IntroML
ML. 1.3 Recap and art

DataLab ICMAT

Conceptual Recap

Recap: Classical vs Bayesian

Most approaches in ML (but not all, recall SVMs, RL…)

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
Plug in MLE to make predictions	Use predictive distribution to make predictions

Regularisers as bridges

And then used them for decision support !!!

Inference in ML

Probabilistic model of observed variables x and latent variables z (includes parameters)

$$p(\mathbf{z}, \mathbf{x})$$

ML e

$$\mathbf{z}^\star = \arg\max_{\mathbf{z}} p(\mathbf{x}|\mathbf{z})$$

MAP e

$$\mathbf{z}^\star = \arg\max_{\mathbf{z}} p(\mathbf{z}|\mathbf{x}) = \arg\max_{\mathbf{z}} p(\mathbf{x}|\mathbf{z})p(\mathbf{z})$$

Bayes e

$$p(\mathbf{z}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{p(\mathbf{x})} = \frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{\int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z}}$$

Incorporates prior info
Estimates full distribution
Denominator (evidence) frequently intractable

Recap: ML2

Likelihood

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

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

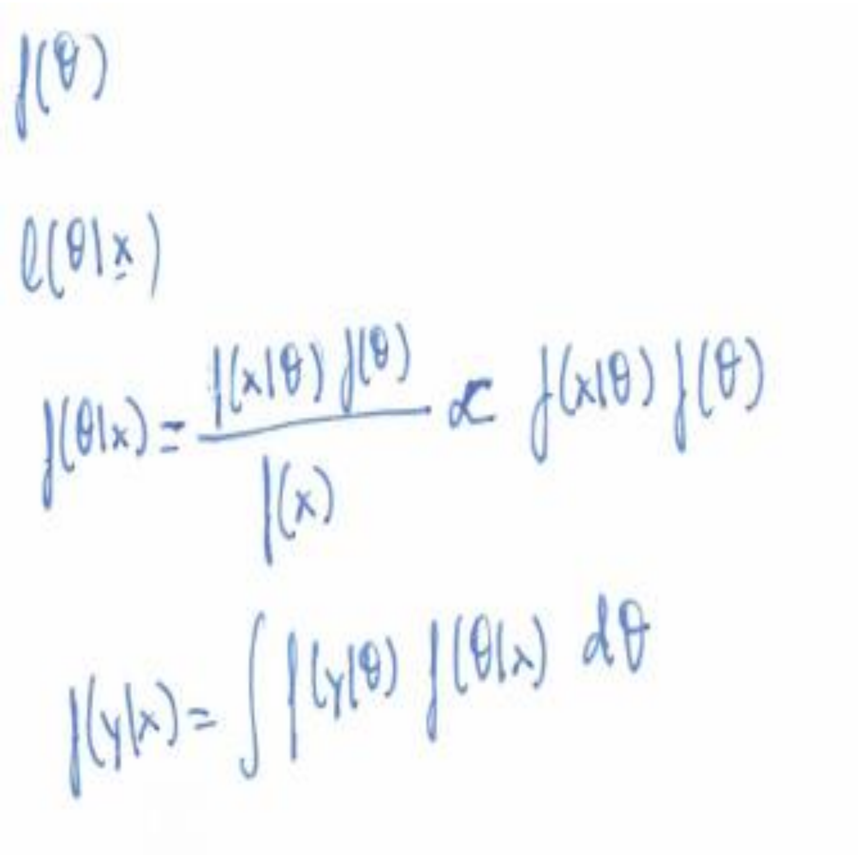
MLE

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

Predictions

$$f(x|\hat{\theta})$$

Recap: BML



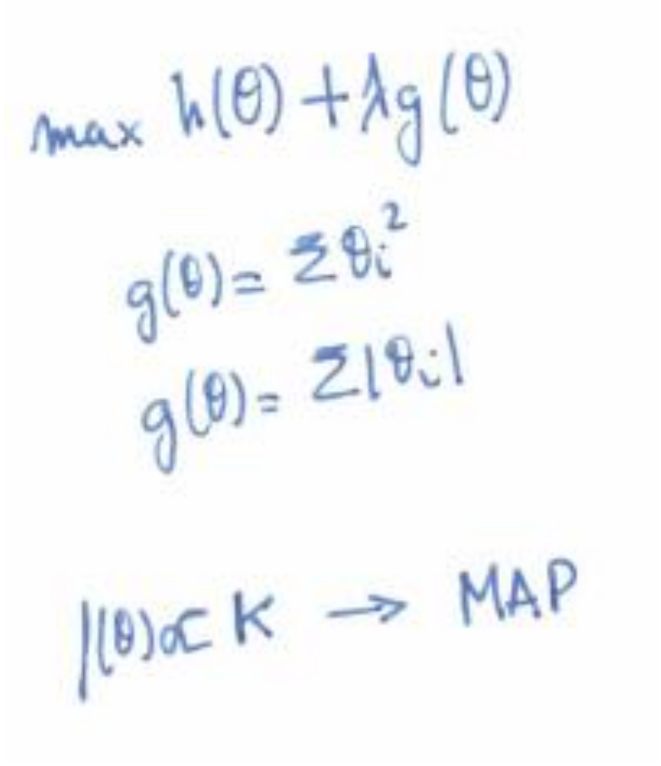
Prior $f(\theta)$

Likelihood $\ell(\theta|\underline{x})$

Posterior $f(\theta|x) = \frac{f(x|\theta)\, f(\theta)}{f(x)} \propto f(x|\theta)\, f(\theta)$

Predictive $f(y|x) = \int f(y|\theta)\, f(\theta|x)\, d\theta$

Recap: RegML



$$\max \; h(\theta) + \lambda g(\theta)$$

$$g(\theta) = \Sigma \theta_i^2$$

$$g(\theta) = \Sigma |\theta_i|$$

$$p(\theta) \propto K \rightarrow \text{MAP}$$

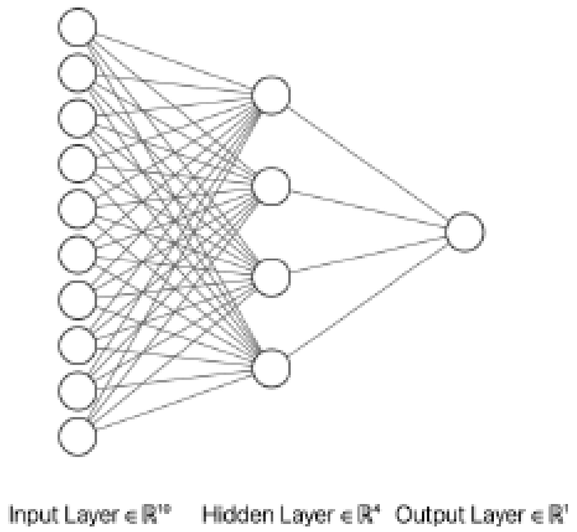
Optimisation. Recall

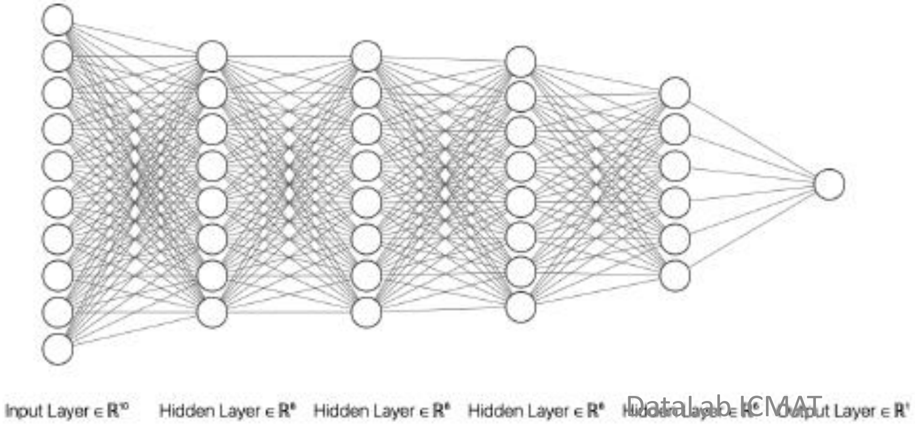
Check e.g. Goodfellow et al Ch. 4 (+8)

Optimization: Fitting neural nets (least squares, maximum likelihood)

$$y_j = \sum_{i=1}^{m} \beta_i \psi(x_k \omega_i) + \varepsilon_j$$

$$\min_{\beta, w} \sum_{k=1}^{n} \left(y_k - \sum_{i=1}^{m} \beta_i \psi(x_k \omega_i) \right)^2$$



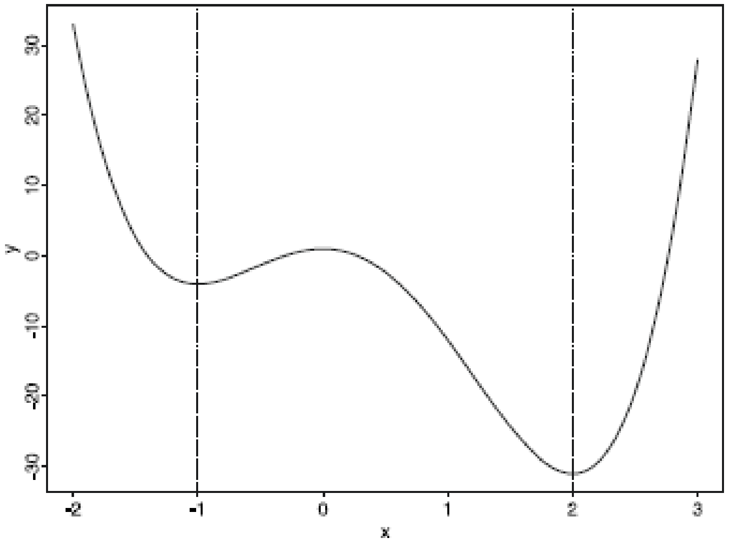


Optimization. Basic definitions

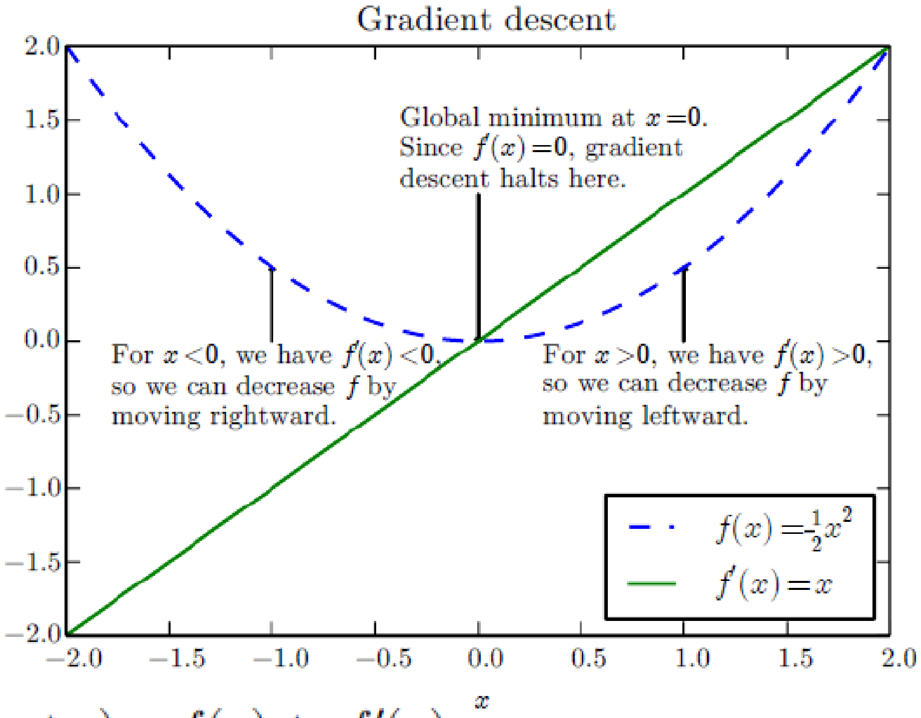
$$\min f(x)$$
$$\text{s.t. } x \in S$$

- Feasible set
- Objective function, criterion
  (cost function, loss, error)
- Local optimum
- Global optimum
- With or without constraints
- Minimization vs maximization

- (Linear programming)
- Quadratic programming
- Nonlinear programming
- Dynamic programming



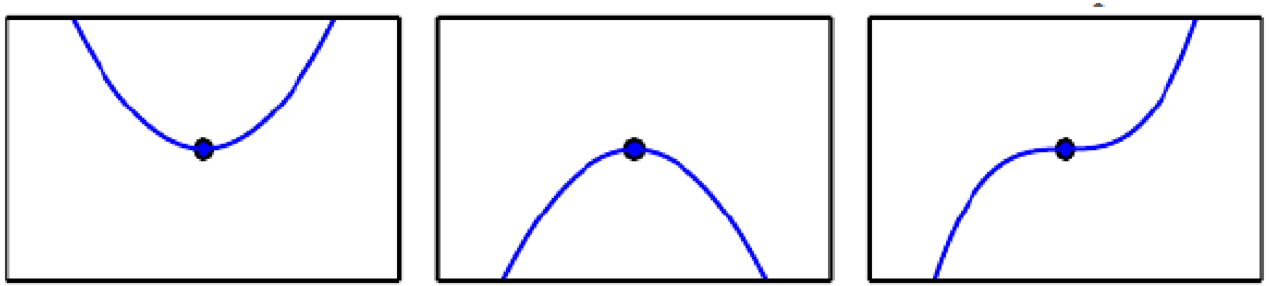
Optimization: Using gradient info



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

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

$f'(x) = 0$ Stationary point



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

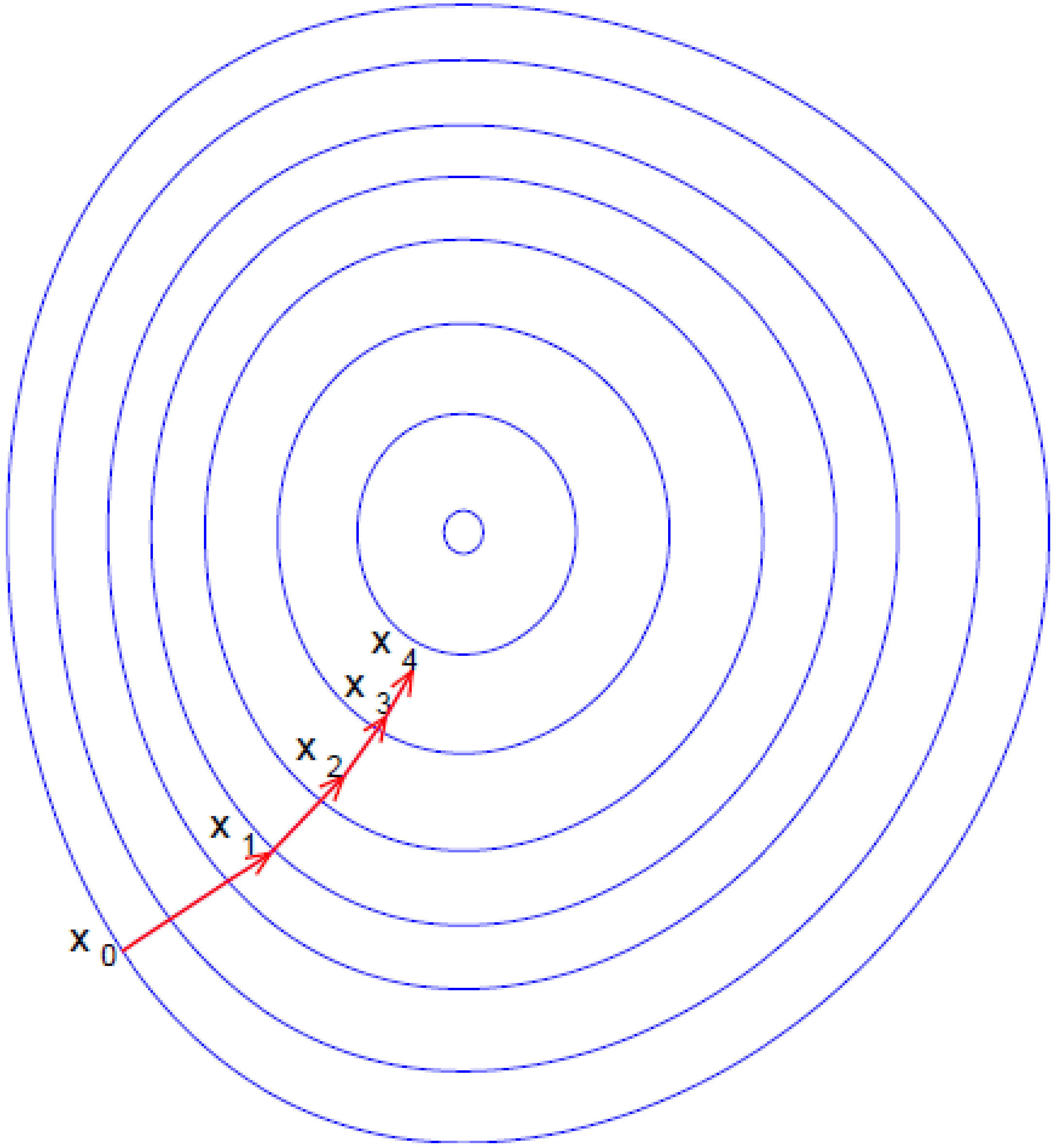
Learning rate

Until stopping condition

Gradient descent

- Fixed and small rate
- Line search

Grad estimation. Backprop for NNs


x_0
x_1
x_2
x_3
x_4

MLE optimization

Problem

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

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

Gradient

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

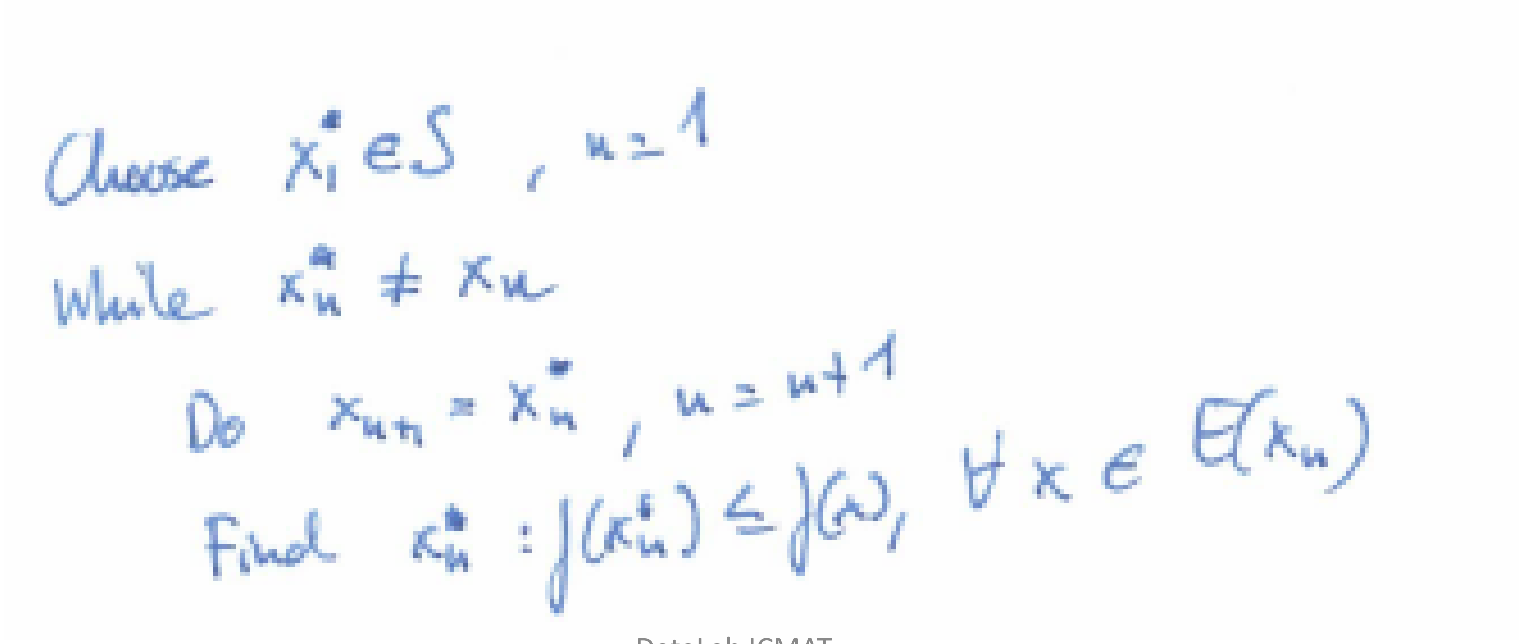
What if also regulariser?

What if m is large?

Stochastic gradient descent….

Additional remarks

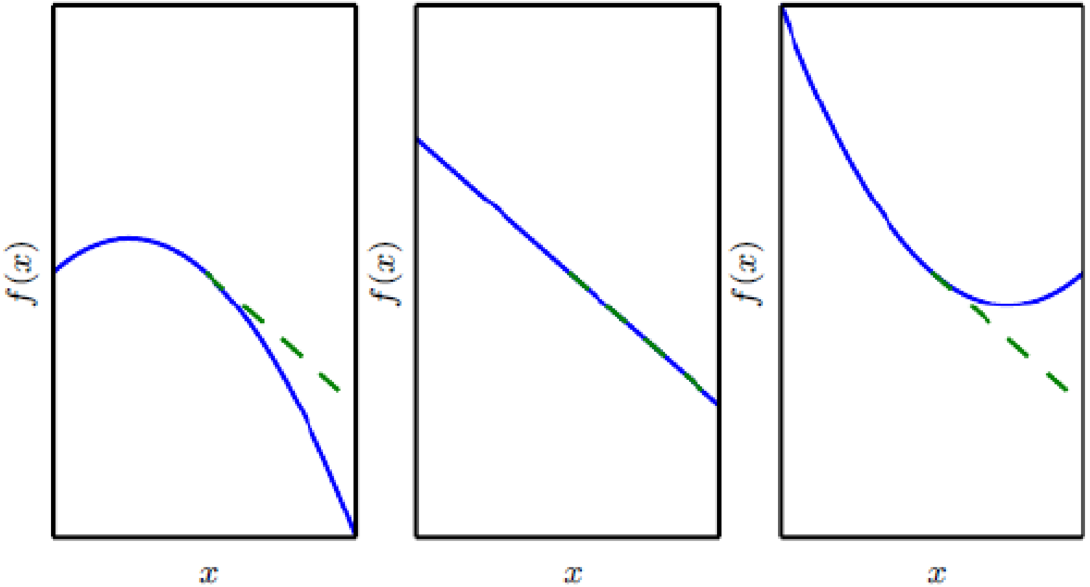
- Sometimes directly solve Gradient = 0
- Local search



```
Choose $x_1^* \in S$ , $n = 1$
While $x_n^* \neq x_n$
    Do $x_{n+1} = x_n^*$ , $n = n+1$
    Find $x_n^* : f(x_n^*) \leq f(x), \ \forall x \in E(x_n)$
```

Optimization: Adding curvature info

$f''(x)$



Hessian

$$\boldsymbol{H}(f)(\boldsymbol{x})_{i,j} = \frac{\partial^2}{\partial x_i \partial x_j} f(\boldsymbol{x})$$

Newton's method

$$f(\boldsymbol{x}) \approx f(\boldsymbol{x}^{(0)}) + (\boldsymbol{x}-\boldsymbol{x}^{(0)})^\top \nabla_{\boldsymbol{x}} f(\boldsymbol{x}^{(0)}) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}^{(0)})^\top \boldsymbol{H}(f)(\boldsymbol{x}^{(0)})(\boldsymbol{x}-\boldsymbol{x}^{(0)})$$

$$f(\boldsymbol{x}^{(0)} - \epsilon \boldsymbol{g}) \approx f(\boldsymbol{x}^{(0)}) - \epsilon \boldsymbol{g}^\top \boldsymbol{g} + \frac{1}{2}\epsilon^2 \boldsymbol{g}^\top \boldsymbol{H} \boldsymbol{g}$$

$$\boldsymbol{x}^* = \boldsymbol{x}^{(0)} - \boldsymbol{H}(f)(\boldsymbol{x}^{(0)})^{-1} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}^{(0)})$$

Optimization: Adding constraints

In many problems, constraints are added: non negative solution; small solution (in some sense, eg small norm);…

Constraints $$\mathbb{S} = \{\boldsymbol{x} \mid \forall i, g^{(i)}(\boldsymbol{x}) = 0 \text{ and } \forall j, h^{(j)}(\boldsymbol{x}) \leq 0\}$$

Lagrangian $$L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\alpha}) = f(\boldsymbol{x}) + \sum \lambda_i g^{(i)}(\boldsymbol{x}) + \sum \alpha_j h^{(j)}(\boldsymbol{x})$$

$\min_{\boldsymbol{x} \in \mathbb{S}} f(\boldsymbol{x})$ equivalent to $\min_{\boldsymbol{x}} \max_{\boldsymbol{\lambda}} \max_{\boldsymbol{\alpha}, \boldsymbol{\alpha} \geq 0} L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\alpha})$

If feasible $$\max_{\boldsymbol{\lambda}} \max_{\boldsymbol{\alpha}, \boldsymbol{\alpha} \geq 0} L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\alpha}) = f(\boldsymbol{x})$$

If unfeasible $$\max_{\boldsymbol{\lambda}} \max_{\boldsymbol{\alpha}, \boldsymbol{\alpha} \geq 0} L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\alpha}) = \infty$$

KKT conditions

Optimization: minimizing linear least squares

Minimize

$$f(\boldsymbol{x}) = \frac{1}{2}||\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}||_2^2$$

Gradient

$$\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) = \boldsymbol{A}^\top (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}) = \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{x} - \boldsymbol{A}^\top \boldsymbol{b}$$

Gradient descent

```
while $||\boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{x} - \boldsymbol{A}^\top \boldsymbol{b}||_2 > \delta$ do
  $\boldsymbol{x} \leftarrow \boldsymbol{x} - \epsilon \left(\boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{x} - \boldsymbol{A}^\top \boldsymbol{b}\right)$
end while
```

Add constraints

$$\boldsymbol{x}^\top \boldsymbol{x} \leq 1$$

Lagrangian

$$L(\boldsymbol{x}, \lambda) = f(\boldsymbol{x}) + \lambda \left(\boldsymbol{x}^\top \boldsymbol{x} - 1\right) \qquad \min_{\boldsymbol{x}} \max_{\lambda, \lambda \geq 0} L(\boldsymbol{x}, \lambda).$$

Diff in x and making 0

$$\boldsymbol{x} = (\boldsymbol{A}^\top \boldsymbol{A} + 2\lambda \boldsymbol{I})^{-1} \boldsymbol{A}^\top \boldsymbol{b}.$$

Choose lambda to satisfy constraint (iteratively)

Optimization in ML

- Multimodality
- Large scale
- Gradients expensive
- Hessian superexpensive
- …

Optimization: To be seen

Stochastic gradient descent

Variants to decrease learning rate

- (Adagrad)
- (RSMProp)
- Adam
- (Adadelta)
- ….

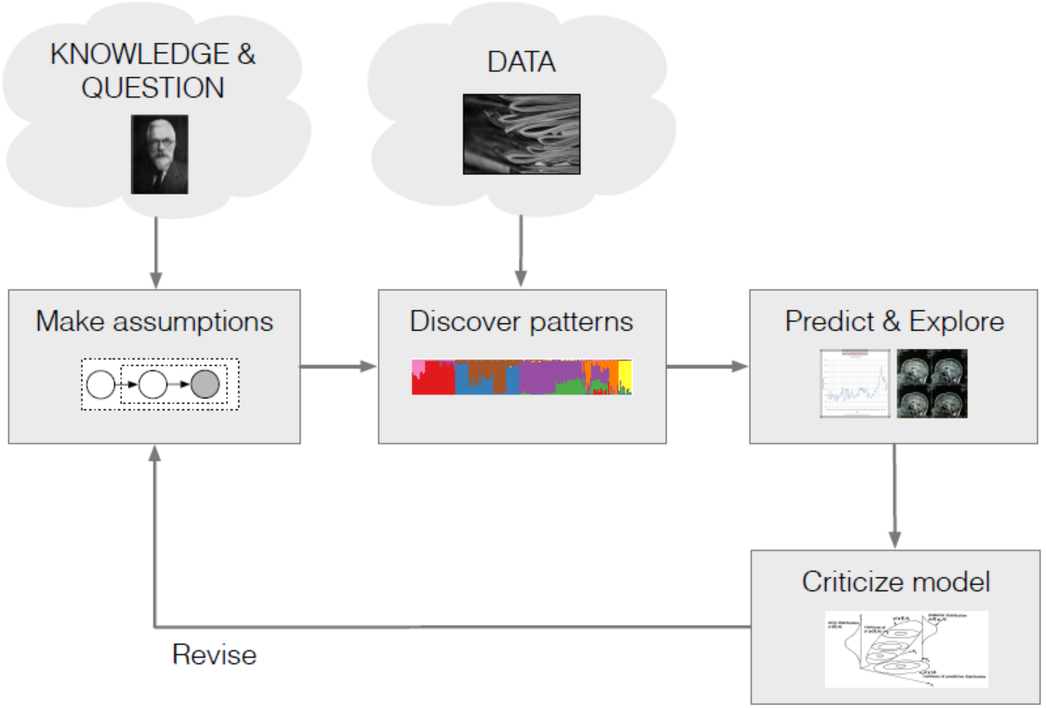
Bayesian computations to be mentioned.

But recall MAP as optimization

And variational Bayes as optimization

Data flow in machine learning

Broad learning scheme



Hypothesis

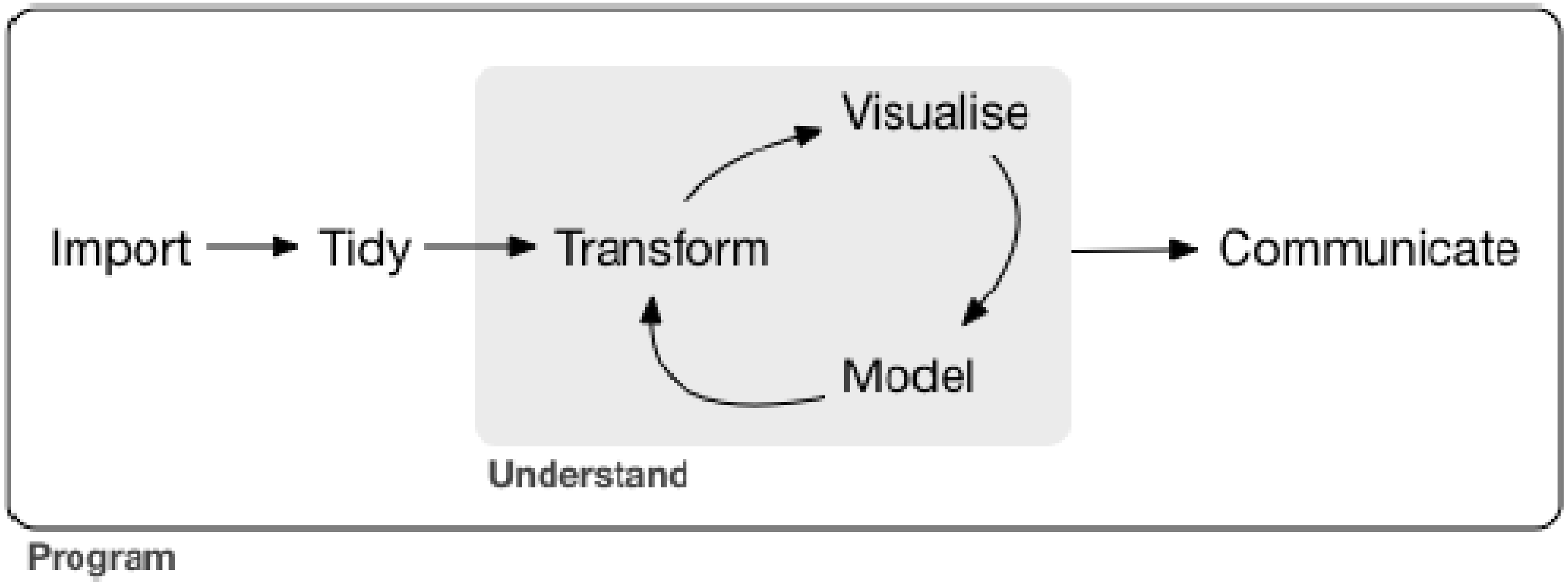
Computing

Application

Inference

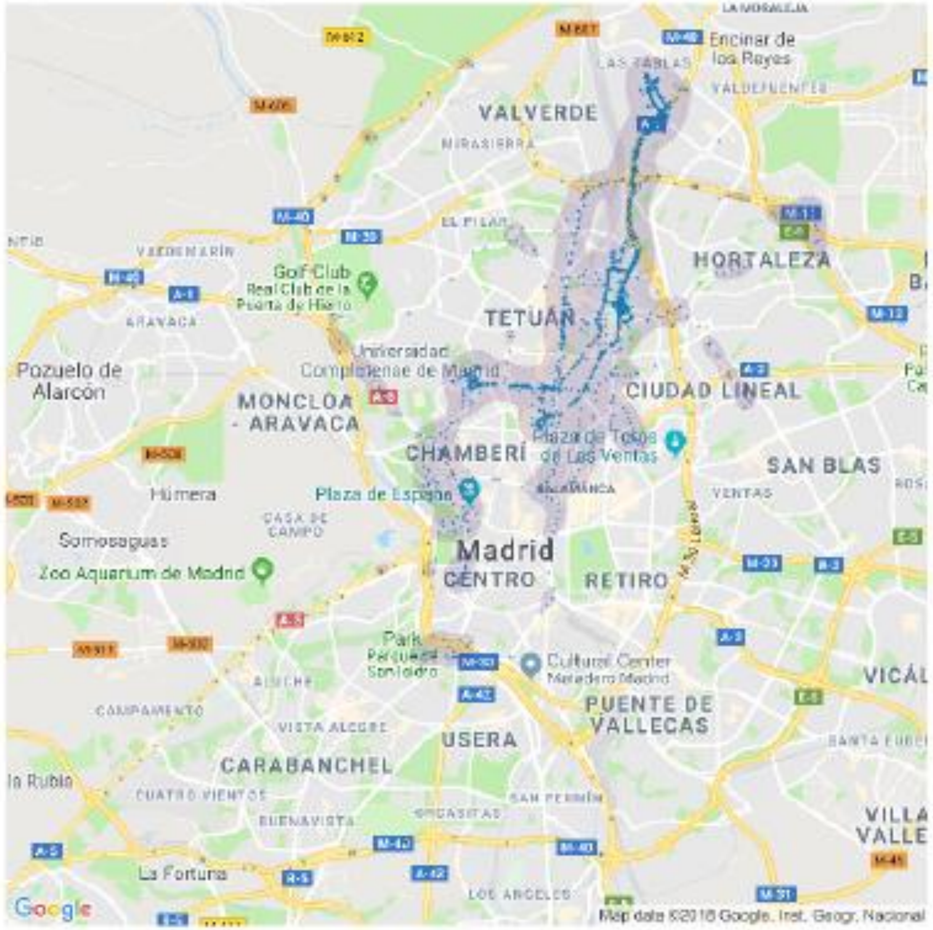
What does my model say about data?

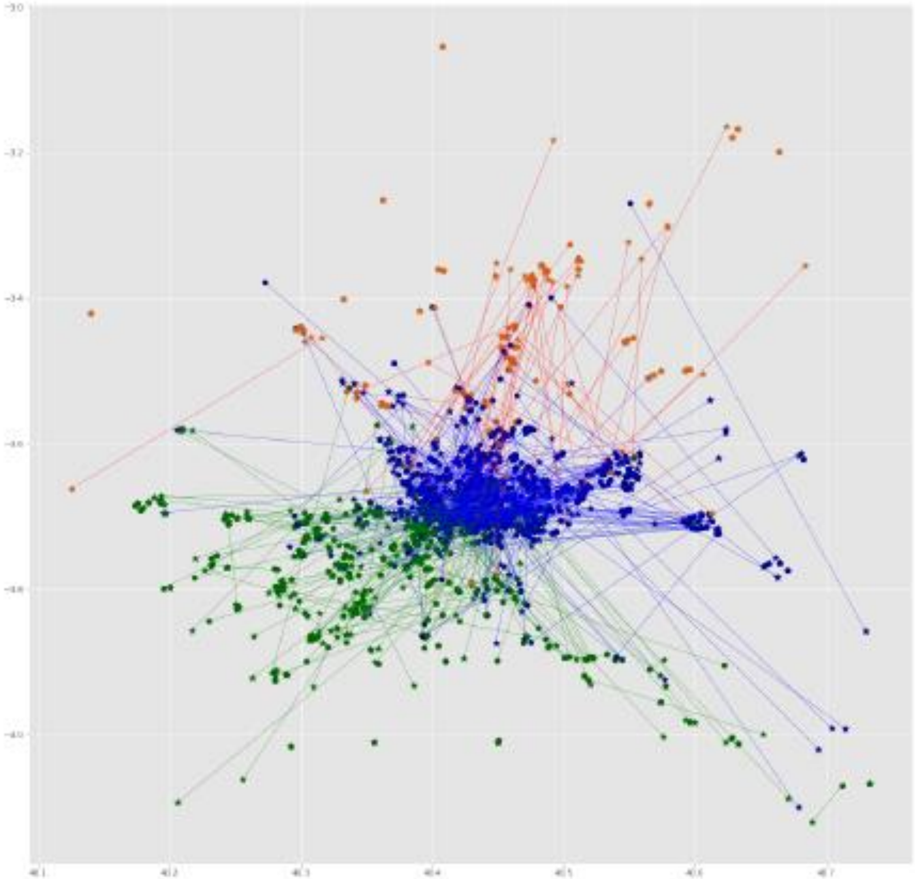
General, Scalable



Just some ideas briefly. Many more during labs and at final part

ML and BD


LA MORALEJA
Encinar de los Reyes
LAS TABLAS
VALDEFUENTES
VALVERDE
MIRASIERRA
EL PILAR
HORTALEZA
VALDEMARÍN
Golf Club Real Club de la Puerta de Hierro
ARAVACA
TETUÁN
Universidad Complutense de Madrid
Pozuelo de Alarcón
MONCLOA - ARAVACA
CIUDAD LINEAL
CHAMBERÍ
Plaza de Toros de Las Ventas
SAN BLAS
Húmera
Plaza de España
SALAMANCA
VENTAS
CASA DE CAMPO
Somosaguas
Madrid
CENTRO
RETIRO
Zoo Aquarium de Madrid
Park Parque de San Isidro
Cultural Center Matadero Madrid
VICÁLVARO
CAMPAMENTO
PUENTE DE VALLECAS
VISTA ALEGRE
USERA
CARABANCHEL
la Rubia
CUATRO VIENTOS
SAN FERMÍN
BUENAVISTA
ORCASITAS
VILLA DE VALLECAS
La Fortuna
LOS ÁNGELES
Google
Map data ©2018 Google, Inst. Geogr. Nacional



First steps. Preprocessing

- Data from heterogeneous sources (social networks, sensors, samples,….) in different support (text, data bases, streams, images,…)
- First: identify problem to be solved, available variables that may provide information
- Combine available info in a coherent manner
- Final objective of preprocessing: organise data in tensorial/tabular form

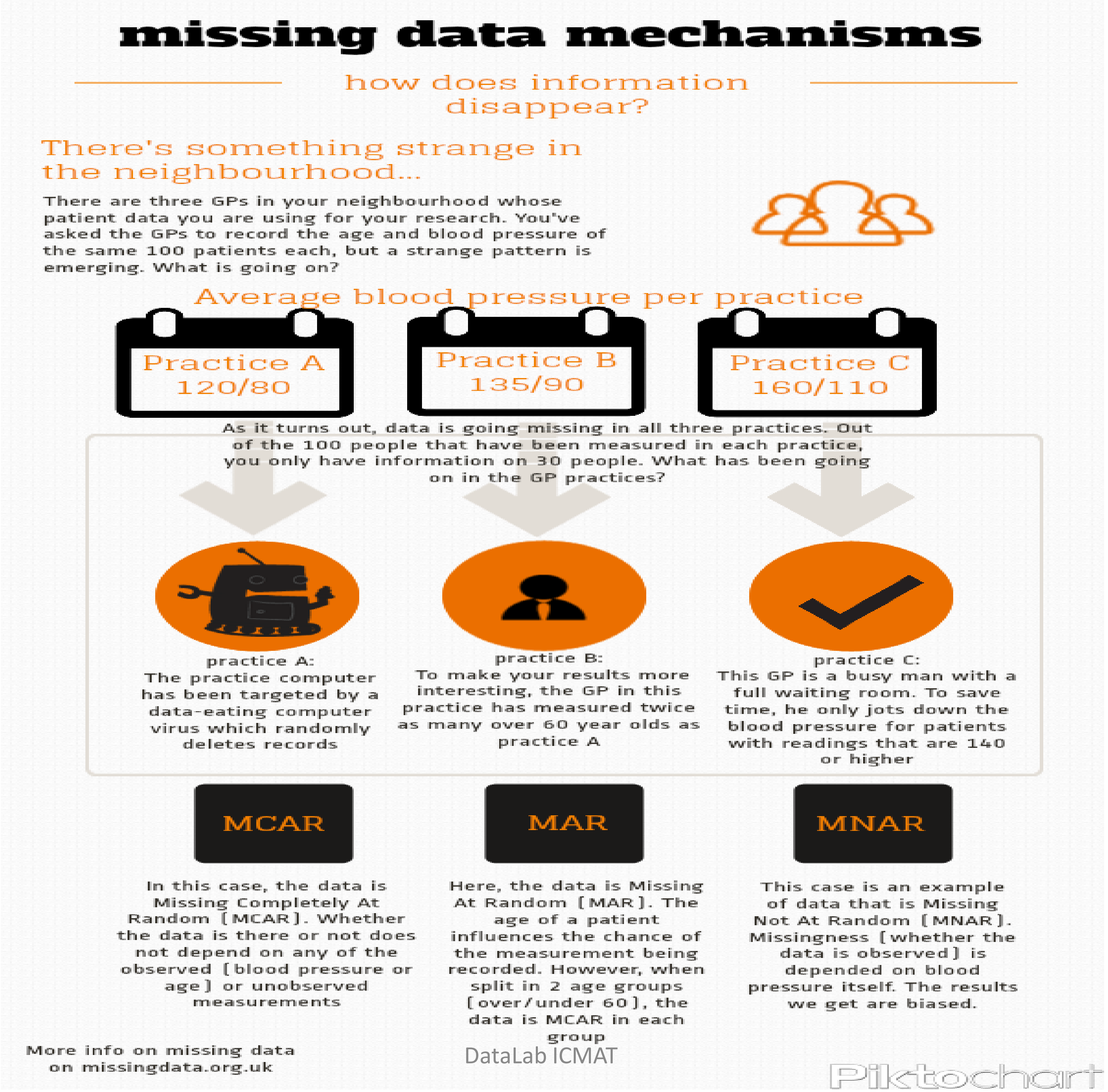
Different types of data

- Not always trivial to transform data in numerical and/or categorical variables
- Extra preprocessing required
- Examples
  - Text (tweets, web pages,…) word2vec, bag-of-words, n-grams
  - Images: RGB values of pixels, grey intensities
  - Audio: Fourier transform, MFCC (Mel Frequency Cepstral coeffs)
  - Video: sequences of frames
  - SMILE codes in chemoinformatics
  - Facebook likes

Missing values

- Examples: in a medical review a patient lacks value in Medicines (none or forgotten by doctor?); FB likes of user

  (0 likes or no likes?)
- Underlying mechanism of missing values? Examples next slide
- Potentially coded differently (in R NA)



missing data mechanisms

how does information disappear?

There's something strange in the neighbourhood...

There are three GPs in your neighbourhood whose patient data you are using for your research. You've asked the GPs to record the age and blood pressure of the same 100 patients each, but a strange pattern is emerging. What is going on?

Average blood pressure per practice

Practice A	Practice B	Practice C
120/80	135/90	160/110

As it turns out, data is going missing in all three practices. Out of the 100 people that have been measured in each practice, you only have information on 30 people. What has been going on in the GP practices?

practice A:
The practice computer has been targeted by a data-eating computer virus which randomly deletes records

practice B:
To make your results more interesting, the GP in this practice has measured twice as many over 60 year olds as practice A

practice C:
This GP is a busy man with a full waiting room. To save time, he only jots down the blood pressure for patients with readings that are 140 or higher

MCAR

In this case, the data is Missing Completely At Random [MCAR]. Whether the data is there or not does not depend on any of the observed [blood pressure or age] or unobserved measurements

MAR

Here, the data is Missing At Random [MAR]. The age of a patient influences the chance of the measurement being recorded. However, when split in 2 age groups [over/under 60], the data is MCAR in each group

MNAR

This case is an example of data that is Missing Not At Random [MNAR]. Missingness [whether the data is observed] is depended on blood pressure itself. The results we get are biased.

More info on missing data on missingdata.org.uk

DataLab ICMAT

Treating missing values

- Ignore observations lacking some variables….
- Complete observation with mean, median (continuous), mode (categorical) from all observations with that variable or observations grouped according to some criteria
- Predictive models
  - R: mice
  - KNN
  - …

Extreme values

- Distinguish if mistake or valid. Case: Fraud detection
- Rules to identify them. In R, outliers
- May affect learning algos (although others are robust)
- Check they are not impossible values

  Example: In medical project an individual with BMI 50

Treating Extreme values

- Eliminate observation
- Assign as new value: lower or upper limit of standard values
- Assign NA and use imputation
- Robust models. Normal vs t
- Mixture models

Standardization

- Very different ranges frequently
  (salary vs age)
- This impacts many models/techniques: e.g. PCA
- Standardization also favours numerical stability
- Recall when interpreting

Treating standardization

- Mean 0, variance 1
- Scale to interval [-1,1]
- Robust standardization
- …

Categorical values

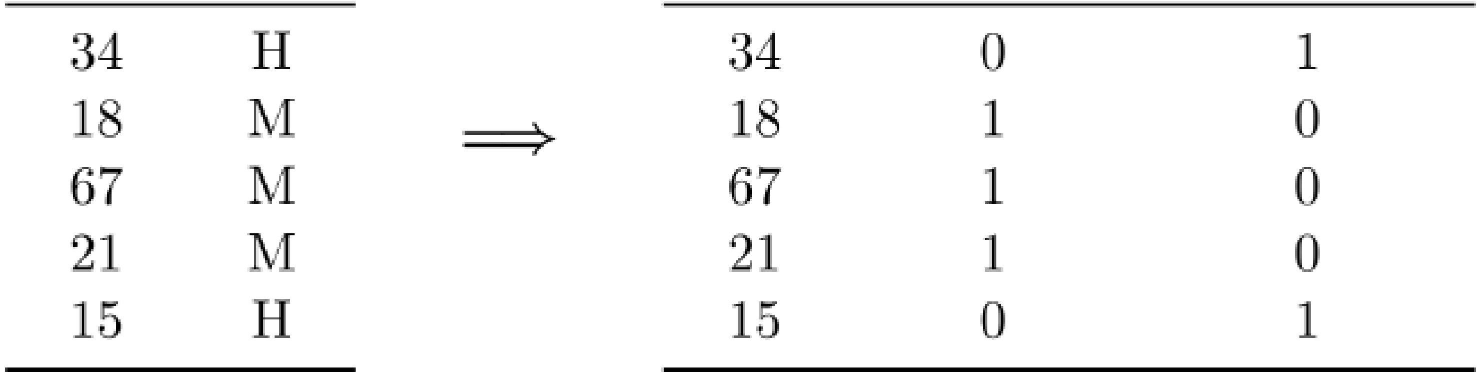
- Quite common

- Pose difficulties to some algos (e.g. linear regression)

- Convert them in numerical (but care with order induced…)

- Dummy encoding (or one hot encoding)

Dummy encoding



Age	Gender
34	H
18	M
67	M
21	M
15	H

⟹

Age	Is M?	Is H?
34	0	1
18	1	0
67	1	0
21	1	0
15	0	1

If categorical variable with p values, add p-1 new variables

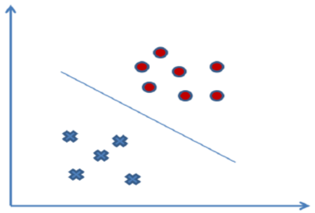
In R, factor

But see the lab

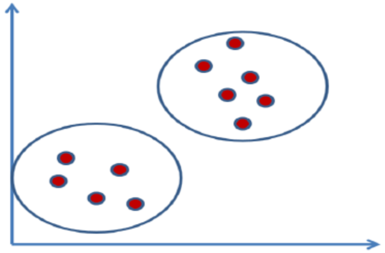
Summing up

Recap

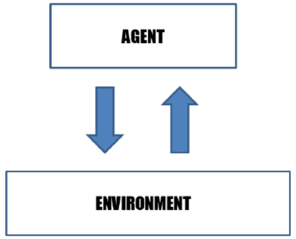
- Supervised



- Unsupervised



- Reinforced



- ML vs Bayes
- Challenges due to BD