Team meeting

Pierre Dangauthier

24 avril 2006

Outline

- Machine learning
- Philosophy of probability
- Bayesian Learning
- 3 debates in machine learning
- Entropy maximization difficulties
- Pictures of my last holidays

Machine learning

- Goal
  - machine or agents that learn from experience,
  - that performs better days after days.

- Necessity of
  - Memorization: summing up past **relevant** information
  - Generalization: **over fitting** is a big issue
  - Prediction (what decision to take in front of a new situation)
  - A model of the phenomena
  - A fitness function (for evaluation)

- Probabilities: 2 purposes:
  - Obvious: Analyze performances of learning algorithms,
    - ex: how often they takes good decisions on a benchmark set
  - One approach probability rules = the way the agent represents its knowledge
    - ”probability as logic”

- Those two applications are compatible, and both imply that probability is an important notion for machine learning.

Philosophy of probability

- What is a probability ? Definition ?
  - Kolmogorov axiomatic is (almost) uncontroversial.

- What meaning, what interpretation ?
  - Different philosophical schools
  - Practical methods depend on interpretation
  - Numerical results are different
  - Open debates

Philosophy of probability

- **Kolmogorov** axiomatic: measure theory
- P is a probability if
  - P is a function from a convenient set of events
    - (sigma algebra)
  - to [0,1]
  - Sum to one
  - Countable additivity

$$\sum_{Events} P(event_i)=1$$

$$A_i)=\sum_{j=1}^{\infty} P(A_i)\ \ if\ \ A_j intersect A_i=0 ¿$$

- That’s all, no interpretation implied !
  - Random variables = measurable function
    - From probabilized space to measurable space
    - P(a<X<b) = P(X^-1 ([a,b]))
  - No definition of “randomness”, just measure theory.
  - Random variables are useful (also in bayesianism)

Philosophy of probability

- Different interpretations
  - **Classical probabilities**
    - based on the principle of indifference.
  - **Logical probabilities**
    - notion of “degree of implication” in a formal language.
  - **Frequency interpretation**
    - limiting relative frequencies in an hypothetical infinite sequence of trials.
  - **Propensity**
    - quality of the physical word. It represents an intrinsic tendency to behave in a certain way.
  - **Subjective interpretation**
    - degree of belief of a suitable agent.

Bayesian learning

- **Rational belief updating**
  - Of a model in front of data.
- **Methodology**
  - 1: Make a model, formalize prior knowledge
    - Parametrical model ex: X <- N(m,s2)
    - Ex: Prior knowledge S2 = 3, m <- uniform[0,5.5]
  - 2: collect data
  - 3: Compute posterior proba. of parameters.
  - 4: Use posterior for:
    - Point estimation of parameters
    - Prediction
    - Model comparison
    - Decide with expected loss minimization

Bayesian learning

$$P(param|data) = \frac{P(data|param)P(param)}{P(data)}$$

$$P(d|\theta)P(|\theta)$$

$$Posterior \propto Likelihood * prior$$

- Prediction (marginalization over parameter space)

$$P(new\ x|data) = \int_{param} P(new\ x|param)P(param|data)$$

$$P(x|d,M) = \int_{\theta} P(x|\theta)P(\theta|d,M)d\theta$$

Bayesian learning: Definitions

- **Maximum Likelihood (ML) Learning**

$$\theta = ArgMax\ P(d|\theta)$$

- **Maximum a Posteriori (MAP) Learning:**

$$\theta = ArgMax\ P(\theta|d) = ArgMax\ P(d|\theta)P(\theta)$$

- **Bayesian Learning:** We keep all

$$P(\theta|d) \propto P(d|\theta)P(\theta)$$

Bayesian learning: Priors

- **Informative**
  - Translate expert knowledge in prior
- **“non informative”**
  - Respect invariance of model (Jeffrey)
  - Minimize effect of prior (Bernardo)
  - Maximize “uncertainty” entropy (Jaynes)
- **Hierarchical priors, empirical**
  - Hyper parameters estimated on data
- **Conjugate priors**
  - Mathematical convenience, usable if flexible enough
  - Virtual data

Bayesian learning: inference

- **Exact**
  - tractable integration on discrete variable
  - Analytical (easy if conjugate prior)
- **Approximate**
  - Sampling (all MCMC stuff)
  - Variational approaches
    - Message passing (loopy belief propagation, expectation propagation, mean field) [belief propagation is exact on trees) cf Statistical physics
    - Min d(real posterior, a class of candidate distrib)
  - Posterior modes
    - Laplace, BIC
    - EM

Bayesian learning : Pros and cons

- Pros
  - Coherent, rational framework
  - Uncertainty
  - Pretty good philosophical foundations
  - Doesn’t over fit, if good prior
  - Automatic Occam razor
  - Generalization of (ML, Kalman, Markov, PCA, ICA, EM, neural nets)
  - Rational agent model : psychology and machine learning
- Cons
  - Subjectivity
  - Formulating a good prior is **THE** issue
  - Computationally demanding
  - Easy to do bad Bayesian :
    - No distribution over models
    - Improper Posteriors
    - Stupid priors ridicules, just convenient (conjugate prior)
    - Only MAP
    - Always uniform prior

3 debates in machine learning

- **Vapnik or Bayes?**
  - Vapnik (SVM): minimize future risk of error
  - Bayes : Probability = knowledge
- **Frequencies or degree of belief?**
  - Orthodoxy isn’t Bayesianism with flat prior !!
- **Subjective or objective Bayesianism?**
  - *Will two agents with the same information choose the same prior ?*
    - *Objectivist: **YES** (but which one ? Formal rule ? “know nothing ?)*
    - *Subjectivists: **NO**, matter of conventions, only “reference”, “defaults”* priors

Entropy maximization difficulties

- **Problems of MaxEnt**
  - Which entropy function ?
  - Sensitive to discretization
  - Invariance by reparametrization
  - Pb. with continuous variable
    - Choice of base measure reports the problem (Jaynes: invariance)
  - History: Shannon != inference
  - Uncertainty not well defined ! Uniform is information !
  - Conflict with Bayesian updating
    - Why not ? That’s different theories !

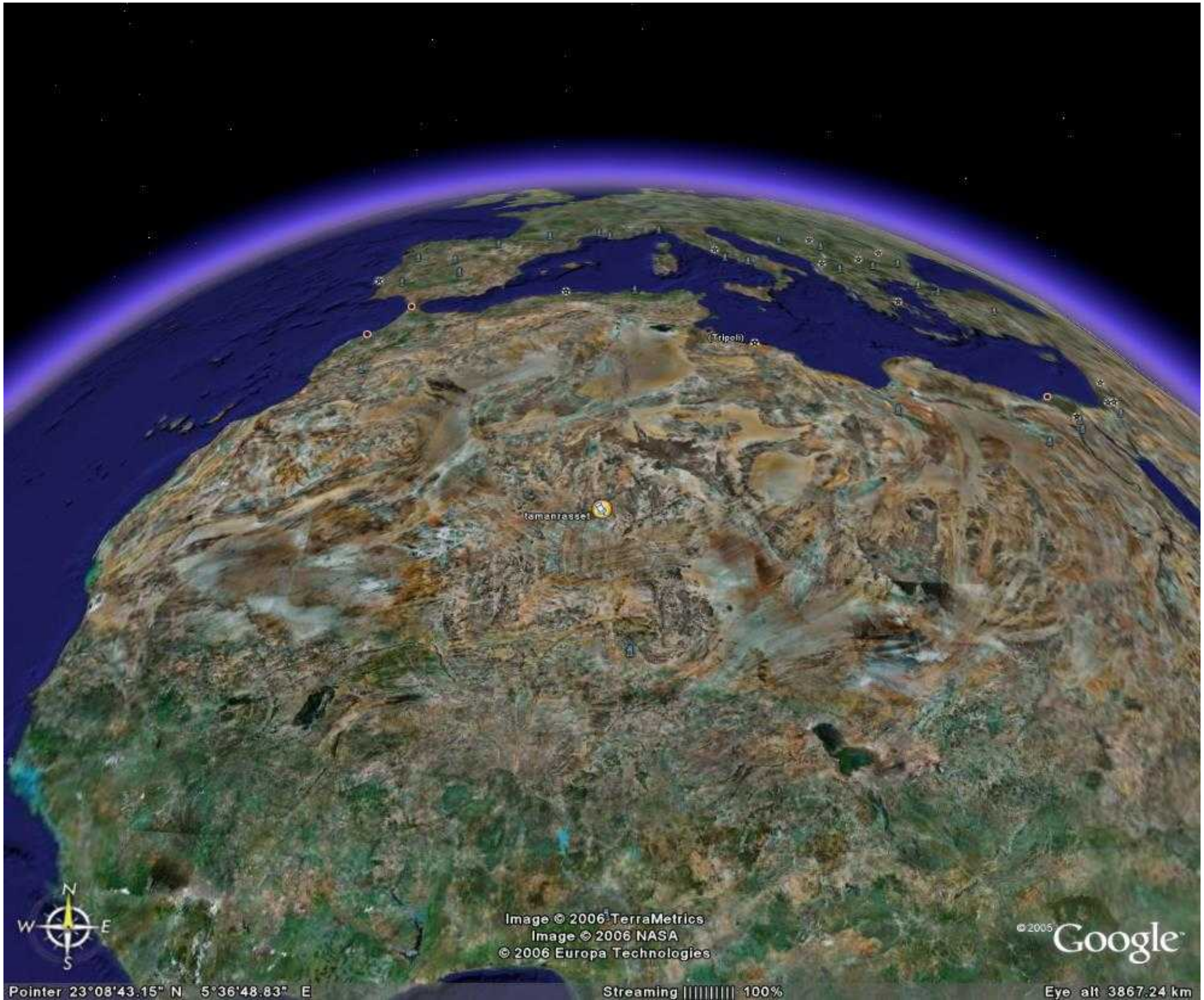
$$-\sum p_i \log p_i$$

$$-\int_x p(x) \log(p(x))\, d\mu$$

Entropy maximization difficulties

- Take a die X = 1,2,3,4,5,6
- Prior information: E(X)=3.5
  - MaxEnt -> p1=..=p6=1/6
- New information: “A”= only odd numbers
  - MaxEnt E(X)=3.5 and E(1A)=1
    - -> p1=0.22 p2=0 p3=0.32 p4=0 p5=0.47 p6=0
  - Bayesian updating
    - P(X|A)=0 for 2,4,6 and 1/3 for others
- Conflict ! Explanation: 2 different ways of belief updating, nothing say they must agree !

Pictures of my last holidays



References

Yet Another Machine Learning Blog

3 debates about machine learning

Ultimate machine learning algorithm ?

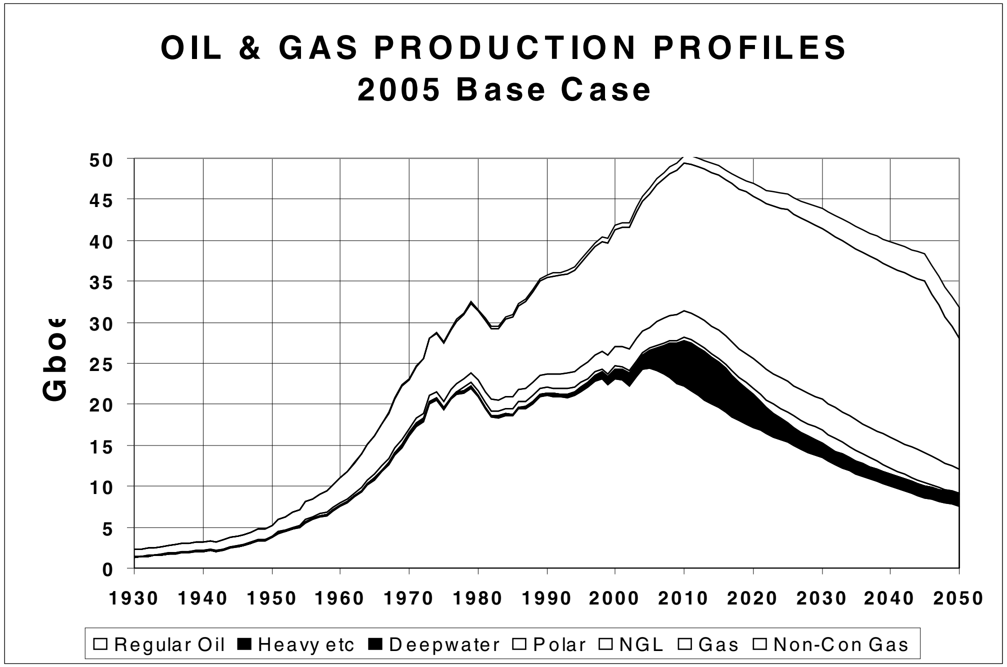
http://emotion.inrialpes.fr/~dangauthier/blog/

- Historical
  - Laplace / Jeffrey / Jaynes / Good
- Modern bayesians
  - MacKay / Neal / Ghahramani / Pearl / Minka
- More «mathematical» bayesians
  - Robert / Gelman / Bernardo
- Physicists
  - Baez / Jaynes / Jefferys/

The end

« Toute personne croyant qu'une croissance exponentielle peut durer indéfiniment dans un monde fini est soit un fou, soit un économiste. »

Kenneth Boulding



Comparaison de modèles

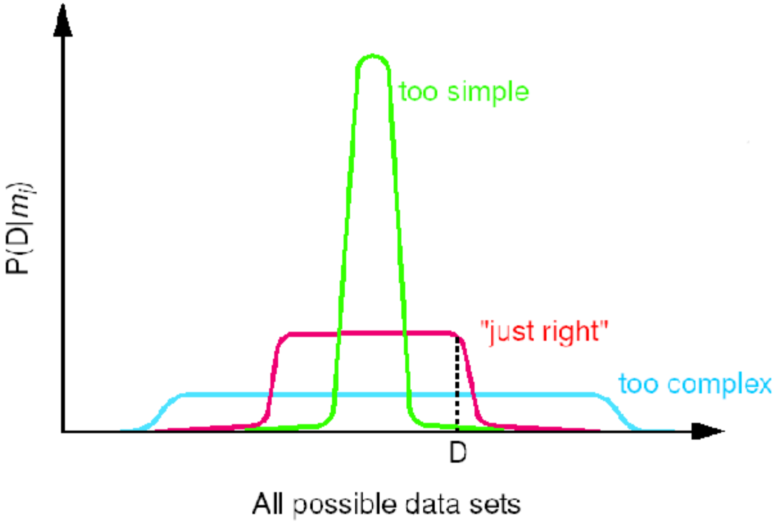
- Choix entre deux modèles

$$\frac{P(M_1|d)}{P(M_2|d)}=\frac{P(d|M_1)P(M_1)}{P(d|M_2)P(M_2)}=\frac{\int P(d|\theta,M_1)P(\theta|M_1)d\theta}{\int P(d|\theta,M_2)P(\theta|M_2)d\theta}\frac{P(M_1)}{P(M_2)}$$

- Rasoir d’Occam automatique

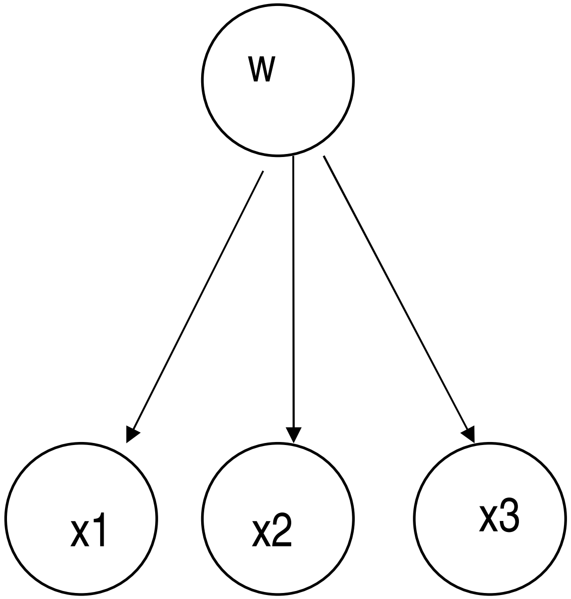
Une classe de modèles trop simple ou trop complexe donnera une faible probabilité au jeu de données

La classe de modèles la plus probable pour un jeu de données aura une complexité adaptée.

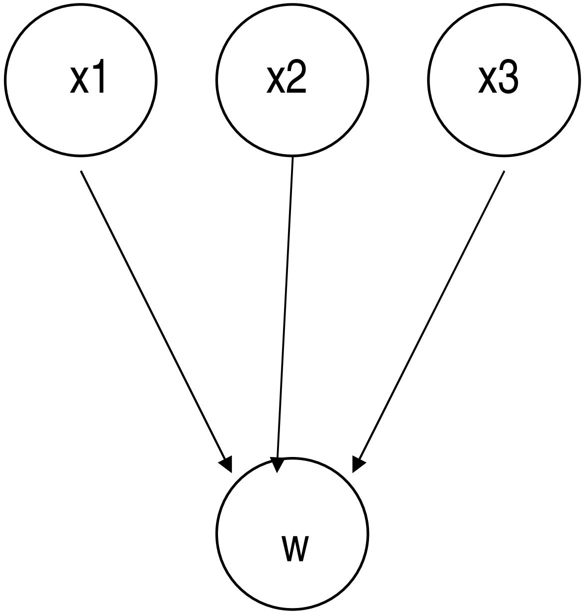


Modèle génératif / discriminatif

- Une classification binaire peut s’exprimer
  - Génératif - Discriminatif



$$P\left(x,w\right)=P\left(w\right)P\left(x\middle|w\right)$$



$$P\left(x,w\right)=P\left(x\right)P\left(w\middle|x\right)$$

Modèle génératif

$$P(\vec{x}|w_i)=\frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\vec{x}-\eta_i)^T\Sigma^{-1}(\vec{x}-\eta_i)\right)$$

$$P(w_0|\vec{x})=\frac{P(\vec{x}|w_0)P(w_0)}{P(\vec{x})}=\frac{P(\vec{x}|w_0)P(w_0)}{P(\vec{x}|w_0)P(w_0)+P(\vec{x}|w_1)P(w_1)}$$

$$\frac{1}{1+\exp\left(-\log\left[\frac{P(\vec{x}|w_0)}{P(\vec{x}|w_1)}\right]-\log\left[\frac{P(w_0)}{P(w_1)}\right]\right)}$$

$$\frac{1}{1+e^{-(w^T\vec{x}+b)}}$$