PML 1. Intro and Overview

Probabilistic Machine Learning Reading Group

David Rios Insua

November 5, 2025

Inst. C. Matemáticas (CSIC)

Motivation

- Aihub-CSIC connection interdisciplinary discussions
- Consulting for other CSIC intitutes
- Momentum+ALLIES programs @CSIC
- CSIC statistical/ML courses ‘kind of oldies’
- Internal needs within ICMAT and some of our sponsors
- Community building

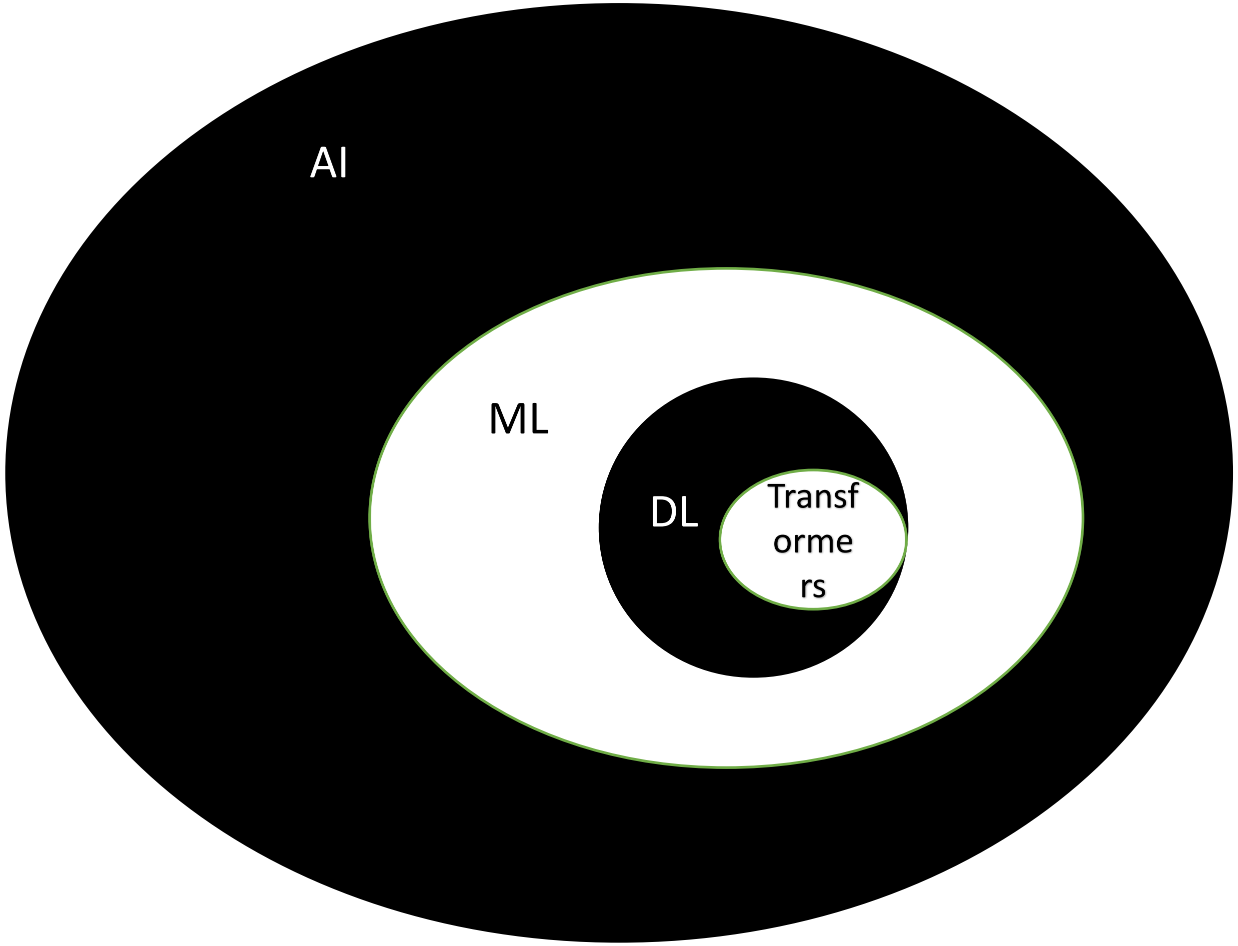
Contents

- Of Artificial and Natural Intelligence
- Basic PML concepts through an example
- Probabilistic Graphical Models
- Intro to ML with neural nets. Optimization with stochastic gradient descent
- Intro to PML with neural nets. MCMC and variational inference
- The road ahead

A brief distillation of KM 1-7

Of Artificial and Natural Intelligence

AI



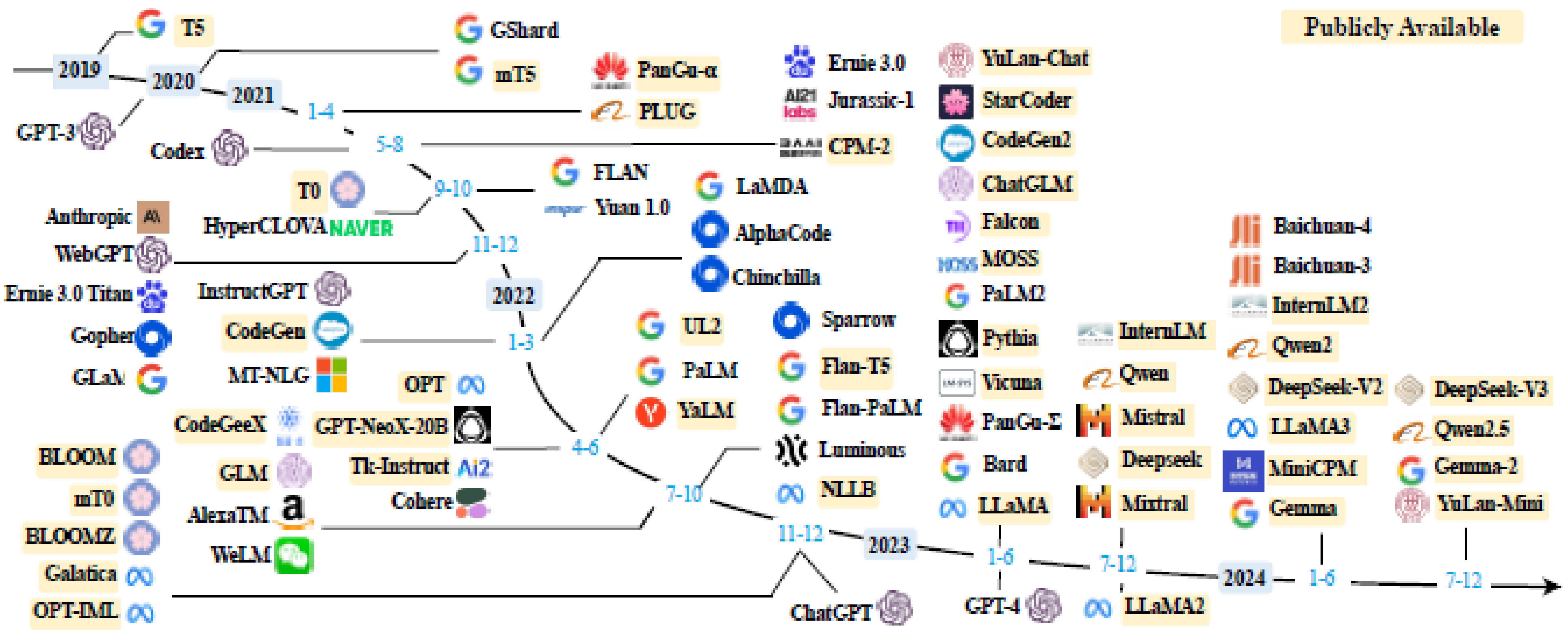
EU AI Act (Sep'23 version)

'**artificial intelligence system**' (AI system) means **software** that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, **generate outputs such as content, predictions, recommendations, or decisions** influencing the environments they interact with;

AnnexI: **ML…, logic…, statistics (bayes)…**

In final version: ML, logic+KB

The LLM rush



A survey of large language models https://arxiv.org/abs/2303.18223
Eight things to know about LLMs https://arxiv.org/abs/2304.00612
Attention is all you need (2017)
https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

ML

A computer program learns from experience E with respect to task T and performance measure P,

If its behaviour with respect to T, measured according to P, improves with experience E

Representation-Evaluation-Optimization

Goodfellow et al DL book

Some ML examples. Red matters!!!

Uncertainty is almost ubiquitous in ML:

- Given the monitoring trace of an Inet device, are we facing an attack? Should I stop operations?
- A person with these FB likes will buy this type of beer? Should I send her my brand add?
- If robot performs this, How will the user react? And the environment? Consequently, what should the robot do?

In many applications, we’ll need to go beyond

- Beyond a model with good fit…
- Beyond a model that predicts well…

In many applications, we'll need to go beyond

- Beyond a model with good fit…
- Beyond a model that predicts well…

- Fraud detection. Classification problem
  - Few false positives. FPR
  - Few false negatives. FNR

In applications, we'll need to go beyond

- Beyond a model with good fit…
- Beyond a model that predicts well…

- Fraud detection. Classification problem
  - Few false positives. FPR
  - Few false negatives. FNR
  - But what really matters are minimising monetary losses!!!

In applications, we'll need to go beyond

- Beyond a model with good fit…
- Beyond a model that predicts well…
- Fraud detection. Classification problem
    - Few false positives. FPR
    - Few false negatives. FNR
    - But what really matters are minimising monetary losses!!!

- Reservoir system management. Forecasting model for inputs and demands

Feeds decision model e.g to minimize energy deficit, wasted water, given constraints…..

- Aviation safety risk management. Forecasting models for accidents and incidents, as well as their multiple impacts

Feed a risk managment model: optimal safety resource allocation gven constraints…

- Robot control. Forecasting model for user and environments

Feeds robot control model: optimal robot decisions over time, given constraints…

PML. NI meets AI

- Bayesian inference provides a unified and coherent approach to problems of interest in Statistics, inference, prediction and decision support. Thus, to (most) ML problems

Yet mainstream ML focuses on MLE or MLE+regularization, check IntroML at https://datalab-icmat.github.io/courses_stats.html

- But things are changing slowly…
- PML leads to complex computational problems, some of which yet to be solved.
- An introduction to what is known (and what is yet to be discovered)
- But also intro to key Bayesian concepts
- In relation to key models in ML applications: supervised, unsupervised, reinforced, semisupervised,…

Some objectives of PML-RG

- Introduce key concepts in PML as well as key models motivated by real problems
- Introduce key computational methods
- Showcase methods in realistic problems
- A Bayesian view on popular ML models (supervised, unsupervised, reinforced, semisupervised)
- Community building

Basic PML concepts through an example

Basic concepts!!!

- Inference/Learning: Beyond Point Estimation, Interval estimation, Hypothesis testing
- Prediction
- Decision Support

- Uncertainty almost ubiquitous
  - Inherent to system
  - Incomplete observability
  - Incomplete modelling

- Probability as measure of degree of uncertainty with certain mathematical properties

- Interpretations
  - Classical
  - Frequentist
  - ***Subjective***

https://www.youtube.com/watch?v=KxV5kckOVeA

https://www.youtube.com/watch?v=L1Q7w3ch3

https://www.youtube.com/watch?v=OWjWYyG4Oys

Basic concepts!!!!

- Conditional probability

$$P(\mathrm{y} = y \mid \mathrm{x} = x) = \frac{P(\mathrm{y} = y, \mathrm{x} = x)}{P(\mathrm{x} = x)}$$

- Independence $\mathrm{x} \perp \mathrm{y}$

$$\forall x \in \mathrm{x}, y \in \mathrm{y},\ p(\mathrm{x} = x, \mathrm{y} = y) = p(\mathrm{x} = x) p(\mathrm{y} = y)$$

- Conditional independence $\mathrm{x} \perp \mathrm{y} \mid \mathrm{z}$

$$\forall x \in \mathrm{x}, y \in \mathrm{y}, z \in \mathrm{z},\ p(\mathrm{x} = x, \mathrm{y} = y \mid \mathrm{z} = z) = p(\mathrm{x} = x \mid \mathrm{z} = z) p(\mathrm{y} = y \mid \mathrm{z} = z)$$

Ultrabasic concepts!!!!

- Marginal distribution

$$P(\mathrm{x}, \mathrm{y})$$

$$\forall x \in \mathrm{x}, P(\mathrm{x} = x) = \sum_y P(\mathrm{x} = x, \mathrm{y} = y).$$

$$p(x) = \int p(x, y) dy$$

- Bayes rule

$$P(\mathrm{x} \mid \mathrm{y}) = \frac{P(\mathrm{x}) P(\mathrm{y} \mid \mathrm{x})}{P(\mathrm{y})}$$

$$P(\mathrm{y}) = \sum_x P(\mathrm{y} \mid x) P(x)$$

Beta-binomial model: A typical example

Consider recovery protocols for an SME computer service after a cyber attack. We introduce one protocol and wish to assess it, e.g. to be compared with another one.

Protocol tested in 12 attacks. Effective in 9 (e.g. attack duration was less than one hour)

Let's start with the model

A typical example

- n trials (identical, independent). Two results: success, failure
- Number X of successes in n trials
- Success probability in a trial θ_1
- Distribution of number of successes in n trials $X|\theta_1 \sim Bin(12, \theta_1)$
- For X=9,

$$Pr(X = 9|\theta_1) \propto \theta_1^9 (1 - \theta_1)^3, \ \theta_1 \in [0, 1]$$

A typical example

Likelihood $Pr(X = 9|\theta_1) \propto \theta_1^9(1 - \theta_1)^3,\ \theta_1 \in [0, 1]$

First approach: Maximise likelihood --→ Maximum likelihood estimator MLE

The MLE is 9/12

But MLE has several defects...

A typical example

We may use another source of information about the parameter. The prior distribution, e.g.

$$p(\theta_1) = 1, \theta_1 \in [0,1].$$

Update it through Bayes formula, to get the posterior

$$p(\theta_1|x=9) \propto p(\theta_1) \times Pr(X=9|\theta_1) \propto \theta_1^9(1-\theta_1)^3, \ \theta_1 \in [0,1]$$

which summarises all the info available about the parameter in a distribution

$$\text{Beta } (10,4)$$

Check

http://en.wikipedia.org/wiki/Beta_distribution

A typical example

The posterior serves as prior for subsequent studies. E.g., if in the following 5 applications there are 3 successes the new posterior is

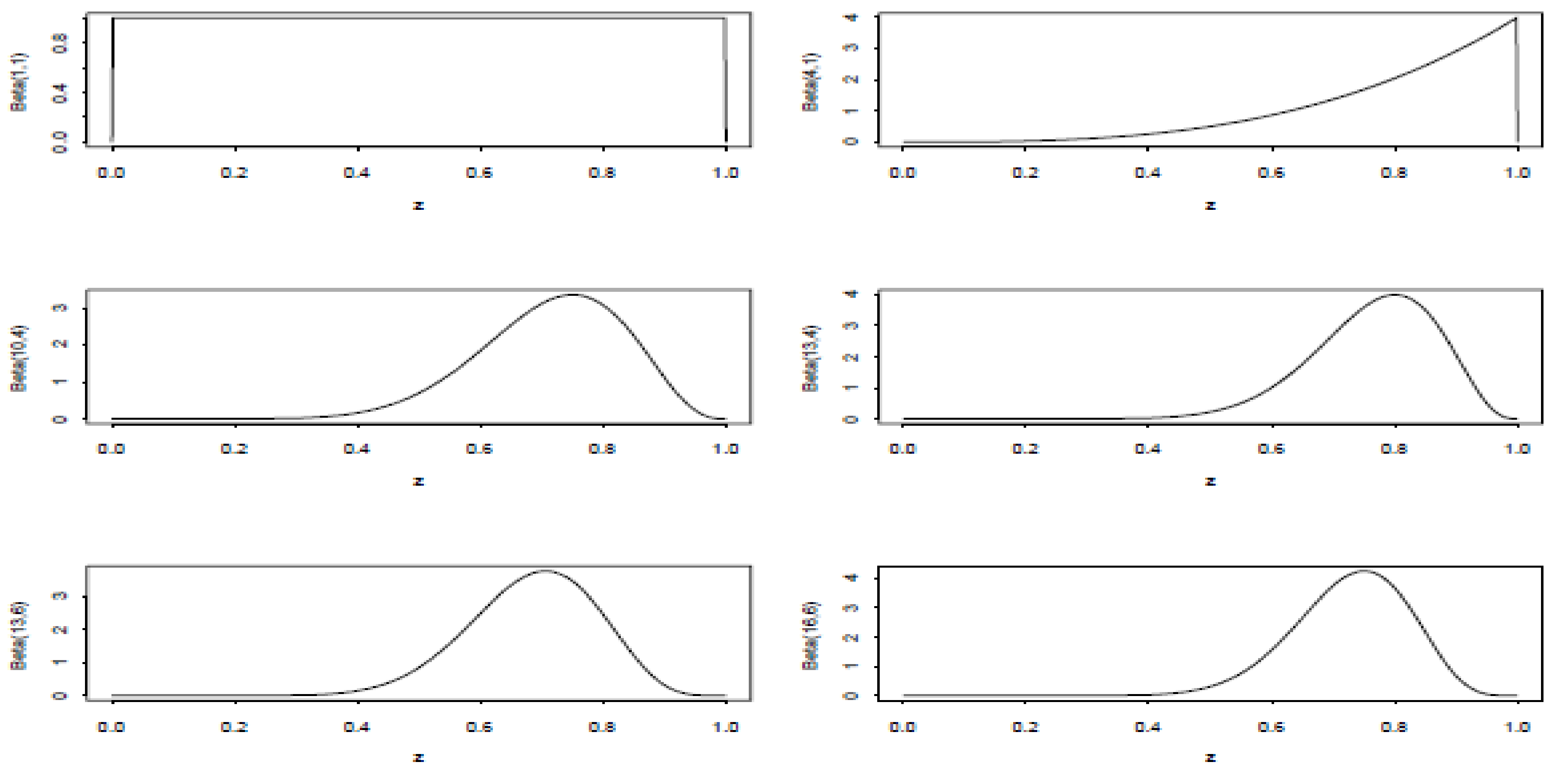
$$p(\theta_1|x=3) \propto [\theta_1^9(1-\theta_1)^3] \times [\theta_1^3(1-\theta_1)^2] = \theta_1^{12}(1-\theta_1)^5\,,\ \theta_1 \in [0,1] \qquad \text{Beta }(13{,}6)$$

Suppose that a priori, the probability is around 80% and bigger values are more likely, the learning goes through

$$\text{Beta }(4{,}1) \longrightarrow \text{Beta}(13{,}4) \longrightarrow \text{Beta}(16{,}6),$$

Sequential nature of Bayes rule

A typical example



Convergence in learning, consensus, asymptotic behavior

A typical example

- Focus on Beta (10,4). Try to use simulation for all computations also!!!

Point estimate. Summarise in a value, e.g. the posterior mean

$$\frac{10}{10+4} = 0.72$$

Why not the posterior median? Or the posterior mode (MAP)!!!

Interval estimate. Summarise interval with high probability e.g. 0.9.

- Symmetric probability wise

$$[0.505, .887]$$

- Highest posterior density interval. HDI

A typical example

- Focus on Beta (10,4)

Hypothesis testing. E.g Is the protocol effective? Null: Is the proportion bigger than 0.5

1-pbeta(0.5,10,4)=0.953

Predictions Probability of more than 4 successes in 7 trials

$$
\begin{aligned}
Pr(X = k|x = 9) &= \int Pr(X = k|\theta_1)p(\theta_1|x = 9)d\theta_1 = \\
&= \int \binom{7}{k} \theta_1^k (1-\theta_1)^{7-k} \binom{13}{3} \theta_1^9 (1-\theta_1)^3 d\theta_1 = \\
&= \frac{\binom{7}{k}\binom{13}{3}}{\binom{20}{9+k}}.
\end{aligned}
$$

$$Pr(X \geq 5|x = 9) = \sum_{k=5}^{7} Pr(X = k|x = 9) = 0.6641.$$

A typical example

Consider a second protocol. 10 opportunities, successful in 6. θ_2

Model

$$X|\theta_1 \sim Bin(12, \theta_1)$$

$$Y|\theta_2 \sim Bin(10, \theta_2)$$

$$\theta_1, \theta_2 \sim Unif[0, 1]$$

independent

Want

$$r = Pr(\theta_1 \geq \theta_2 | x = 9, y = 6)$$

A typical example

$$\theta_1 \sim Beta(10, 4),\ \theta_2 \sim Beta(7, 5)$$

- Distribution of $\theta_1 - \theta_2$????

- Through simulation. E.g 1000 observations, compute differences, count those bigger than 0, divide by 1000.

- Which protocol is better?

$$r \approx 0.772.$$

A typical example

- Utility structure

	succeeds	does not succeed
Plan A	0.8	0
Plan B	1	0.2

- Expected utilities given probabilities

$$0.8\theta_1 + 0(1 - \theta_1) = 0.8\theta_1$$

$$\theta_2 + 0.2(1 - \theta_2) = 0.2 + 0.8\theta_2$$

- Epected utilities

$$0.8E(\theta_1|x = 9) = 0.8 \times \frac{10}{14} = \frac{4}{7}$$

$$0.2 + 0.8E(\theta_2|y = 6) = 0.2 + 0.8 \times \frac{7}{12} = \frac{2}{3}.$$

Recap: Bayesian inference with the beta-binomial model

Parameter θ

Model $\Pr(X = k \mid \theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k}, \quad k = 0, 1, \dots, n$

Data x

Likelihood $\ell(\theta \mid x) \propto \theta^x (1-\theta)^{n-x}$

$h(\theta) = \log \ell(\theta \mid x) = x \log \theta + (n-x) \log(1-\theta)$

(MLE) $h'(\theta) = 0 \Rightarrow \frac{x}{\theta} - \frac{n-x}{1-\theta} = 0 \Rightarrow \hat{\theta} = x/n$

Recap: Bayesian inference with the beta-binomial model

Likelihood

$$\ell(\theta|x) \propto \theta^x (1-\theta)^{n-x}$$

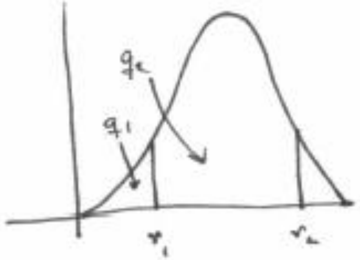
Prior

$$\pi(\theta) \propto \theta^{\alpha-1} (1-\theta)^{\beta-1} \qquad \theta \sim Be(\alpha,\beta)$$

Noninformative prior

$$\pi(\theta) = I_{[0,1]}(\theta) \, .$$

Eliciting the prior



$$\left.\begin{matrix} r_1 \\ r_2 \end{matrix}\right| \rightarrow \alpha, \beta$$

Posterior

Sequential update

Likelihood

Prior

$$\pi(\theta|x) = \frac{\pi(\theta)\, p(\theta|x)}{p(x)} \propto \pi(\theta)\, p(\theta|x) \propto \theta^x (1-\theta)^{n-x}\, \theta^{\alpha-1} (1-\theta)^{\beta-1}$$

$$= \theta^{x+\alpha-1} (1-\theta)^{n-x+\beta-1}$$

$$Be(\alpha,\beta) \rightarrow Be(\alpha+x, \beta+(n-x))$$

Recap: Bayesian inference with the beta-binomial model
(in parallel think of simulation based solutions)

Point estimation

Posterior mean

$$E(\theta|x) = \frac{\alpha+x}{\alpha+\beta+n}$$

Mix of prior and data

$$\frac{n}{\alpha+\beta+n}\,\frac{x}{n} + \frac{\alpha+\beta}{\alpha+\beta+n}\,\frac{\alpha}{\alpha+\beta}$$

What if n grows??

$$\xrightarrow[n]{} \approx \frac{x}{n} \qquad Var(\theta|x) = \frac{(\alpha+x)(\beta+n-x)}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)} \xrightarrow[n]{} 0$$

Posterior median

$$med(\theta|x) \approx \frac{\alpha+x-\frac{1}{3}}{\alpha+\beta+n-\frac{2}{3}} \qquad qbeta(0.5, \alpha+x, \alpha+\beta+n-x)$$

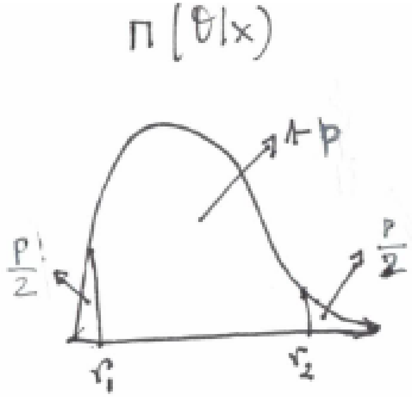
Posterior mode

$$mode(\theta|x) = \frac{\alpha+x-1}{\alpha+\beta+n-2}$$

Recap: Bayesian inference with the beta-binomial model

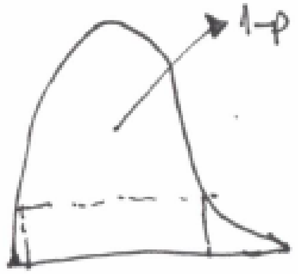
Credible interval

Symmetric interval



$$\left[qbeta\left(p/2, \alpha+x, \beta+n-x\right), qbeta\left(1-\frac{p}{2}, \alpha+x, \beta+n-x\right) \right]$$

HPD



Recap: Bayesian inference with the beta-binomial model

$\pi(\theta|x)$

Hypothesis testing

Testing three hypothesis

$H_1: \theta \in \Theta_1$ $\quad$ $H_2: \theta \in \Theta_2$ $\quad$ $H_3: \theta \in \Theta_3$

$Pr(\theta \in \Theta_1 | x)$ $\quad$ $Pr(\theta \in \Theta_2 | x)$ $\quad$ $Pr(\theta \in \Theta_3 | x)$

0-1 Loss !!!

Choose Θ_i : $\max Pr(\theta \in \Theta_i | x)$

Point nulls

??

$H_0: \theta = \theta_0$ vs $H_1: \theta \neq \theta_1$.

Credible interval R for θ.

Accept H_0 if $\theta_0 \in R$.

Evidence supports H_0

Recap: Bayesian inference with the beta-binomial model

Forecasting. The predictive distribution

m future trials

$$Pr(Y=k|x) = \int Pr(Y=k|\theta)\ \pi(\theta|x)\ d\theta \quad k=0,\dots,m$$

$$= \int \binom{m}{k} \theta^{k} (1-\theta)^{m-k} \beta(\cdot,\cdot)\ \theta^{\alpha+x-1} (1-\theta)^{\beta+(n-x)-1} d\theta$$

$$= \frac{\binom{m}{k} \beta(\alpha+x, \beta+(n-x))}{\beta(k+\alpha+x, (m-k)+\beta+(n-x))}$$

Summarising the predictive distribution

$$E(Y|x) = \sum y\ Pr(Y=k|x)$$

$$= \int \pi(\theta|x) \left(\sum y\ Pr(Y=k|\theta)\right) d\theta$$

$$= \int m\theta\ \pi(\theta|x)\ d\theta = m \frac{\alpha+x}{\alpha+\beta+n}$$

...

Recap. Exchangeability

Of data, models and parameters…..

Observations from random phenomena: independent given a certain parameter (conditionally independent)→exchangeability

http://en.wikipedia.org/wiki/Exchangeable_random_variables

Finite set of rvs exchangeable: any two permutations have the same distribution

Infinite set of rvs exchangeable: any finite subset is exchangeable

De Finetti’s theorem: set of rvs exchangeable iff ciid given a certain parametrisation

Recap: Classical vs Bayesian

Once model fixed, we want to learn about it (its parameters)

Classical	Bayesian
Parameters fixed	Parameters uncertain, prior
Given data, formulate likelihood	Given data, formulate likelihood
Maximize likelihood to find MLE (mimimum least squares, cross entropy,…)	Aggregate likelihood and prior to get posterior

Recap: ML inference

Likelihood

$$\ell(\theta|\underline{x}) = \prod_{i=1}^{n} f(x_i|\theta)$$

$$h(\theta) = \log(\ell(\theta|\underline{x}))$$

MLE

$$\max_{\theta} h(\theta) \longrightarrow \hat{\theta}$$

Recap: Bayesian inference

Prior

Noninformative prior

$f(\theta)$

Conjugate prior

Eliciting the prior

Recap: Bayesian inference

Posterior distribution. Bayes formula

$$f(\theta|x) = \frac{f(\theta) \times l(\theta|x)}{\boxed{\int f(\theta) \times l(\theta|x)\, d\theta}} = \frac{f(\theta)\, l(\theta|x)}{f(x)} \propto f(\theta)\, l(\theta|x)$$

Recap: Bayesian inference. Recall in parallel simulations for this

Point estimation

Posterior mean

$$\boxed{E(\theta|\underline{x}) = \int \theta \, f(\theta|x) \, d\theta}$$

Posterior median

$$Pr(\theta \leq med|\underline{x}) \geq \frac{1}{2} \qquad Pr(\theta \geq med|\underline{x}) \geq \frac{1}{2}$$

Posterior mode. MAP

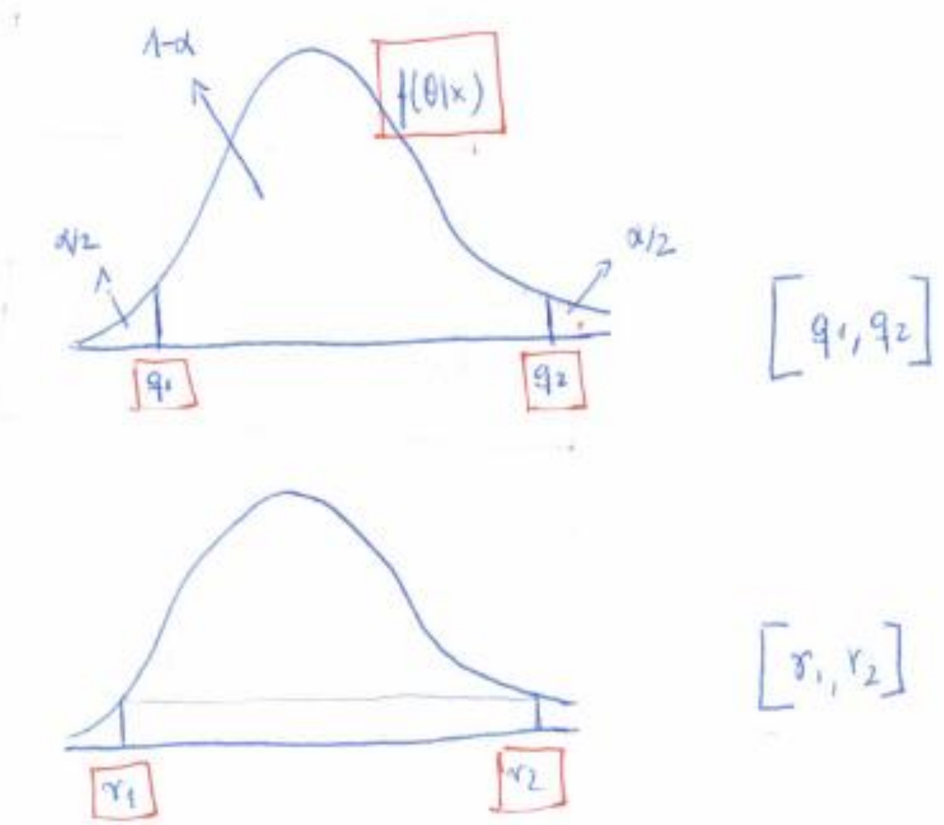
$$\operatorname{argmax} f(\theta|x)$$
$$=$$
$$\operatorname{argmax} \ell(\theta|x) f(\theta)$$
$$=$$
$$\operatorname{argmax} h(\theta) + \log f(\theta)$$

Recap: Bayesian inference

Credible interval

Symmetric interval

HPD



Recap: Bayesian inference

Hypothesis testing

$H_i : \theta \in \Theta_i$

Utility. SAYS H_i, IS $H_j \rightarrow u_{ij}$

SAY $\boxed{\arg\max_i \sum u_{ij} \Pr(\theta_j \mid \underline{x})}$

0-1 utility $u_{ij} = \delta_{ij}$

SAY $\boxed{\arg\max_i \Pr(\theta_i \mid \underline{x})}$

Recap: Bayesian inference

Forecasting

$$\boxed{f(y|\underline{x}) = \int f(y|\theta)\, f(\theta|\underline{x})\, d\theta}$$

Recap: Bayesian decision analysis

Decision analysis

$$a \in A$$

$$(a, \theta) \longrightarrow u(a, \theta)$$

$$\boxed{\arg\max_a \boxed{\int u(a, \theta)\, p(\theta \mid \underline{x})\, d\theta}}$$

$$(a, y) \longrightarrow u(a, y)$$

$$\boxed{\arg\max_a \boxed{\int u(y, \theta)\, p(y \mid \underline{x})\, d\theta}}$$

Bayes in core themes in ML. PML

- Supervised learning: Pairs input-output available

  Regression, Classification

- Unsupervised learning: Outputs not available (or the inputs are the outputs)

  Density estimation, clustering, outlier detection, Visualisation

- Reinforcement learning: Decisions impacting outputs on-the-fly

  Markov decision processes

- Semisupervised,…

Recap. Computational problems in BML

Plagued by complex integrals with complex integrands

+ optimisations

Easy conceptually… tough computationally

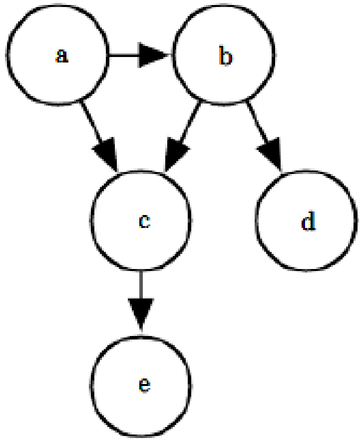
Probabilistic Graphical Models

Motivation

- Simple way to visualize structure of probabilistic models
- Designing and motivating new models
- Understanding properties like conditional independence
- Complex computations viewed through simple graphical manipulations
- Explainable and interpretable. Easy to communicate
- Classification, generation
- Deep belief nets in deep learning….

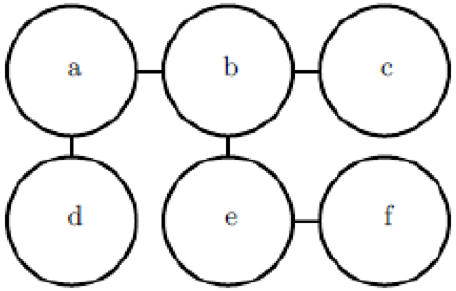
Concept

$$p(\mathbf{x}) = \prod_i p\left(\mathrm{x}_i \mid Pa_{\mathcal{G}}(\mathrm{x}_i)\right)$$



$$p(\mathrm{a}, \mathrm{b}, \mathrm{c}, \mathrm{d}, \mathrm{e}) = p(\mathrm{a})p(\mathrm{b} \mid \mathrm{a})p(\mathrm{c} \mid \mathrm{a}, \mathrm{b})p(\mathrm{d} \mid \mathrm{b})p(\mathrm{e} \mid \mathrm{c})$$

$$\tilde{p}(\mathbf{x}) = \Pi_{\mathcal{C} \in \mathcal{G}}\, \phi(\mathcal{C})$$



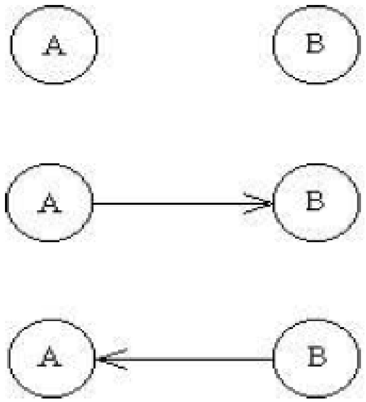
$$p(\mathrm{a}, \mathrm{b}, \mathrm{c}, \mathrm{d}, \mathrm{e}, \mathrm{f}) \quad \frac{1}{Z}\phi_{\mathrm{a},\mathrm{b}}(\mathrm{a}, \mathrm{b})\phi_{\mathrm{b},\mathrm{c}}(\mathrm{b}, \mathrm{c})\phi_{\mathrm{a},\mathrm{d}}(\mathrm{a}, \mathrm{d})\phi_{\mathrm{b},\mathrm{e}}(\mathrm{b}, \mathrm{e})\phi_{\mathrm{e},\mathrm{f}}(\mathrm{e}, \mathrm{f})$$

Bayesian networks. Directed, Acyclic

Markov fields. Undirected

Probabilistic diagrams with two nodes

Model for P(A,B)



P(A)P(B)

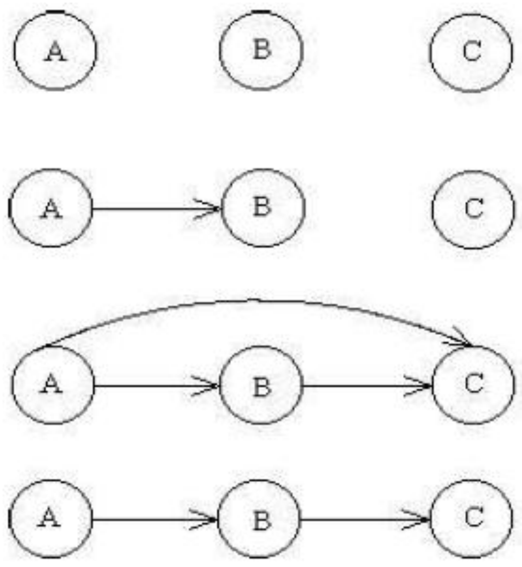
P(A) P(B|A)

P(B) P(A|B)

First case, A and B are independent. We move from second to third, and viceversa, via Bayes formula

Probabilistic diagrams with three nodes

Model P(A ,B ,C)



P(A)P(B)P(C)

P(A) P(B|A) P(C)

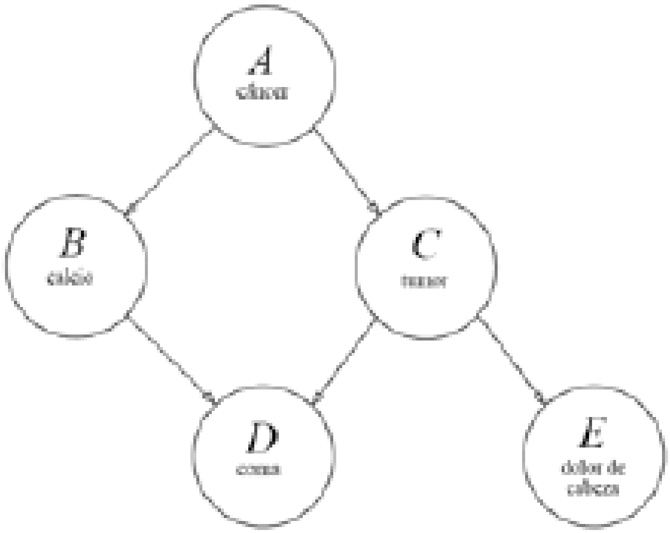
P(A)P(B|A)P(C|A,B)

P(A)P(B|A)P(C|B)

First case, independence. Fourth case, A and C are conditionally independent given B.

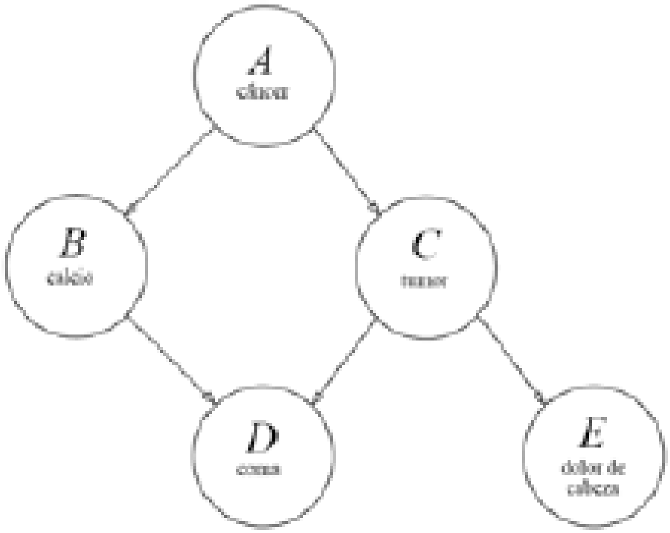
Read http://en.wikipedia.org/wiki/Conditional_independence

The hidden info



$$P(A, B, C, D, E) = P(A)P(B|A)P(C|A)P(D|B, C)P(E|C)$$

The hidden info



a	0.2

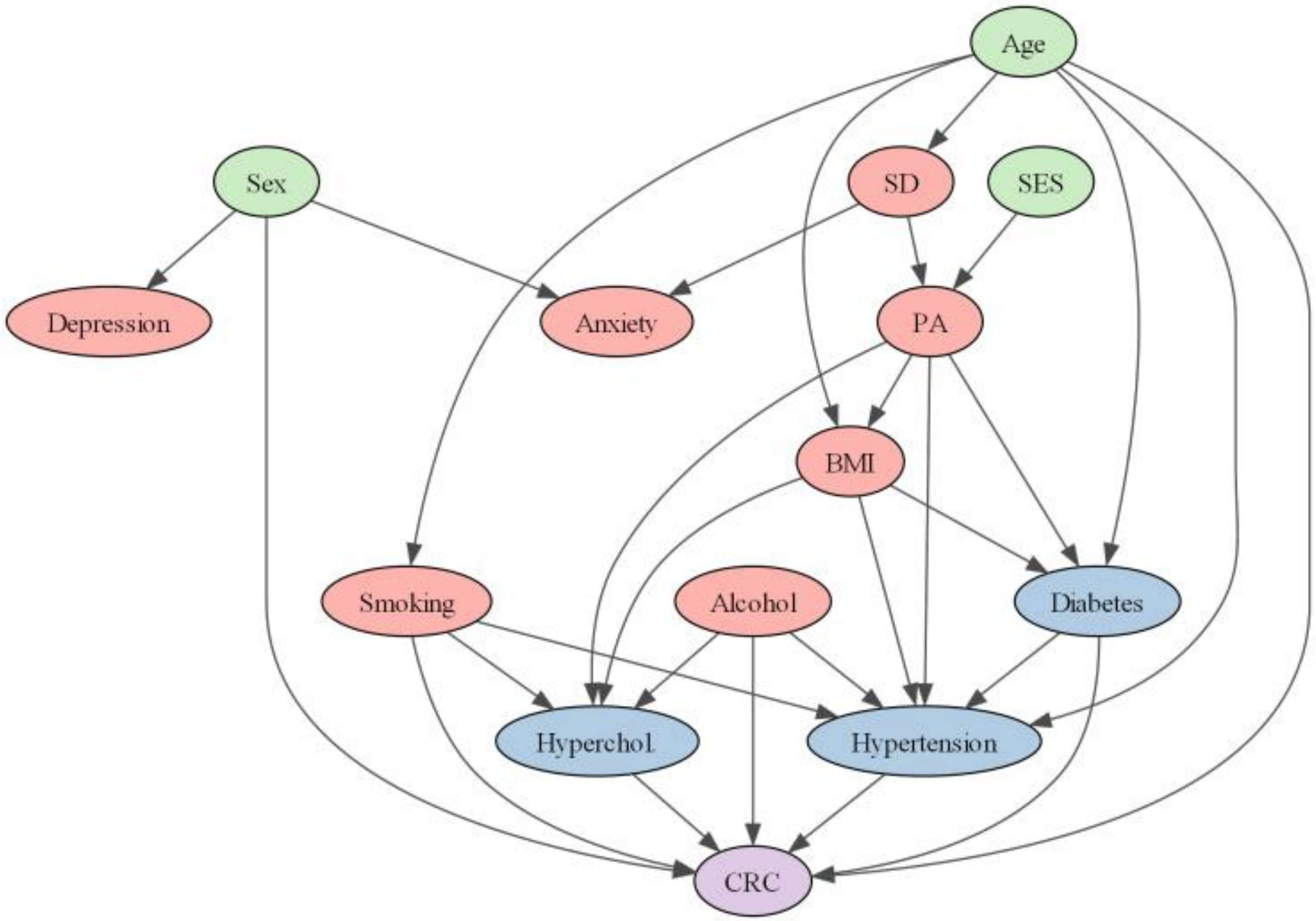
	a	$\overline{a}$
b	0.8	0.2

	a	$\overline{a}$
c	0.2	0.05

	b, c	$\overline{b}, c$	$b, \overline{c}$	$\overline{b}, \overline{c}$
d	0.8	0.8	0.8	0.05

	c	$\overline{c}$
e	0.8	0.6

$$P(A,B,C,D,E) = P(A)P(B|A)P(C|A)P(D|B,C)P(E|C)$$



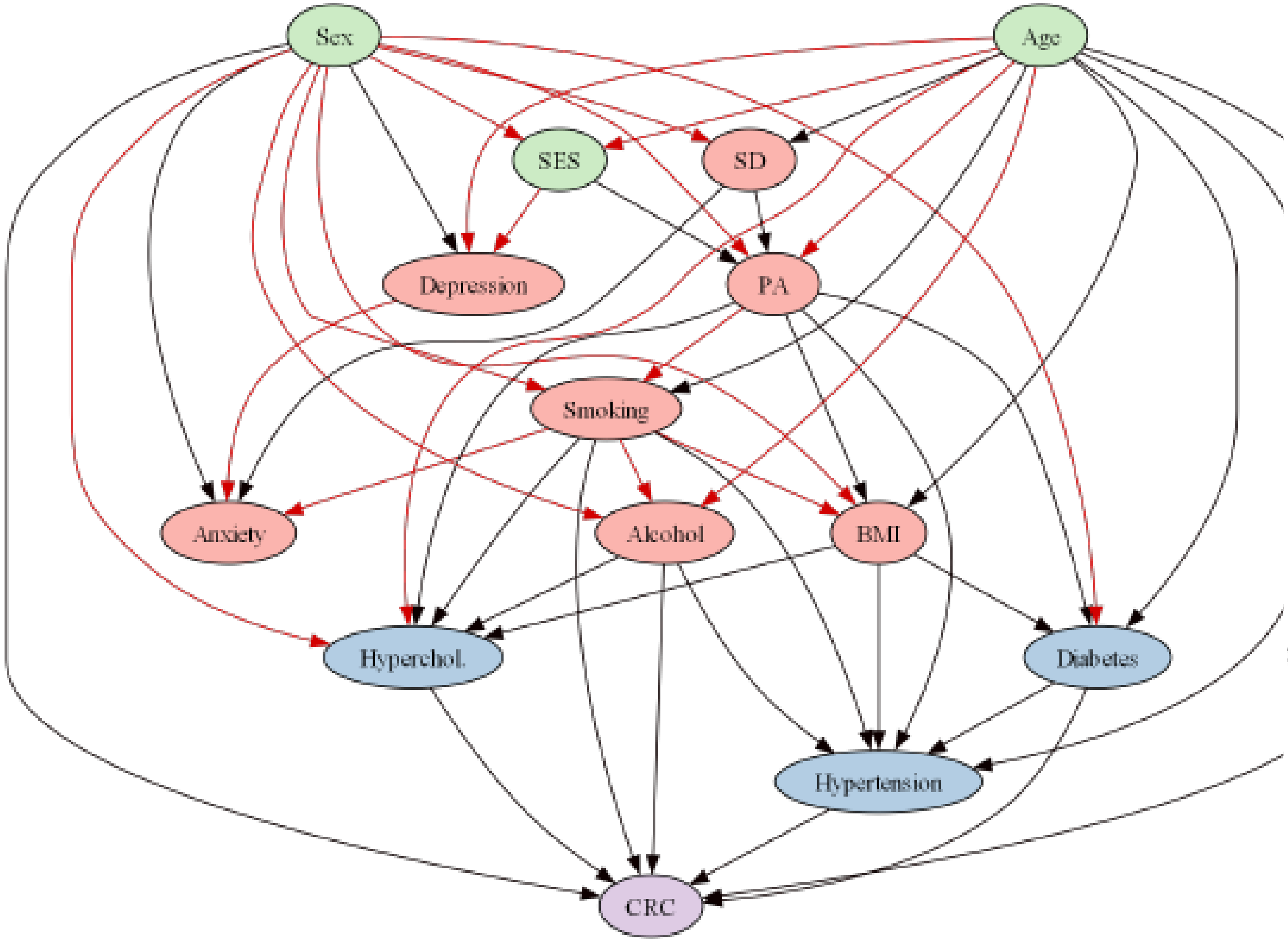
Non-modifiable
Modifiable
Medical conditions
CRC

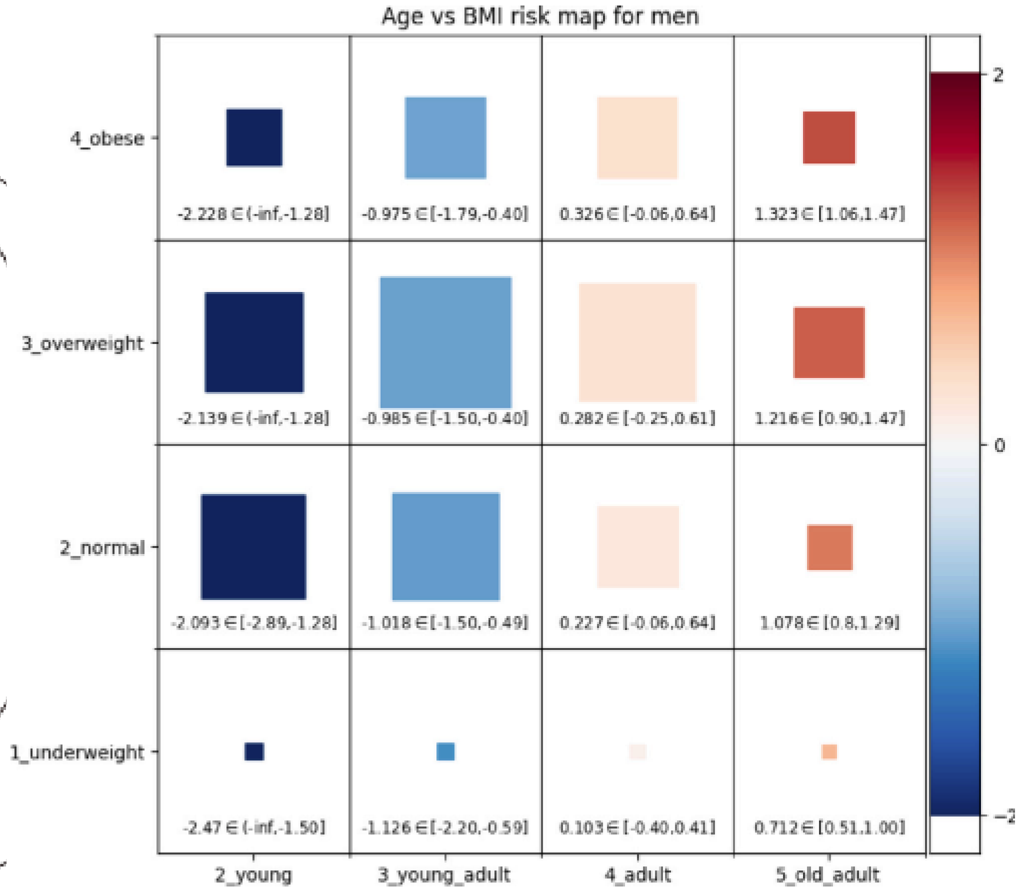
Through standard bibsearch
(Now with ChatGPT et al)

Then reviewed by experts

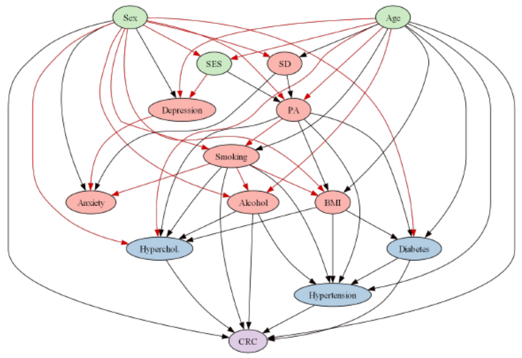
ONCOSCREEN

$$
\begin{aligned}
p(v_{sex},\ldots,v_{depression}) = & \Big[p(v_{sex})p(v_{age})p(v_{SES}|v_{sex},v_{age})\Big]\times \\
& \Big[p(v_{SD}|v_{sex},v_{age})p(v_{PA}|v_{sex},v_{age},v_{SD},v_{SES})p(v_{depr}|v_{sex},v_{age},v_{SES}) \\
& p(v_{smok}|v_{sex},v_{age},v_{PA})p(v_{alc}|v_{sex},v_{age},v_{smok}) \\
& p(v_{BMI}|v_{sex},v_{age},v_{PA},v_{smok})p(v_{anx}|v_{sex},v_{SD},v_{smok},v_{depr})\Big]\times \\
& \Big[p(v_{hypchol}|v_{sex},v_{age},v_{PA},v_{smok},v_{BMI},v_{alc})p(v_{diab}|v_{sex},v_{age},v_{PA},v_{BMI}) \\
& p(v_{hypten}|v_{age},v_{PA},v_{smok},v_{BMI},v_{alc},v_{diab})\Big]\times \\
& p(v_{CRC}|v_{sex},v_{age},v_{alc},v_{smok},v_{hypchol},v_{hypten},v_{diab}) \qquad (1)
\end{aligned}
$$





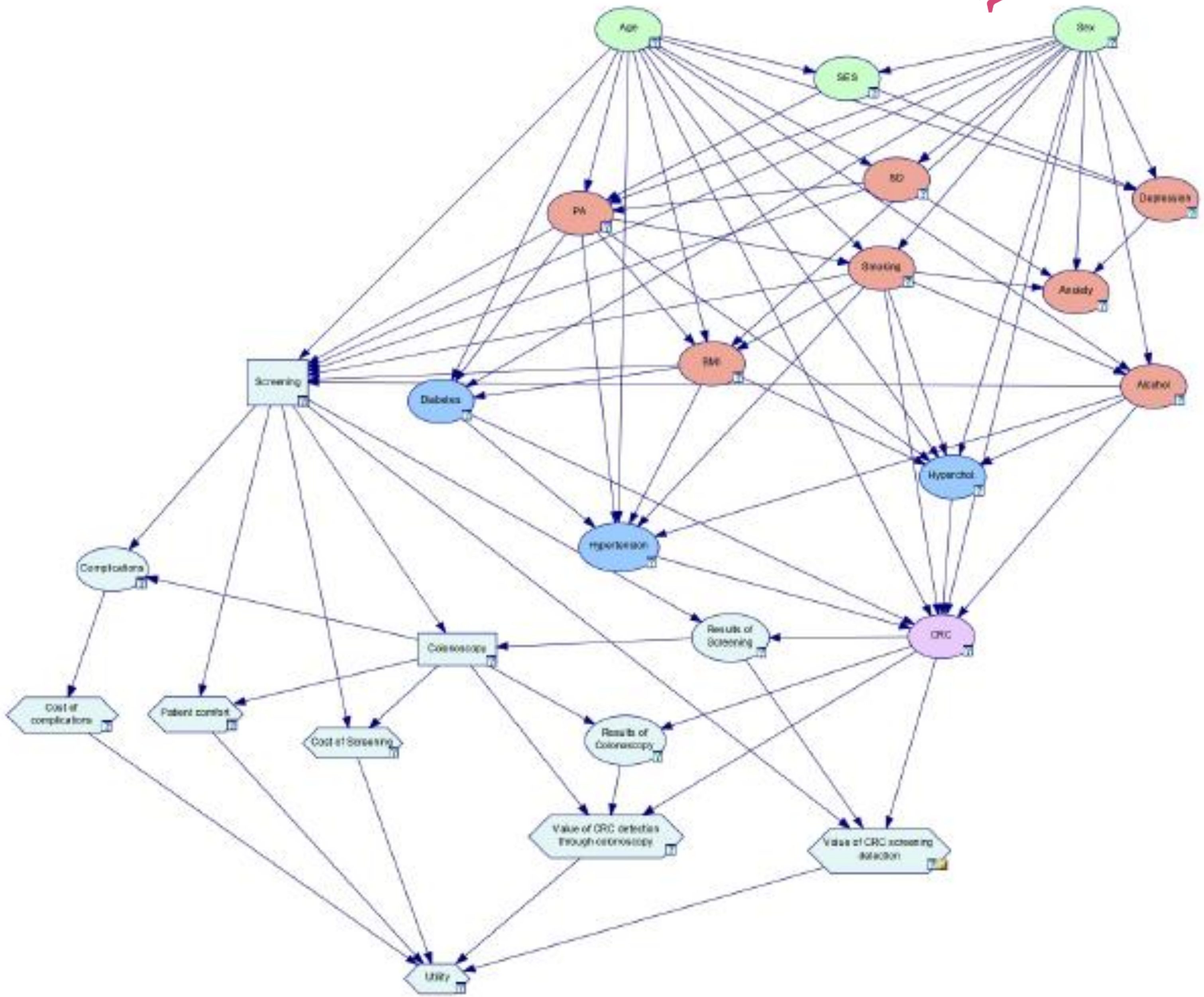
Risk maps
Influencial variables



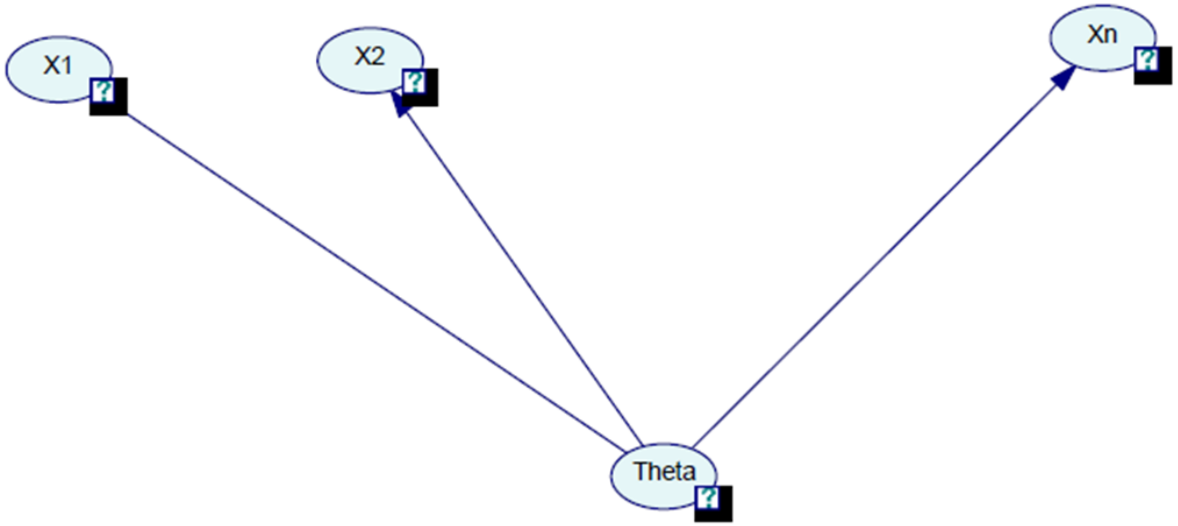
Cost intervention
Cost complications
Comfort
Information

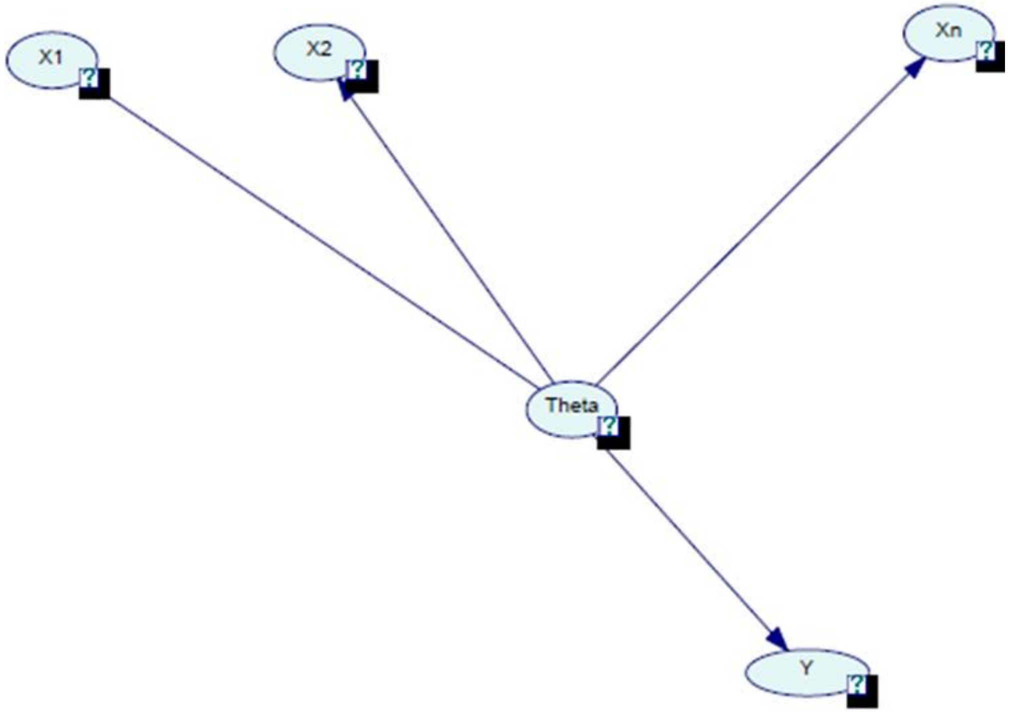
MC Value function
Risk averse utility function

Corrales et al (2024,2025)

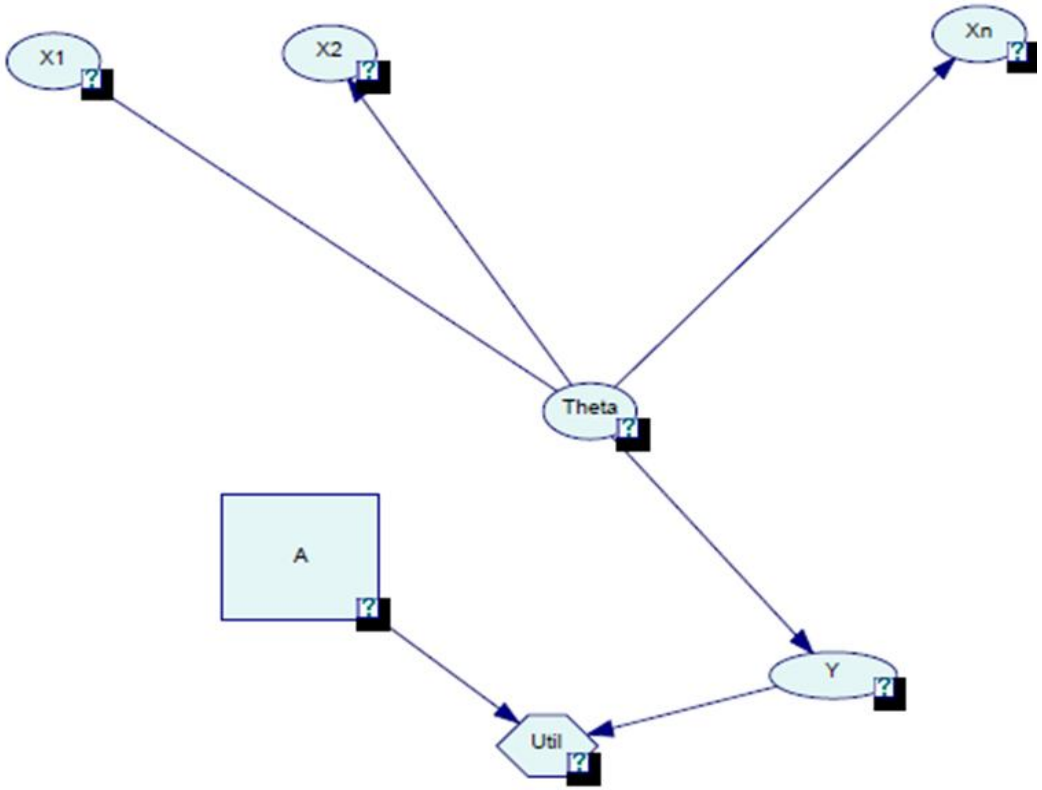


ML models as PGMs. Inference and Prediction

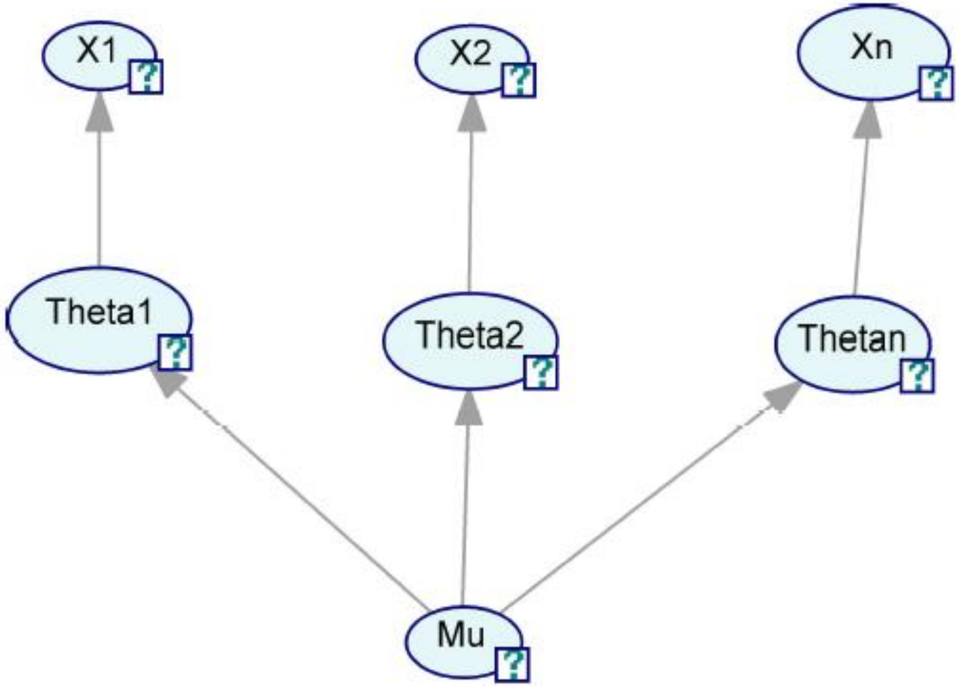




ML models as PGMs. Decision Analysis



ML models as PGMs. Hierarchical models

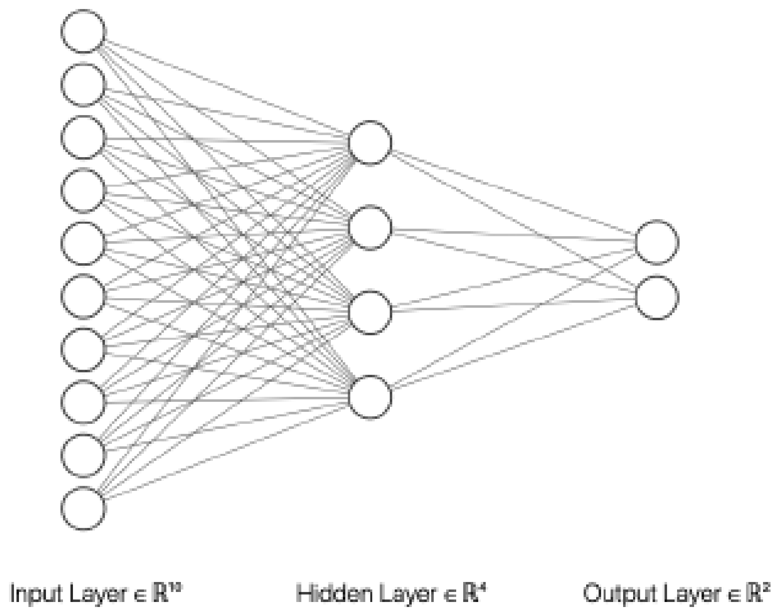


Intro to ML with neural nets. Optimization with stochastic gradient descent

Brief history of NNs

When	What	Why	Why not
End of 50's, Beg of 60's	Rosenblatt's perceptron	Efficiente scheme Good branding	Minsky& Papert (1968)
End of 80's, Beg of 90's	Cybenko's representation Shallow NNs	Good branding Impulse from CS comm	Tech problems (vanishing gradient) Emergence of SVM and others
2010's on	Deep learning, variants Outstanding aplications	Massive labeled data Rediscovery of SGD GPUs ReLUs et al Domain specific architectures Winning Imagenet comp Transformer, LLMs	

Formulation



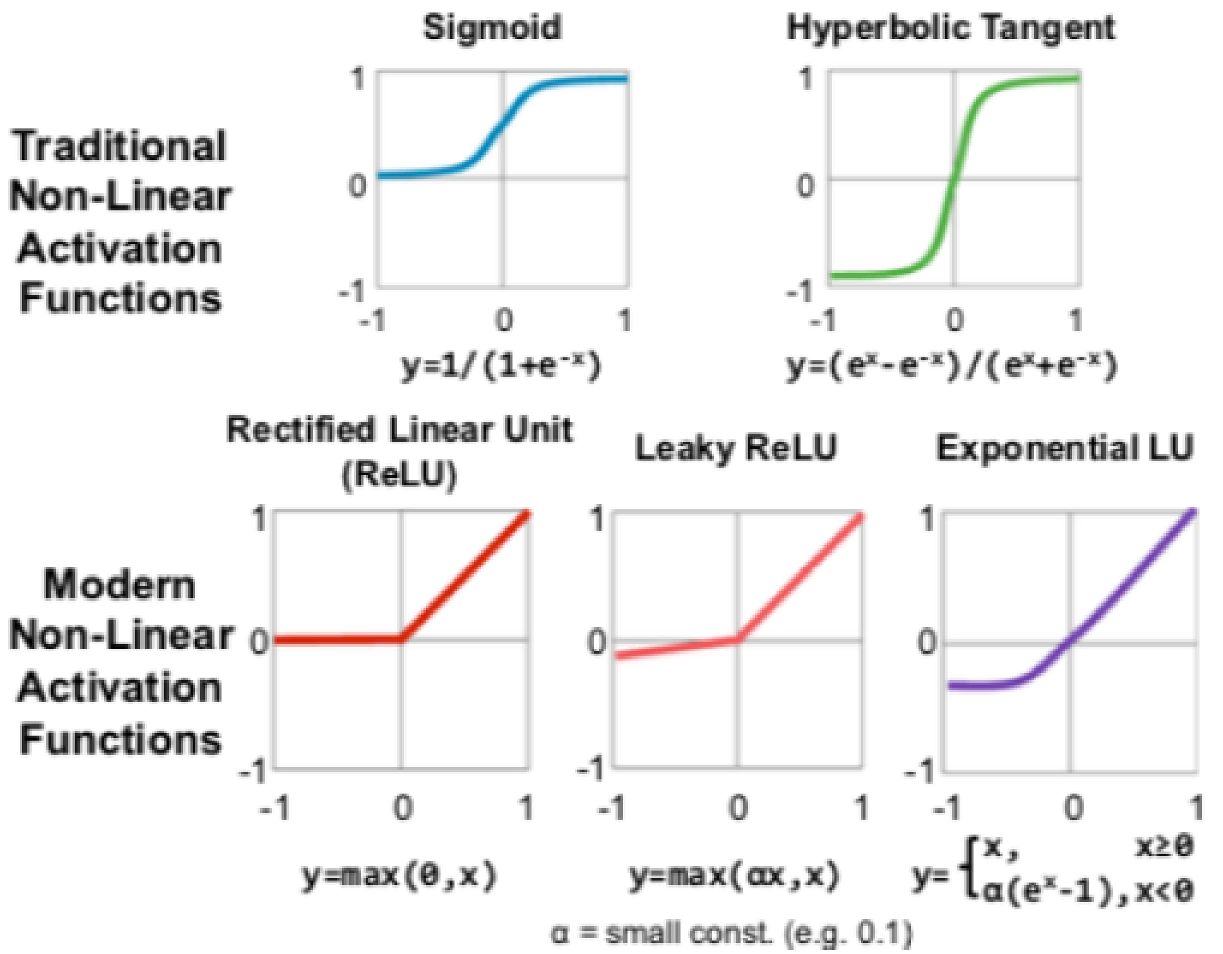
$$y = \sum_{j=1}^{m} \beta_j \psi(x'\gamma_j) + \epsilon$$

$$\epsilon \sim N(0, \sigma^2),$$

$$\psi(\eta) = \exp(\eta)/(1 + \exp(\eta))$$

Linear in beta's, nonlinear in gamma's

The evolution in activation functions



Training

Given training data, maximise log-likelihood

$$\min_{\beta,\gamma} f(\beta,\gamma) = \sum_{i=1}^{n} f_i(\beta,\gamma) = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} \beta_j \psi(x_i' \gamma_j) \right)^2$$

Gradient descent

Backpropagation to estimate gradient

Training with regularisation

$$\min_{\beta,\gamma} f(\beta,\gamma) = \sum_{i=1}^{n} f_i(\beta,\gamma) = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} \beta_j \psi(x_i' \gamma_j) \right)^2$$

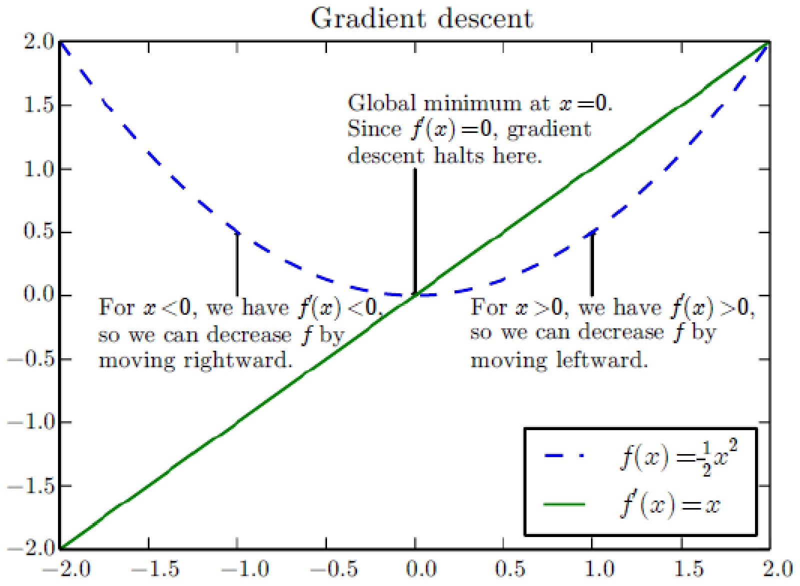
$$\min g(\beta,\gamma) = f(\beta,\gamma) + h(\beta,\gamma),$$

Weight decay

$$h(\beta,\gamma) = \lambda_1 \sum \beta_i^2 + \lambda_2 \sum \sum \gamma_{ji}^2$$

Ridge

Optimization: Using gradient info



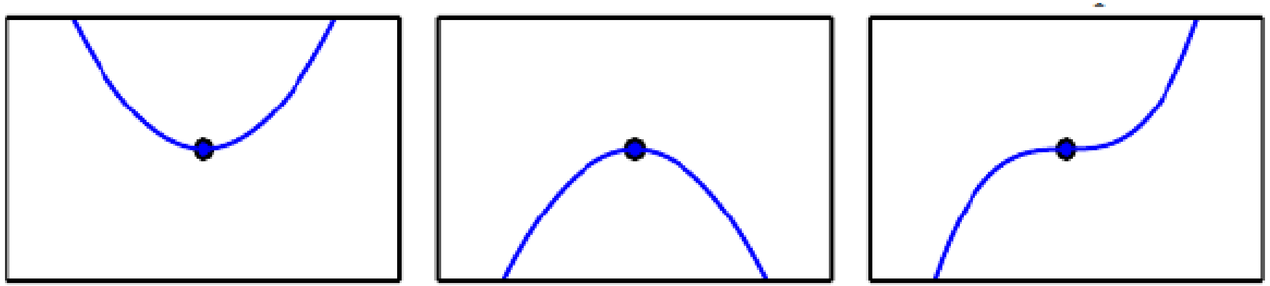
$$f(x+\epsilon) \approx f(x) + \epsilon f'(x)$$

$f(x - \epsilon\, \mathrm{sign}(f'(x)))$ < $f(x)$

$$x' = x - \epsilon \nabla_x f(x)$$

Learning rate

$f'(x) = 0$ Stationary point



Until stopping condition

Gradient descent

- Fixed and small rate
- Line search

Grad estimation. Backprop for NNs

Gradient descent

$$(\beta, \gamma)_{k+1} = (\beta, \gamma)_k - \eta \nabla g((\beta, \gamma)_k)$$

$$\nabla g((\beta, \gamma)) = \sum_{i=1}^{n} \nabla f_i(\beta, \gamma) + \nabla h(\beta, \gamma)$$

$$(\nabla_\beta f_i)_k = -2 \left(y_i - \sum_{j=1}^{m} \beta_j \psi(x_i' \gamma_j) \right) \psi(x_i' \gamma_k)$$

$$(\nabla_\gamma f_i)_{k,l} = -2 \left(y_i - \sum_{j=1}^{m} \beta_j \psi(x_i' \gamma_j) \right) \beta_l \psi(x_i' \gamma_l)(1 - \psi(x_i' \gamma_l)) x_k$$

$$(\nabla_\beta h)_k = 2\lambda_1 \beta_k \qquad (\nabla_\gamma h)_{k,l} = 2\lambda_2 \gamma_{k,l}.$$

Backpropagation (CASI 18, care with notation)

Algorithm 18.1 BACKPROPAGATION

1 Given a pair x, y, perform a "feedforward pass," computing the activations $a_\ell^{(k)}$ at each of the layers $L_2, L_3, \ldots, L_K$; i.e. compute $f(x;\mathcal{W})$ at x using the current $\mathcal{W}$, saving each of the intermediary quantities along the way.

2 For each output unit ℓ in layer L_K, compute

$$\delta_\ell^{(K)} = \frac{\partial L[y, f(x,\mathcal{W})]}{\partial z_\ell^{(K)}} = \frac{\partial L[y, f(x;\mathcal{W})]}{\partial a_\ell^{(K)}} \dot{g}^{(K)}(z_\ell^{(K)}), \tag{18.10}$$

where $\dot{g}$ denotes the derivative of $g(z)$ wrt z. For example for $L(y, f) = \frac{1}{2}\|y - f\|_2^2$, (18.10) becomes $-(y_\ell - f_\ell) \cdot \dot{g}^{(K)}(z_\ell^{(K)})$.

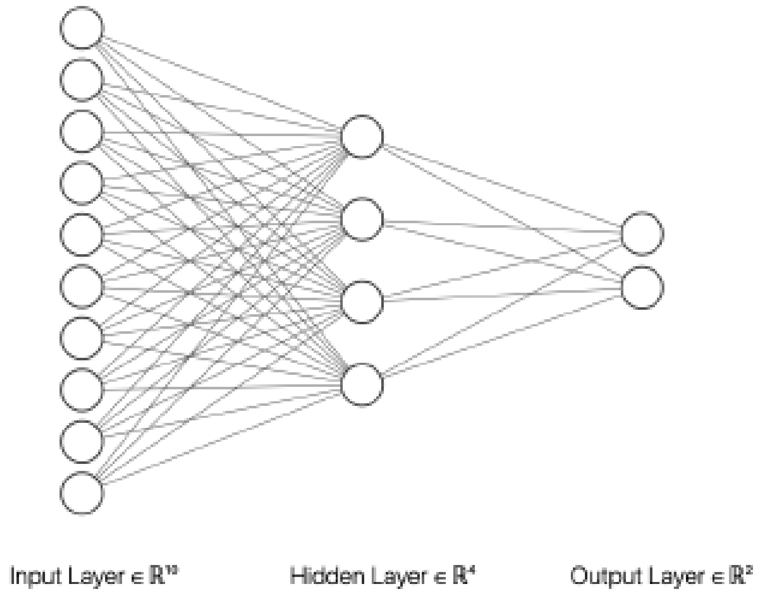
3 For layers $k = K-1,\ K-2, \ldots, 2$, and for each node ℓ in layer k, set

$$\delta_\ell^{(k)} = \left(\sum_{j=1}^{p_{k+1}} w_{j\ell}^{(k)} \delta_j^{(k+1)} \right) \dot{g}^{(k)}(z_\ell^{(k)}). \tag{18.11}$$

4 The partial derivatives are given by

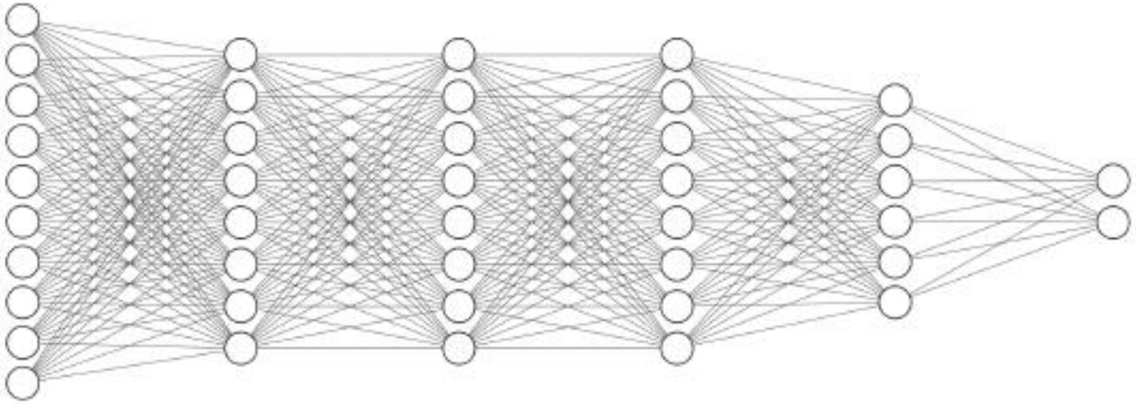
$$\frac{\partial L[y, f(x;\mathcal{W})]}{\partial w_{\ell j}^{(k)}} = a_j^{(k)} \delta_\ell^{(k+1)}. \tag{18.12}$$

Concept



$$y = \sum_{j=1}^{m} \beta_j \psi(x'\gamma_j) + \epsilon$$
$$\epsilon \sim N(0, \sigma^2),$$
$$\psi(\eta) = \exp(\eta)/(1 + \exp(\eta))$$

(Shallow) Neural nets



$$\{f_0, f_1, ..., f_{L-1}\}$$

$$z_{l+1} = f_l(z_l, \gamma_l).$$

$$y = \sum_{j=1}^{m_L} \beta_j z_{L,j} + \epsilon$$
$$\epsilon \sim N(0, \sigma^2),$$

Deep neural nets

Problems

Evaluating the objective function. Depends on n

$$\sum_{i=1}^{n} f_i(\beta, \gamma)$$

Evaluating the gradient. Depends on n

$$\sum_{i=1}^{n} \nabla f_i(\beta, \gamma)$$

Each gradient sub-term $\nabla f_i(\beta, \gamma)$ over a large number of parameters and over a long backwards recurrence

Complexity was O(w) and w is getting pretty big in Deep networks

From gradient descent…

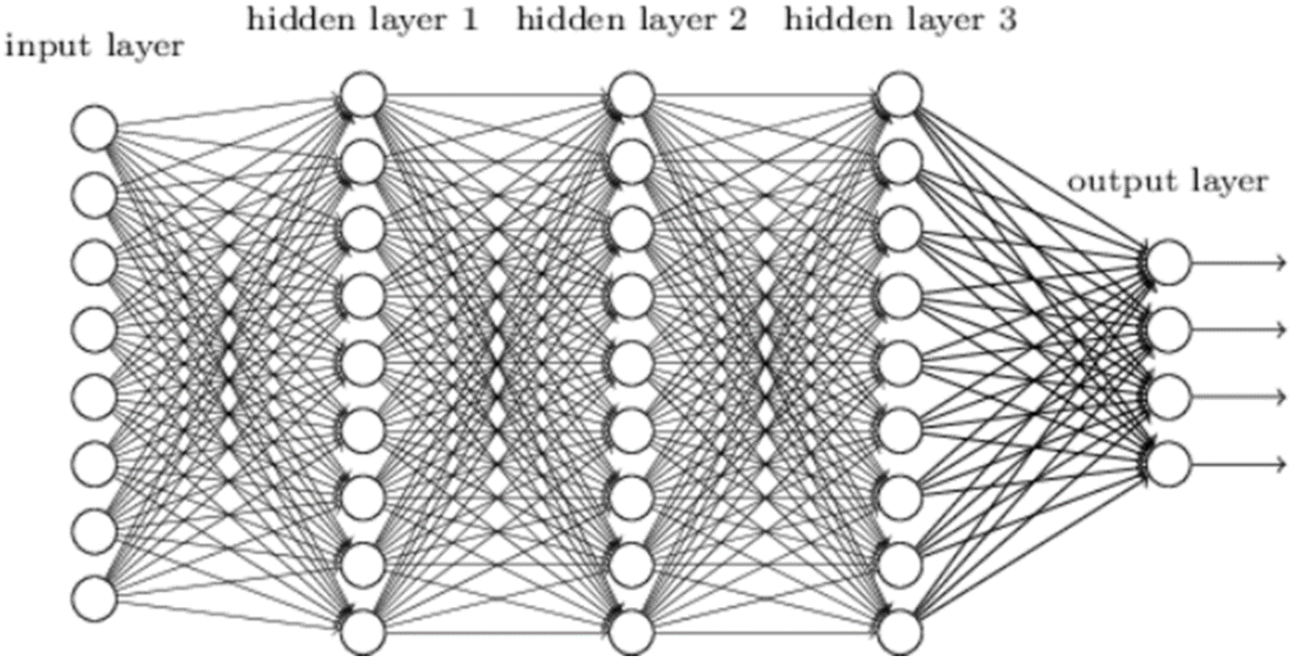
Training goes through minimising

$$J(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x},\mathrm{y}\sim\hat{p}_{\mathrm{data}}} L(\boldsymbol{x}, y, \boldsymbol{\theta}) = \frac{1}{m}\sum_{i=1}^{m} L(\boldsymbol{x}^{(i)}, y^{(i)}, \boldsymbol{\theta})$$

$$L(\boldsymbol{x}, y, \boldsymbol{\theta}) = -\log p(y \mid \boldsymbol{x}; \boldsymbol{\theta})$$

Requires gradient

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \frac{1}{m}\sum_{i=1}^{m} \nabla_{\boldsymbol{\theta}} L(\boldsymbol{x}^{(i)}, y^{(i)}, \boldsymbol{\theta})$$



Might not even fit in memory, very slow anyway

… to stochastic gradient descent

(Randomly) sample a minibatch of size m'.

Approximate gradient

$$\boldsymbol{g} = \frac{1}{m'} \nabla_{\boldsymbol{\theta}} \sum_{i=1}^{m'} L(\boldsymbol{x}^{(i)}, y^{(i)}, \boldsymbol{\theta})$$

Update via gradient descent

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \epsilon \boldsymbol{g}$$

Stochastic gradient descent

Require: Learning rate ϵ_k.
Require: Initial parameter $\boldsymbol{\theta}$
while stopping criterion not met **do**
Sample a minibatch of m examples from the training set $\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\}$ with corresponding targets $\boldsymbol{y}^{(i)}$.
Compute gradient estimate: $\hat{\boldsymbol{g}} \leftarrow +\frac{1}{m}\nabla_{\boldsymbol{\theta}} \sum_i L(f(\boldsymbol{x}^{(i)}; \boldsymbol{\theta}), \boldsymbol{y}^{(i)})$
Apply update: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \epsilon \hat{\boldsymbol{g}}$
end while

Use backprop at this stage

SGD. Robbins Monro conditions (1954!!!!)

If the learning parameters are chosen so that $\sum_{k=1}^{\infty} \epsilon_k = \infty \quad \sum_{k=1}^{\infty} \epsilon_k^2 < \infty$

and the gradient estimator is unbiased

then SDG converges a.s. (to a local optimum)

NB: The batch now fits in memory!!!

NB2: Many SGD variants

NB3: Autodiff

Intro to PML with neural nets. MCMC and variational inference

What is to be gained?

- Uncertainties in predictions
- Improved decision making based on above (risk aversion etc…)
- Some explainability via hypothesis testing
- Architecture choice
- Incorporating prior info (at least structurally)

Bayesian analysis of shallow neural nets (fixed arch)

$$y = \sum_{j=1}^{m} \beta_j \psi(x'\gamma_j) + \epsilon$$

$$\epsilon \sim N(0, \sigma^2),$$

$$\psi(\eta) = \exp(\eta)/(1 + \exp(\eta))$$

Bayesian analysis of shallow neural nets (fixed arch)

$$y = \sum_{j=1}^{m} \beta_j \psi(x'\gamma_j) + \epsilon$$

$$\epsilon \sim N(0, \sigma^2),$$

$$\psi(\eta) = \exp(\eta)/(1 + \exp(\eta))$$

$$\beta_i \sim N(\mu_\beta, \sigma_\beta^2) \text{ and } \gamma_i \sim N(\mu_\gamma, S_\gamma^2)$$

$$\mu_\beta \sim N(a_\beta, A_\beta),\ \mu_\gamma \sim N(a_\gamma, A_\gamma),\ \sigma_\beta^{-2} \sim \ Gamma(c_b/2, c_b C_b/2)$$

$$S_\gamma^{-1} \sim Wish(c_\gamma, (c_\gamma C_\gamma)^{-1}) \text{ and } \sigma^{-2} \sim Gamma(s/2, sS/2)$$

Objects of interest

'Indirectly', the posterior

$$p(\beta, \gamma, \nu | D) = \frac{p(\beta, \gamma, \nu) p(D | \beta, \gamma, \nu)}{\int p(\beta, \gamma, \nu) p(D | \beta, \gamma, \nu) d\beta d\gamma d\nu}$$

Directly, the predictive

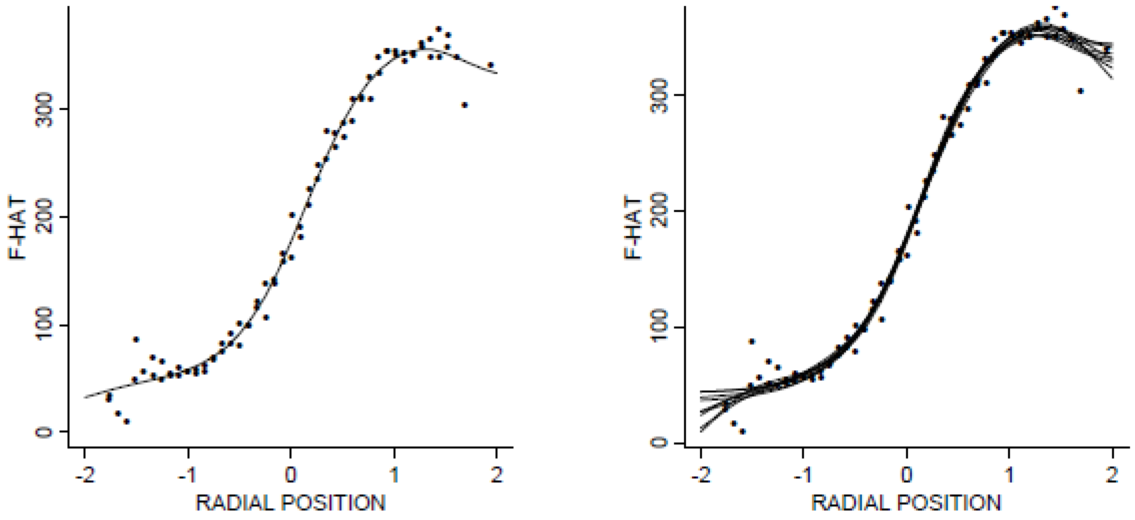
$$p(y_{N+1} | D, x_{N+1}) = \int p(y_{N+1} | \beta, \gamma, \nu, x_{N+1}) p(\beta, \gamma, \nu | D) d\beta d\gamma d\nu$$

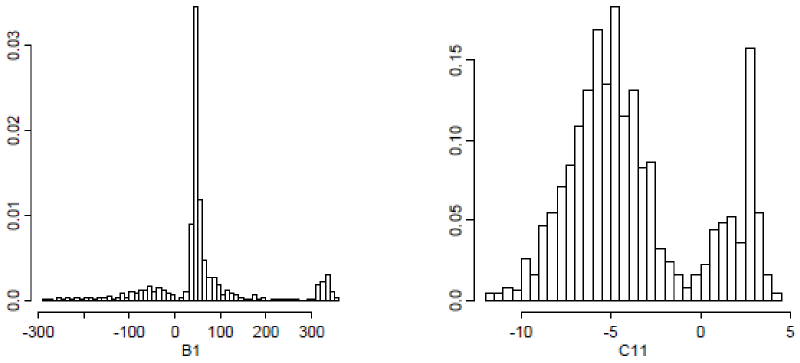
Bayesian analysis of shallow neural nets (fixed arch)

Start with arbitrary (β, γ, ν).
while *not convergence* **do**
  Given current (γ, ν), draw β from $p(\beta|\gamma, \nu, y)$ (a multivariate normal).
  for $j = 1, \ldots, m$, *marginalizing in β and given ν* **do**
    Generate a candidate $\tilde{\gamma}_j \sim g_j(\gamma_j)$.
    Compute $a(\gamma_j, \tilde{\gamma}_j) = \min\left(1, \frac{p(D|\tilde{\gamma},\nu)}{p(D|\gamma,\nu)}\right)$ with $\tilde{\gamma} = (\gamma_1, \gamma_2, \ldots, \tilde{\gamma}_i, \ldots, \gamma_m)$.
    With probability $a(\gamma_j, \tilde{\gamma}_j)$ replace γ_j by $\tilde{\gamma}_j$. If not, preserve γ_j.
  end
  Given β and γ, replace ν based on their posterior conditionals:
  $p(\mu_\beta|\beta, \sigma_\beta)$ is normal; $p(\mu_\gamma|\gamma, S_\gamma)$, multivariate normal; $p(\sigma_\beta^{-2}|\beta, \mu_\beta)$, Gamma; $p(S_\gamma^{-1}|\gamma, \mu_\gamma)$, Wishart; $p(\sigma^{-2}|\beta, \gamma, y)$, Gamma.
end

Uncertainty in predictions, explainability

$$\hat{f}(x) = \hat{E}(y_{n+1} | x_{n+1}, D) = \frac{1}{k} \sum_{t=1}^{k} E(y_{N+1} | x_{n+1}, \theta = \theta_t)$$

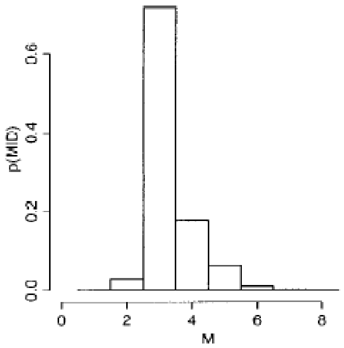




Bayesian analysis of shallow neural nets (var arch)

$$
\begin{aligned}
y &= x_i'a + \sum_{j=1}^{m^*} d_j \beta_j \psi(x'\gamma_j) + \epsilon \\
& \quad \epsilon \sim N(0, \sigma^2), \\
& \quad \psi(\eta) = \exp(\eta)/(1 + \exp(\eta)), \\
Pr(d_j = k | d_{j-1} = 1) &= (1-\alpha)^{1-k} \times \alpha^k, k \in \{0, 1\}
\end{aligned}
$$

$$
\beta_i \sim N(\mu_b, \sigma_\beta^2),\ a \sim N(\mu_a, \sigma_a^2),\ \gamma_i \sim N(\mu_\gamma, \Sigma_\gamma).
$$



Reversible jump algo, Bayesian model averaging

From shallow to deep….

Start with arbitrary (β, γ, ν).
while *not convergence* **do**
  Given current (γ, ν), draw β from $p(\beta|\gamma, \nu, y)$ (a multivariate normal).
  for $j = 1, ..., m$, *marginalizing in* β *and given* ν **do**
    Generate a candidate $\tilde{\gamma}_j \sim g_j(\gamma_j)$.
    Compute $a(\gamma_j, \tilde{\gamma}_j) = \min\left(1, \frac{p(D|\tilde{\gamma},\nu)}{p(D|\gamma,\nu)}\right)$ with $\tilde{\gamma} = (\gamma_1, \gamma_2, \ldots, \tilde{\gamma}_i, ..., \gamma_m)$.
    With probability $a(\gamma_j, \tilde{\gamma}_j)$ replace γ_j by $\tilde{\gamma}_j$. If not, preserve γ_j.
  end
  Given β and γ, replace ν based on their posterior conditionals:
  $p(\mu_\beta|\beta, \sigma_\beta)$ is normal; $p(\mu_\gamma|\gamma, S_\gamma)$, multivariate normal; $p(\sigma_\beta^{-2}|\beta, \mu_\beta)$, Gamma; $p(S_\gamma^{-1}|\gamma, \mu_\gamma)$, Wishart; $p(\sigma^{-2}|\beta, \gamma, y)$, Gamma.
end

Start with arbitrary $\theta_0 = (\beta_0, \gamma_0)$.
while *not convergence* **do**
  Given current θ_t and $q_t \sim \mathcal{N}(0, I)$, perform one or more leapfrog integration steps

$$q_{t+\frac{1}{2}} = q_t - \frac{\epsilon}{2}\nabla U(\theta_t)$$
$$\theta_{t+1} = \theta_t + \epsilon q_{t+\frac{1}{2}}$$
$$q_{t+1} = q_{t+\frac{1}{2}} - \frac{\epsilon}{2}\nabla U(\theta_{t+1})$$

  to reach θ^* and q^*.
  Compute $\alpha(\theta_t, \theta^*) = \min\left\{1, \frac{\exp H(\theta^*, r^*)}{\exp H(\theta_t, r_t)}\right\}$.
  Accept θ^* as θ_{t+1} with probability $\alpha(\theta_t, \theta^*)$, else discard it.
end

Information theoretic concepts. KL. VB

- Start with probabilistic model of observed variables **x** and latent variables **z (labels and pars)** $p(\mathbf{z}, \mathbf{x})$
- Want to estimate $p(\mathbf{z}|\mathbf{x})$ but difficult because of p(x)
- Approximate with a distribution of efficient computation $q(\mathbf{z})$
- Minimise distance so that they resemble
- Use Kullback-Leibler divergence $$KL(q||p) = \int q(\mathbf{z}) \log \frac{q(\mathbf{z})}{p(\mathbf{z}|\mathbf{x})} d\mathbf{z}$$

Disimilarity measure
Non negative
0 iff they coincid

The road ahead

Next sessions

Nov 19. Variational inference. KM-10

Dec 3. MC&Hamitonian Methods. KM-12

Dec 17. Sequential MC. KM-13

Jan 14. Bayesian Neural networks KM-17

…….

Challenges

The limits of VI
New hybrids
New bright ideas from physics?
PPLs
Decision support
The meaning of priors
Bayesian Transformers
Bayesian graduation in LLMs
Architecture selection
Security. BAML
Intepretability

ML and Stats

- J. Friedman. Data Mining and Statistics, What's the connection (1998)
- L. Breiman. Statistical Modeling,. The two cultures (2001)
- Cross Validated. What's the difference between data mining, statistics, machine learning and AI (2010)
- S.D. Sekar What's the difference between Artificial Intelligence, Machine Learning, Statistics and Big Data (2014)
- Cross Validated. What exactly is Big Data? (2015)
- David Donoho. 50 years of data science (2015)
- B. Efron, T. Hastie. Computer Age Statistical Inference (2016)
- *D. Dunson Statistics in the Big Data era: Failures of the Machine* (2019)
- D. Spiegelhalter The Art of Statistics (2020)
- *M. Hernan, J. Hsu, B. Healy A second chance to get causal inference right: a classification of data science tasks (2019)*

ML

- Efron, Hastie (2017) Computer Age Statistical Inference. Camb. UP
- Goodfellow, Bengio, Courville (2017) Deep Learning, MIT Press.
- Hastie, Tibshirani, Friedman (2009) Elements of Statistical Learning. Springer
- James, Hastie, Witten, Tibshirani (2013) An intro to Statistical Learning. Springer.
- James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). An intro to Statistical Learning: with Applications in Python. Springer

https://datalab-icmat.github.io/courses_stats.html#Introduction_to_Machine_Learning

https://llms-cunef-icmat-rg2024.github.io/

Bayes

- French, DRI (2000) Statistical Decision Theory, Wiley
- Hoff, P. (2009) A first course in Bayesian Statistical Methods, Springer
- Berger, J. (2013) Statistical decision theory and Bayesian analysis, Springer
- Robert, C. (2007) The Bayesian choice, Springer
- Gelman, A., Carlin, J….. (2013) Bayesian Data Analysis, CRC
- DRI,Ruggeri, Wiper (2012) BASP, Wiley

https://datalab-icmat.github.io/courses_stats.html#Bayesian_Data_Science

PML/Bayes and ML

- Barber (2020) Bayesian reasoning and machine learning.
- Bishop (2006) Pattern Recognition and Machine Learning. Springer
- Murphy (2014, 2022, 2023) PML
- Naveiro, DRI (2026) BAML

Thanks

- Next meeting Nov 19th