.0	37.6	0.191	30.0	tested_negative
12	10.0	168.0	74.0	0.0	0.0	38.0	0.537	34.0	tested_positive
13	10.0	139.0	80.0	0.0	0.0	27.1	1.441	57.0	tested_negative
14	1.0	189.0	60.0	23.0	846.0	30.1	0.398	59.0	tested_positive
15	5.0	166.0	72.0	19.0	175.0	25.8	0.587	51.0	tested_positive
16	7.0	100.0	0.0	0.0	0.0	30.0	0.484	32.0	tested_positive
17	0.0	118.0	84.0	47.0	230.0	45.8	0.551	31.0	tested_positive
18	7.0	107.0	74.0	0.0	0.0	29.6	0.254	31.0	tested_positive

Undo OK Cancel

preg	*Number of times pregnant*
plas	*Plasma glucose concentration after 2 hours in an oral glucose tolerance test*
pres	*Diastolic blood pressure*
skin	*Triceps skin fold thickness*
insu	*2-hour serum insulin*
mass	*Body mass index*
pedi	*Diabetes pedigree function*
age	
class	*Diabetic or not*

Source: UCI Pima Indian Diabetes dataset

Example of application of Graphical Models: Data Analysis and Decision Theory

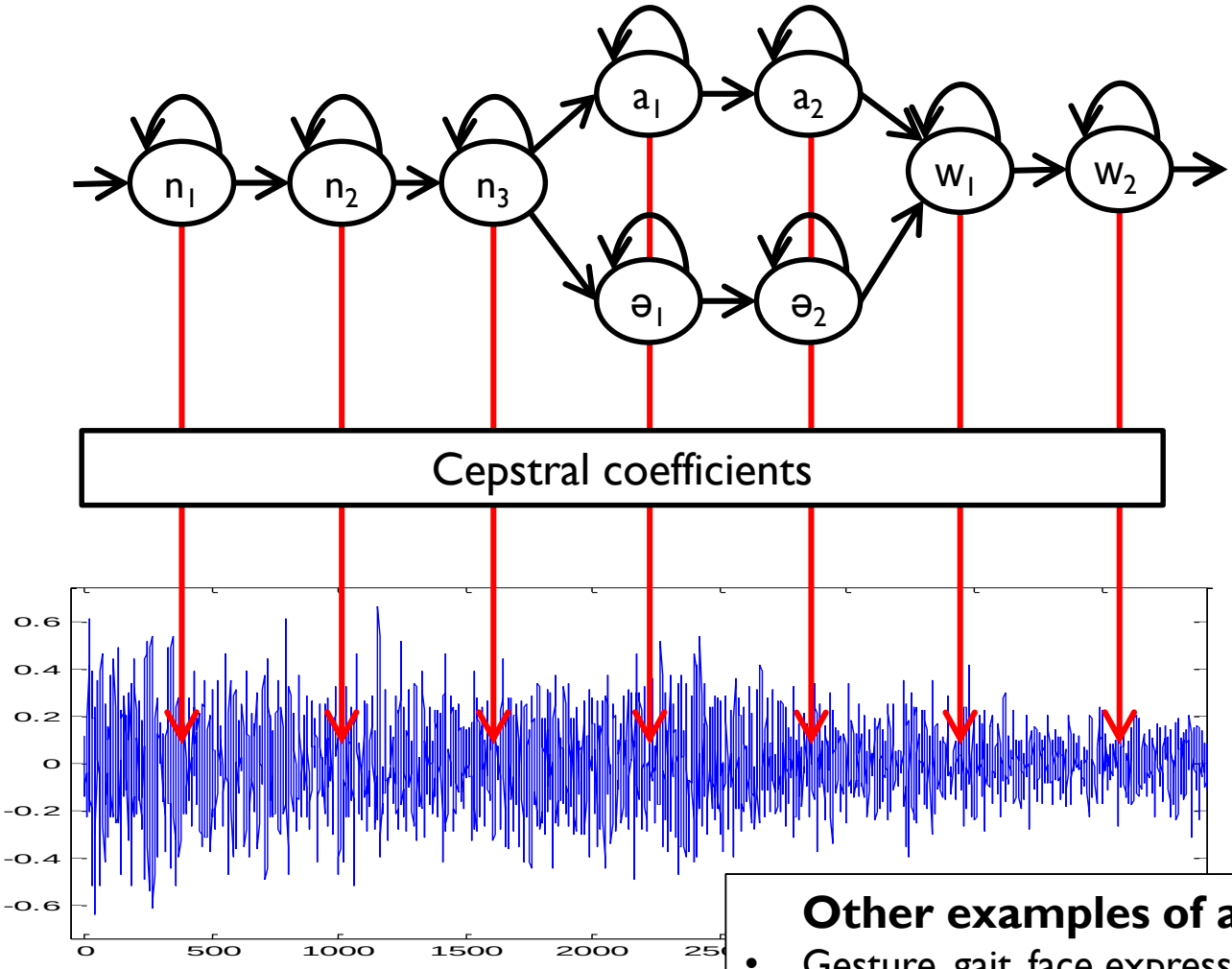


Probability Distribution Table For class

plas	mass	age	tested_negative	tested_positive
'(-inf-99.5]'	'(-inf-27.85]'	'(-inf-28.5]'	0,992	0,008
'(-inf-99.5]'	'(-inf-27.85]'	'(28.5-inf)'	0,921	0,079
'(-inf-99.5]'	'(27.85-inf)'	'(-inf-28.5]'	0,904	0,096
'(-inf-99.5]'	'(27.85-inf)'	'(28.5-inf)'	0,817	0,183
'(99.5-127.5]'	'(-inf-27.85]'	'(-inf-28.5]'	0,971	0,029
'(99.5-127.5]'	'(-inf-27.85]'	'(28.5-inf)'	0,75	0,25
'(99.5-127.5]'	'(27.85-inf)'	'(-inf-28.5]'	0,824	0,176
'(99.5-127.5]'	'(27.85-inf)'	'(28.5-inf)'	0,51	0,49
'(127.5-154.5]'	'(-inf-27.85]'	'(-inf-28.5]'	0,794	0,206
'(127.5-154.5]'	'(-inf-27.85]'	'(28.5-inf)'	0,868	0,132
'(127.5-154.5]'	'(27.85-inf)'	'(-inf-28.5]'	0,582	0,418
'(127.5-154.5]'	'(27.85-inf)'	'(28.5-inf)'	0,356	0,644
'(154.5-inf)'	'(-inf-27.85]'	'(-inf-28.5]'	0,7	0,3
'(154.5-inf)'	'(-inf-27.85]'	'(28.5-inf)'	0,367	0,633
'(154.5-inf)'	'(27.85-inf)'	'(-inf-28.5]'	0,155	0,845
'(154.5-inf)'	'(27.85-inf)'	'(28.5-inf)'	0,162	0,838

Source: UCI Pima Indian Diabetes dataset

Example of application of Hidden Markov Models: Speech recognition systems



Other examples of application:

- Gesture, gait, face expression tracking, etc
- Handwritten text recognition
- Genomics

Example of application of Kalman Filters: Navigation/Tracking Systems and Data Fusion

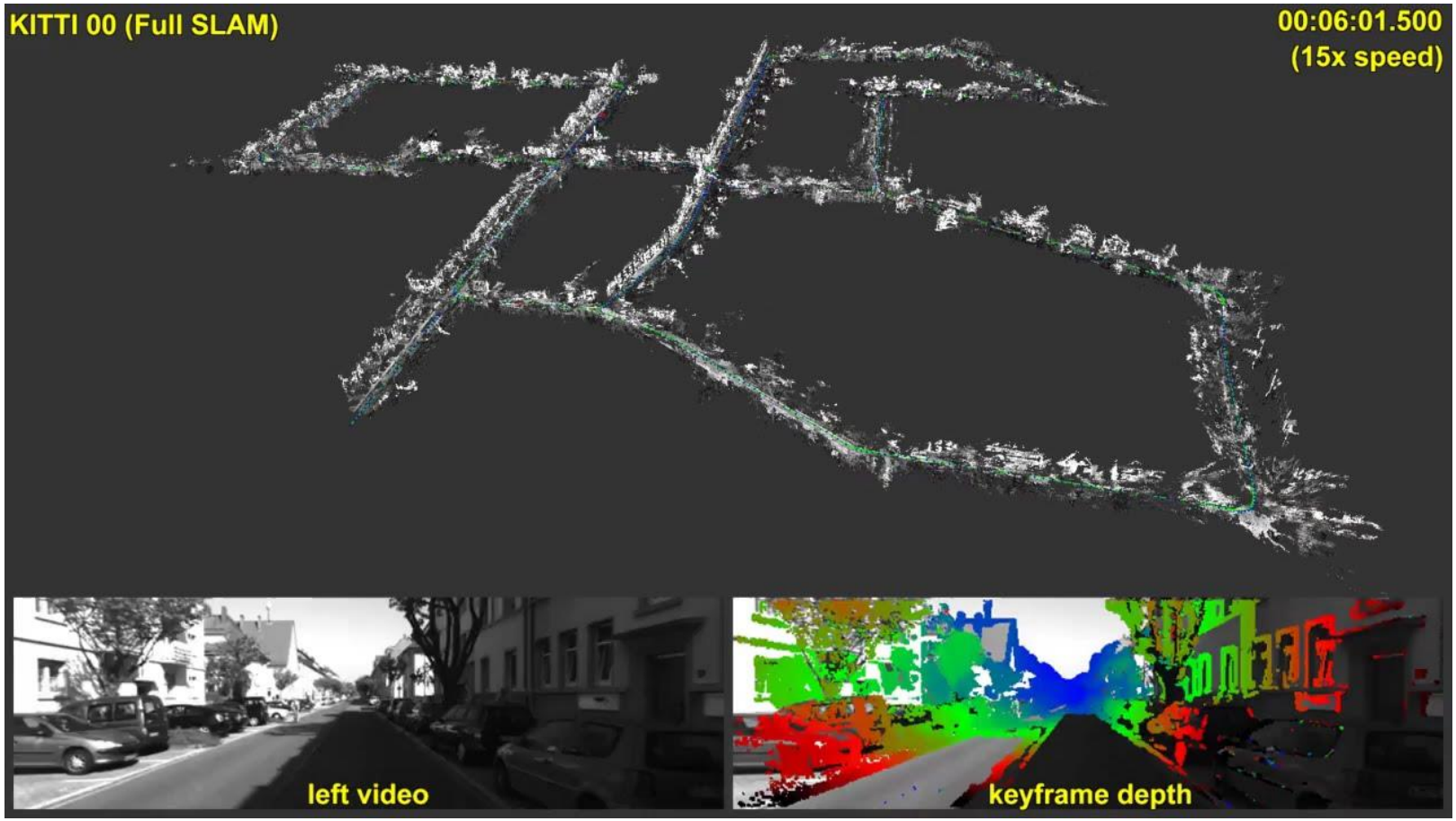
Navigation System:
Position (L, l, z)
Angles (yaw, pitch, roll)
Speed
Angular speed

Altimeter:
☺ No drift
☹ Altitude only
☹ Not accurate

GPS:
☺ No drift
😐 Medium accuracy
☹ Not responsive

Inertial System:
☺ Responsive
☹ Long term drift

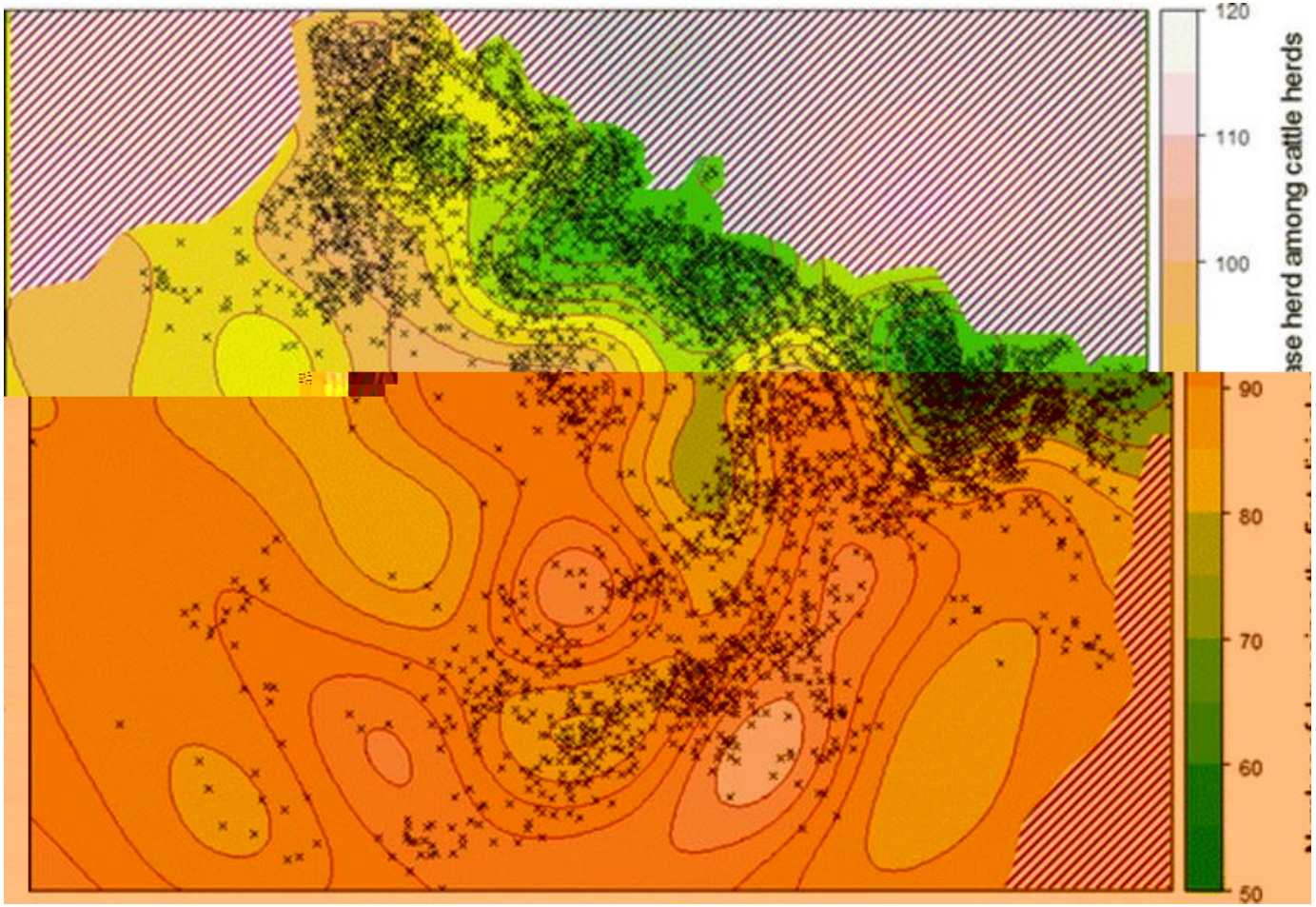
Example of application of Extended Kalman Filters and Particle Filters: Simultaneous Localization and Mapping (SLAM)



Video https://www.youtube.com/watch?v=oJt3Ln8H03s

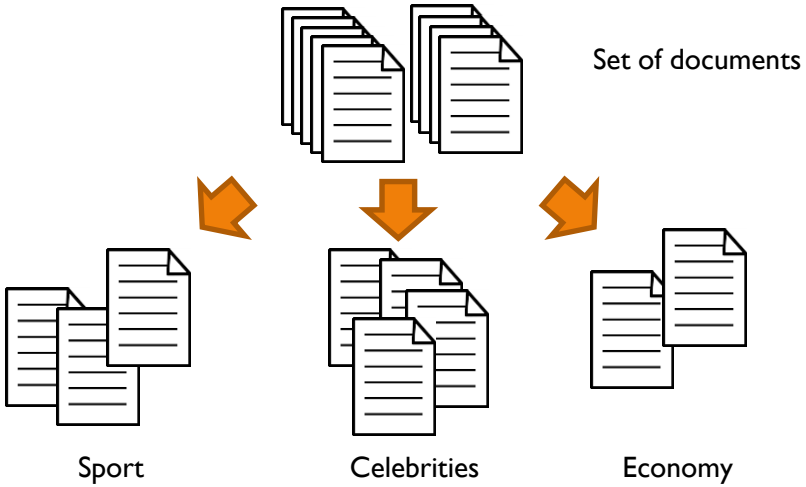
Source: Engel, 2015

Example of application of Gaussian Processes: Map Interpolation (Kridging)



Source: Nusinovici & al, 2016

Example of application of Bayesian Clustering: Topic model and LDA



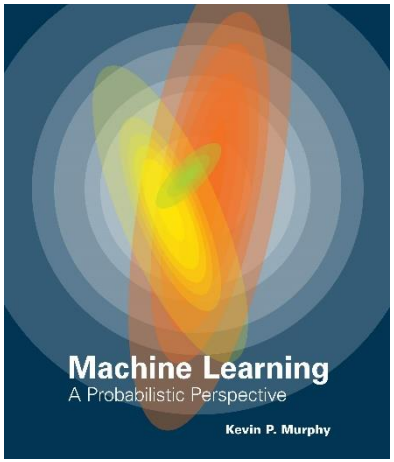
Topic #0:
government people mr law gun state president states public use right rights national new control american security encryption health united
Topic #1:
drive card disk bit scsi use mac memory thanks pc does video hard speed apple problem used data monitor software
Topic #2:
said people armenian armenians turkish did saw went came women killed children turkey told dead didn left started greek war
Topic #3:
year good just time game car team years like think don got new play games ago did season better ll
Topic #4:
10 00 15 25 12 11 20 14 17 16 db 13 18 24 30 19 27 50 21 40
…

Source: 20newsgroups dataset, Scikit-learn

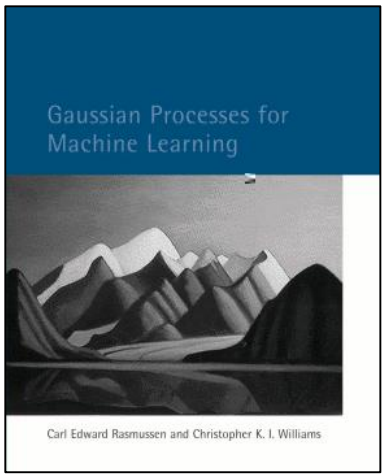
Syllabus

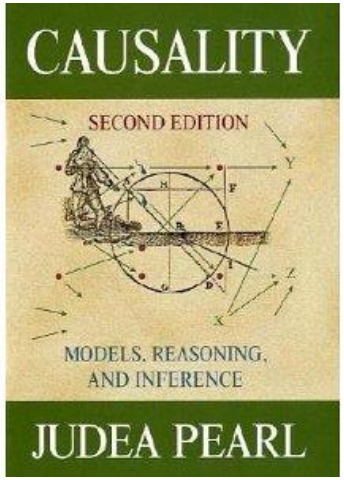
1. Bayesian estimation: general principles
   - Bayes' rule
   - Bayes estimators
   - Introduction to Bayesian Network
2. Elementary Bayesian methods
   - Classification: Naive Bayes
   - Regression: Bayesian linear regression
3. Models with latent variables
   - EM algorithm
   - Mixture models and GMM
4. Non parametric method:
   - Gaussian Processes
5. Markov models:
   - Markov chains
   - Hidden Markov models
   - Kalman filter
6. Approximate estimation and Sampling
   - Gibbs Sampling, MCMC
   - Importance Sampling, particle filtering

References


Machine Learning
A Probabilistic Perspective
Kevin P. Murphy


tracking
uncertainty
time series
inference
data mining
statistics
data
decision
BAYESIAN
finance
kernels
clustering


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


CAUSALITY
SECOND EDITION
MODELS, REASONING,
AND INFERENCE
JUDEA PEARL


Algorithmes
Patrick Naïm, Pierre-Henri Wuillemin,
Philippe Leray, Olivier Pourret, Anna Becker
Réseaux
bayésiens
• Introduction intuitive aux réseaux bayésiens
• Fondements théoriques et algorithmes
• Méthodologie de mise en œuvre
• Domaines d'application et études de cas détaillées
• Outils logiciels : Bayes Net Toolbox,
BayesiaLab, Hugin et Netica
EYROLLES

What types of uncertainty can be modeled by probabilities?

- From hardly predictable **repeatable** events (randomness)
  - Chaotic events like throwing a dice
- To "myopic" views of collections of samples (statistics)
  - Samples are not observable

    E.g. statistical physics: kinetic energy of a given molecule in a gas?
  - Samples are useless/too expensive to be stored

    E.g. data streams (logs etc)
- To states of belief (no samples)
  - Single events like the probability I will pass the exam
  - Distribution of a model parameter given past observations

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 $V \in \Sigma$ is mapped to an event (and thus a probability):

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

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

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

Joint distribution and the curse of dimensionality

Joint distribution :

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

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

Curse of dimensionality:

- 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

→ Needs huge number of samples to avoid overfitting

→ 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**:

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

- 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).