Machine Learning
ML. 1. Intro

David Ríos Insua and Roi Naveiro

Objectives and schedule

A broad overview of Machine Learning

- Supervised learning
- Unsupervised learning
- Reinforcement learning

Machine learning

From Wikipedia

ML: the study of computer algos that improve automatically through experience. It is seen as a part of AI. ML algos build a model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to do so….

A subset of ML is closely related to computational statistics, which focuses on making predictions using computers; but not all ML is statistical learning.

The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning.

Data mining is a related field of study, focusing on exploratory data analysis through unsupervised learning.

In its application across business problems, machine learning is also referred to as predictive analytics.

Some ML examples. Red matters!!!

Uncertainty is almost ubiquitous in ML:

- Given a certain transaction, is it fraudulent or not? If fraudulent should I stop it?
- Given the monitoring trace of an Inet device, are we facing an attack? Should I stop operations?
- Does this medical image correspond to a person with a certain illness? Should I make further tests?
- A person with these FB likes will buy this type of beer? Should I send him my brand add?
- A person with these tweets is conservative? Should I send her Brexit propaganda?
- Robots (or ADS): If robot performs this, How will the user react? And the environment? Consequently, what should the robot do?

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 multple impacts

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

ML, Stats in modern times (Big Data)
Computer Age Statistical Inference

Volume. Space scalability

Variety. Text, images, sound, video,… video,….

Velocity. High frequency for data and decisions, time series, dynamic models. Time scalability

Create value by actually supporting decisions!!!

Core themes

- Supervised learning: Pairs input-output available

  Regression, Classification

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

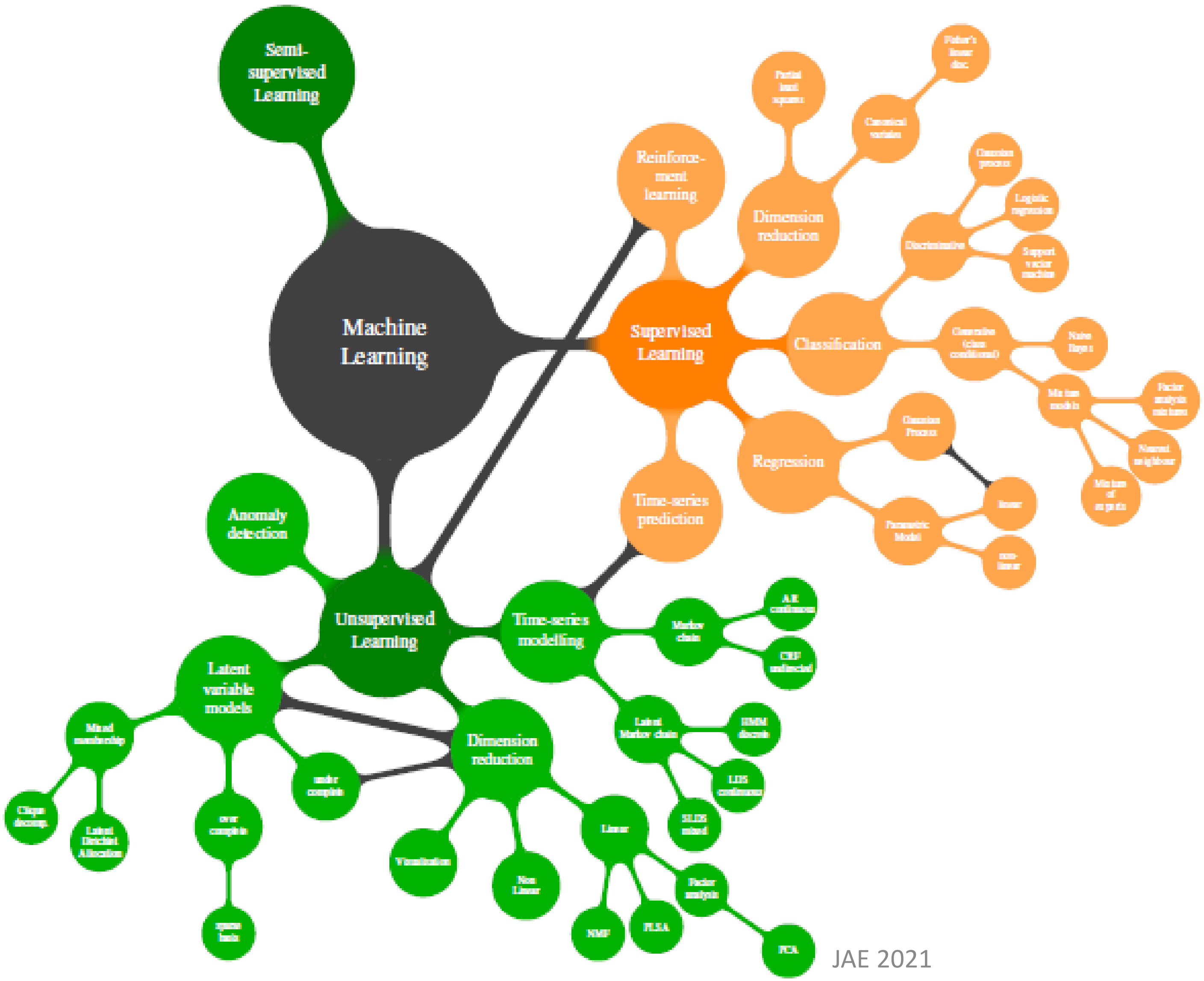
  Density estimation, clustering, outlier detection, Visualisation

  Case: Singular community detection
- Reinforcement learning: Decisions impacting outputs on-the-fly

  Markov decision processes

  Case: Autonomous driving systems

With somewhat blurry borders….



Case: Lookalike modeling

A brief KitKat

vs

A basic example

Consider a cyberattack recovery protocol for SMEs. Introduce a process. Want to assess it. E.g to compare it with another

Test it against 12 attacks, effective 9 times (e.g., system up in less than 1 hour)

A basic example. Model

- Number X of successes in n trials (ii)
- Success probability in one trial
- Distribution of number of successes
- For X=9,

θ_1

$X \mid \theta_1 \sim \text{Bin}(12, \theta_1)$

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

A basic example. MLE

Likelihood

$$\ell(\theta_1) \propto \theta_1^9 (1-\theta_1)^3$$

Log-likelihood

$$L(\theta_1) = \log(\ell(\theta_1)) = 9 \log \theta_1 + 3 \log(1-\theta_1)$$

Maximise likelihood or maximize log likelihood . MLE

In this case, MLE is

$$L'(\theta_1) = 0 \Rightarrow \hat{\theta}_1 = \frac{9}{12} = .75$$

Defects?

For future observations (e.g. 4 succeses in next 7 trials)

$$Pr(Y=4 \mid \hat{\theta}_1, 7) = \binom{7}{4} .75^4 \; .25^3$$

A basic example. Bayes

Prior, e.g.

$$\pi(\theta_1) = 1$$

$$\pi(\theta_1|9) \propto 1 \times \theta_1^9 (1-\theta_1)^3 \sim \mathcal{B}e(10,4)$$

Posterior

- Posterior mean $\frac{10}{14}$
- Posterior mode (MAP) $\frac{9}{12}$

Predictive

$$\Pr(Y=4|9) = \int \binom{7}{4} \theta_1^4 (1-\theta_1)^3 \, \pi(\theta_1|9) \, d\theta_1$$

$$= \frac{\binom{7}{4}\binom{13}{3}}{\binom{20}{12}}$$

And so…

We end up using this to make decisions. Which protocol to implement?

How would you choose between two protocols?

Intro to Supervised Learning

SL: ingredients

- Data available: examples, samples, instances,…
- Several observed variables: predictors, attributes, features, covariates, explanatory variables, independent variables,…
- Some of special interest: response(s), dependent variable(s), target(s), output(s), label(s),…

SL: types of problems

1. Regression, response variable is continuous
2. Classification, response variable is discrete
3. Other:
   – Mixed (some continuous, some discrete)
   – Discrete but ordered
   – …

Predictors

$(x_1, \dots, x_p)$

Dependent variable

y

Some relation

$$Y = f(X) + \epsilon$$

Systematic info

Random term. Zero mean, Indep of x

Inference vs Prediction

How do we estimate f?

Training data $\{(x_1, y_1), \dots, (x_n, y_n)\}$

$\hat{f}(x_i) \approx y_i$

For any observable $\hat{f}(x_0) \approx y_0$

Parametric, e.g. $f(x) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$

$Y \approx \beta_0 + \dots + \beta_p x_p.$

$\hat{f}$ vs $(\beta_0, \dots, \beta_p)$

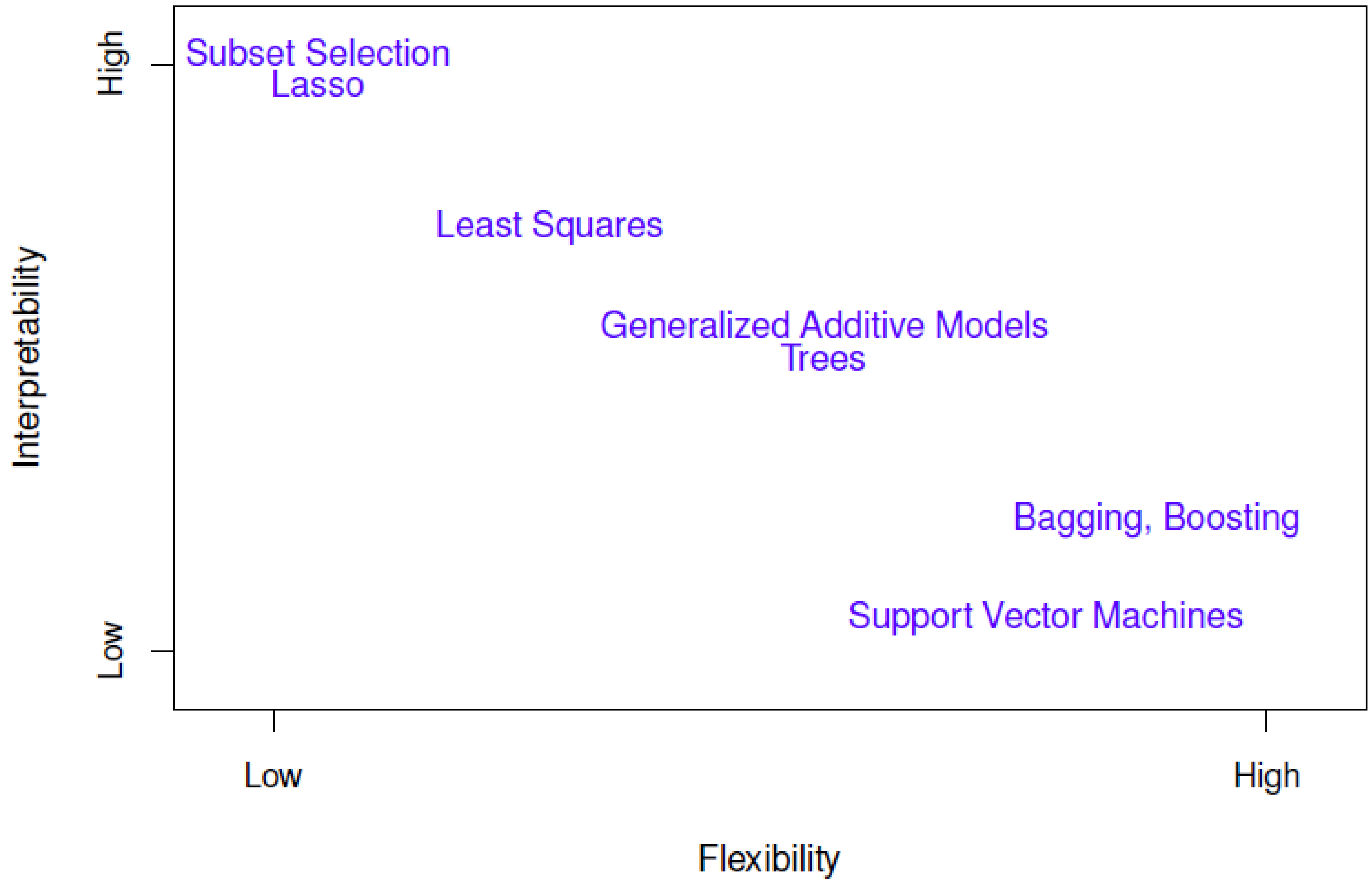
Flexibility and overfitting

Non-parametric

Wider range, much larger #observations

Flexibility vs Interpretability

(somewhat old figure from ISLR)



Deep Models!!!!

Rudin's paper!!!!

Assessing accuracy

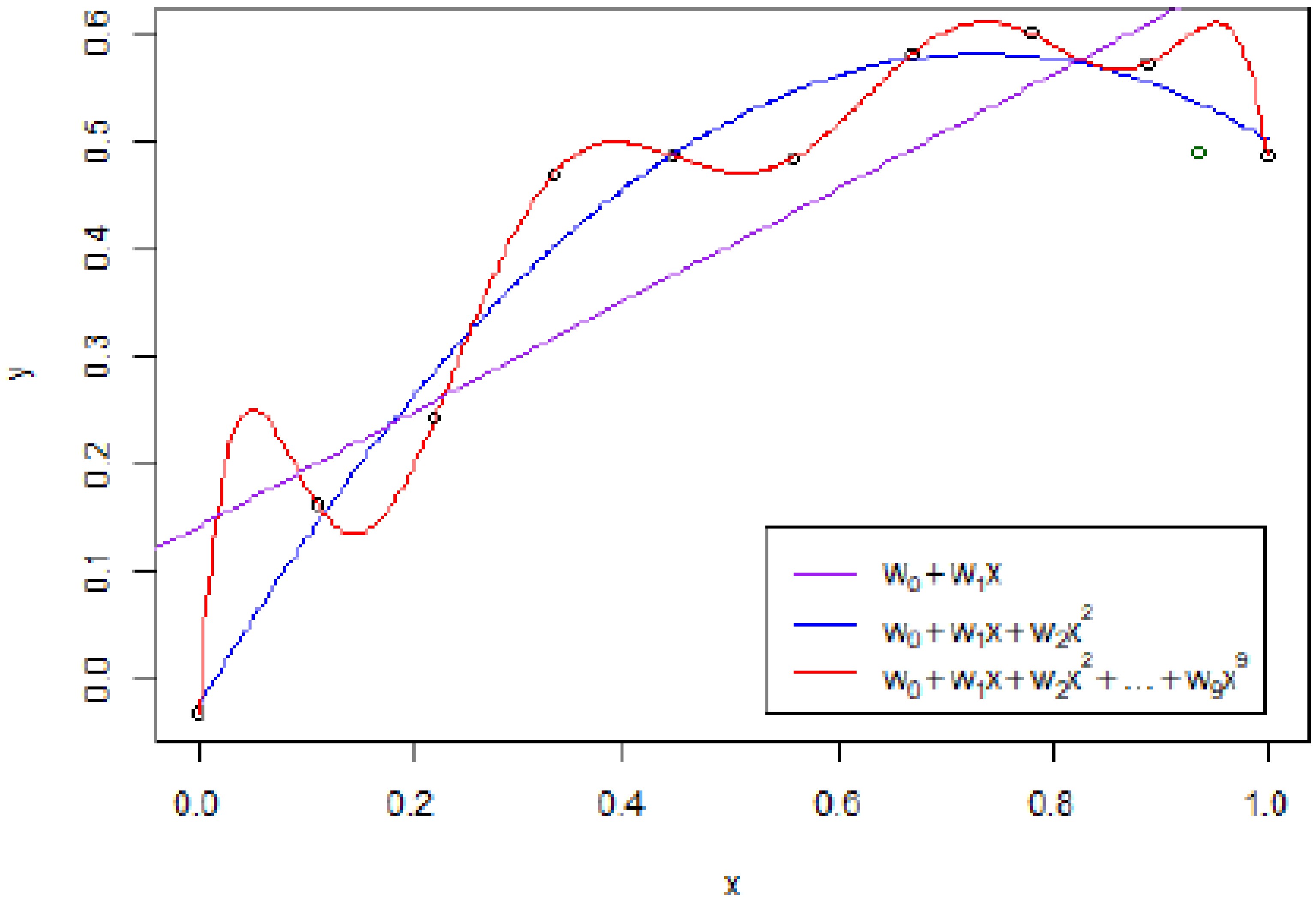
No free lunches

Quality of fit, e.g. $MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{f}(x_i))^2$

Training MSE vs Test MSE

Not $\hat{f}(x_i) \approx y_i$ but $\hat{f}(x_0) \approx y_0$

Small $Ave(y_0 - \hat{f}(x_0))^2$



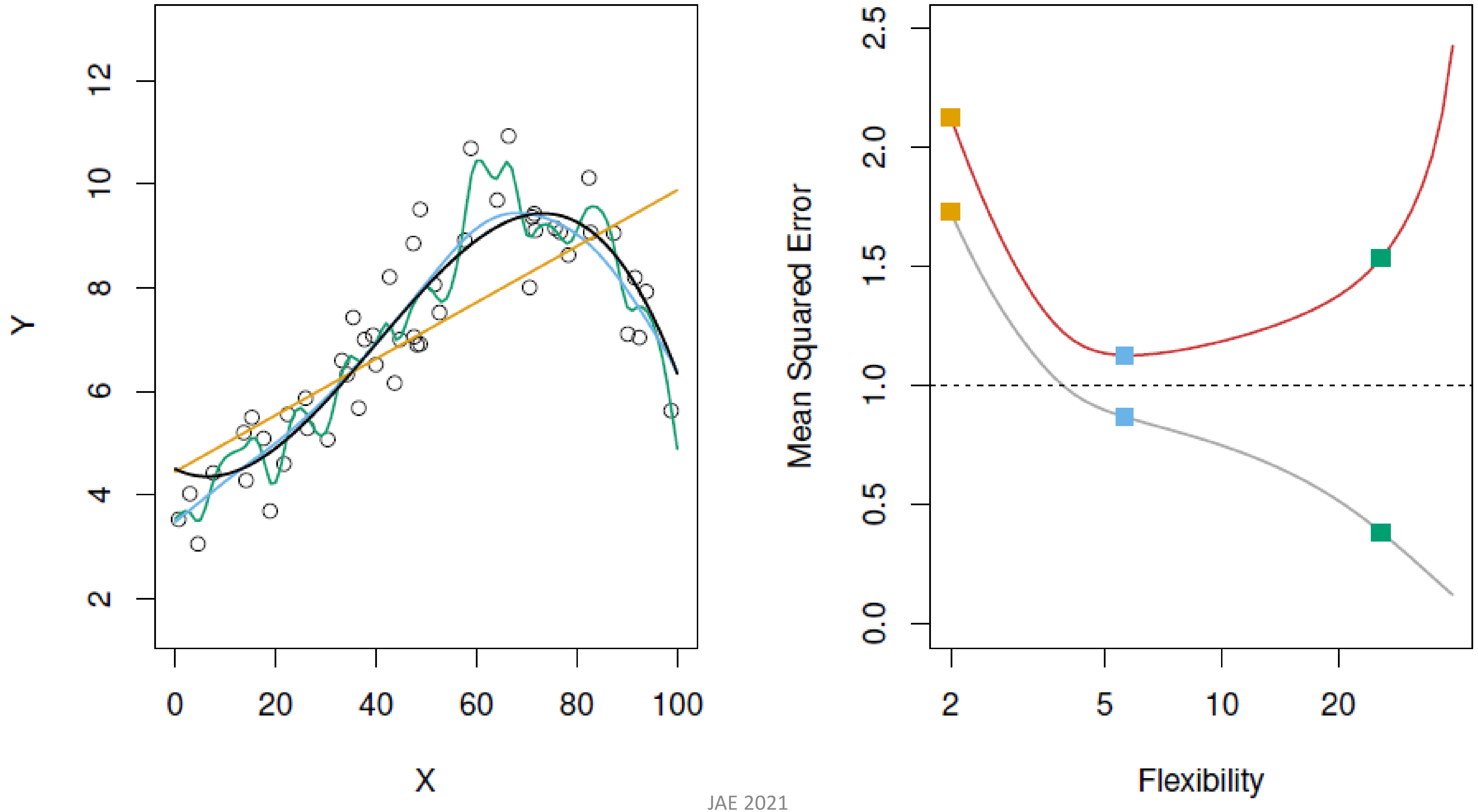
0.0
0.1
0.2
0.3
0.4
0.5
0.6
y
0.0
0.2
0.4
0.6
0.8
1.0
x
$w_0 + w_1x$
$w_0 + w_1x + w_2x^2$
$w_0 + w_1x + w_2x^2 + \ldots + w_9x^9$

Model selection

- Compute empirical risk over training set
- May be reduced almost arbitrarily increasing model complexity

e.g. based on polynomials

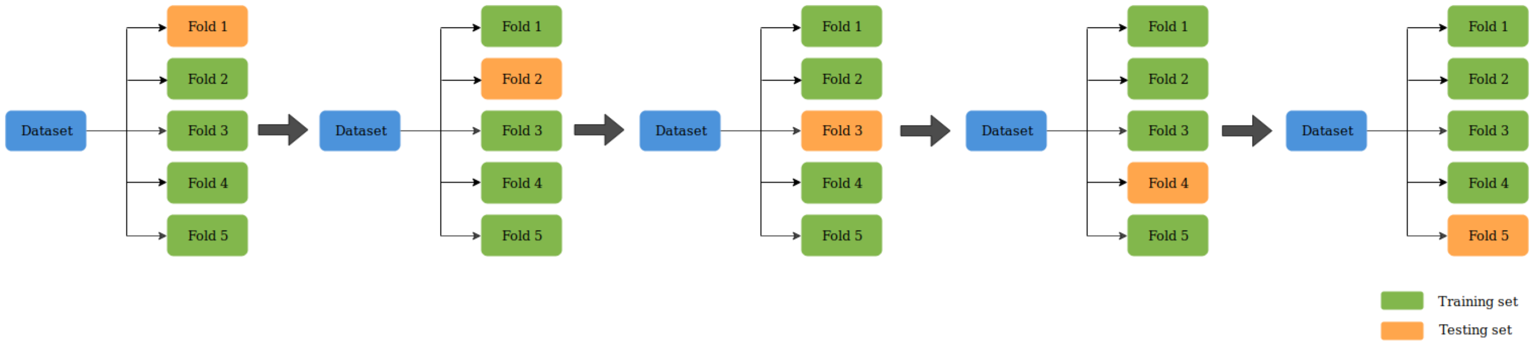
- Generalisation error over observations not used to train the model (cannabinoid project)
- If no test set available, split data in two sets:
  - Training set
  - Test set



Y
2
4
6
8
10
12
0
20
40
60
80
100
X
Mean Squared Error
0.0
0.5
1.0
1.5
2.0
2.5
2
5
10
20
Flexibility

Cross validation

- Hyperparameter choice to control model complexity
- Choose a third set for validation to select and compare models
- If data not plentiful, divide set in k partitions
- Use k-1 to train and the other to test: k models
- Cross validation error



- If k=n leave-one-out cross validation

Bias-variance tradeoff

- Assume model

$$Y = f(X) + \varepsilon \qquad E(\varepsilon) = 0 \qquad Var(\varepsilon) = \sigma^2$$

- Exp. Pred. error (under quad. Loss)

$$EPE = E\left(Y - \hat{f}(x)\right)^2$$

- Decomposed as

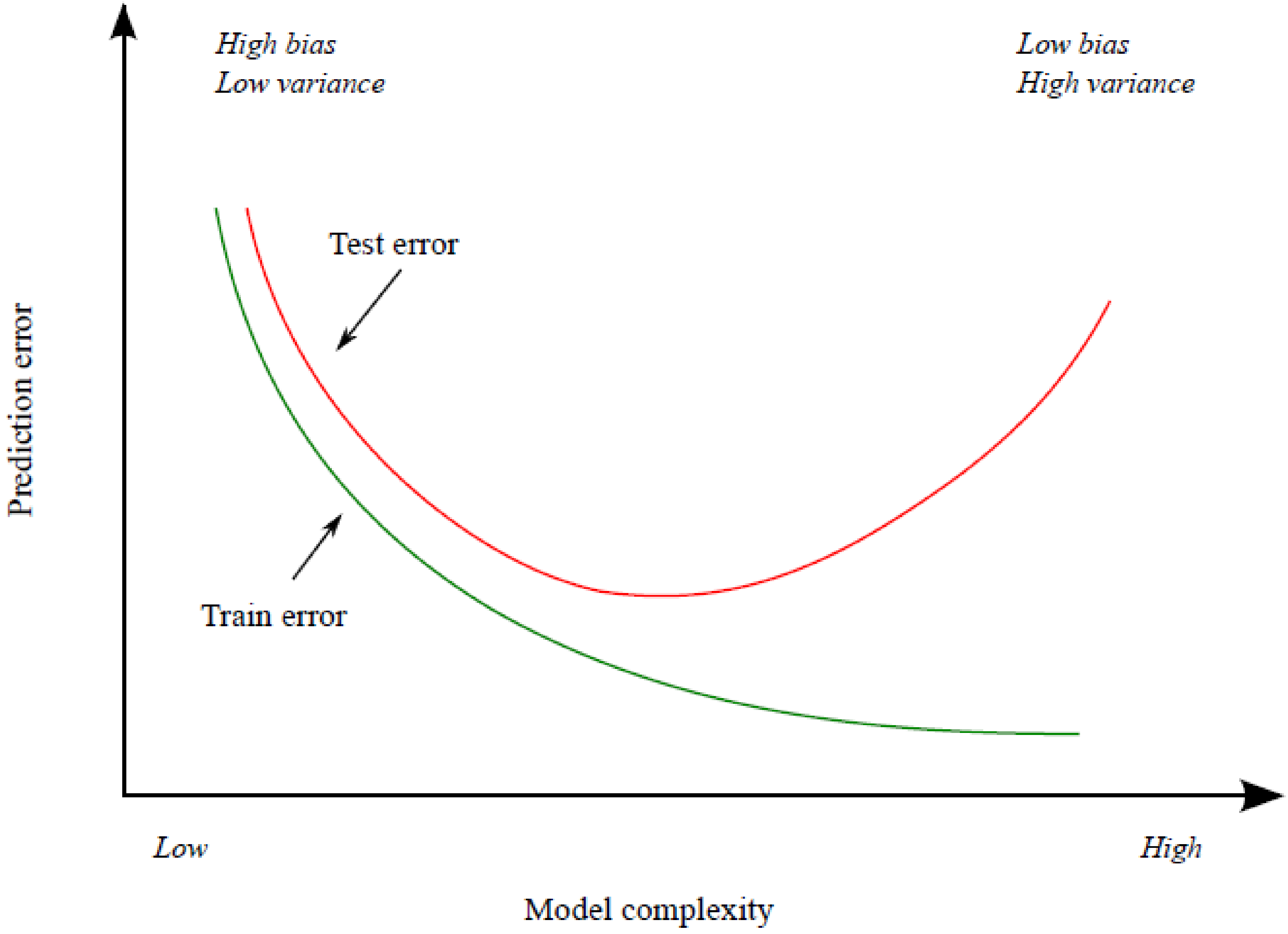
$$EPE = E\left(\hat{f}(x) - f(x)\right)^2 \quad \text{BIAS}^2$$
$$+$$
$$E\left(\hat{f}(x) - E(\hat{f}(x))\right)^2 \quad \text{VAR}$$
$$+$$
$$\sigma^2 \quad \text{NOISE}$$

Variance. How approximation changes if a different training set used

Bias. Error due to using much simpler model

Generally, more flexible method: variance increases, bias decreases

Bias-variance tradeoff


High bias
Low variance
Low bias
High variance
Test error
Train error
Prediction error
Low
High
Model complexity

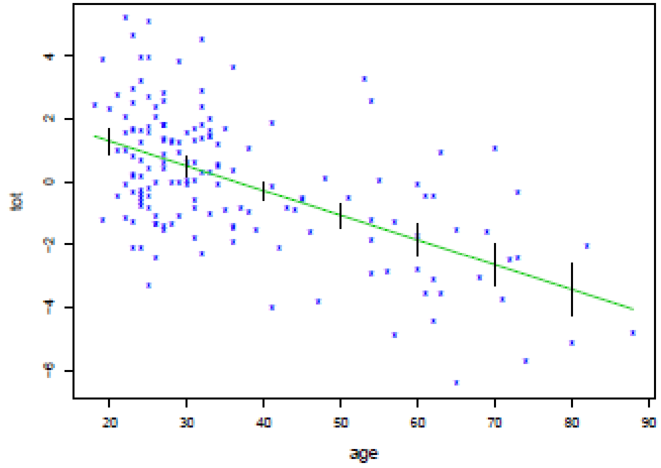
Regularisation

- Aim: reduce variance in exchange of a small bias
- Introduce sparsity
- Limit model complexity by adding a regularisation term

$$\min \sum_{i=1}^{n} (y_i - \beta x_i)^2 + \boxed{\lambda \sum_j \beta_j^L}$$

Linear regression model. A typical example

Consider a study of kidney function. Data represent (x=age of person, y=tot, a composite measure of the overall function). Kidney function declines with age. We need to provide additional information concerning decline rate. This is important in managing kidney transplants.



Check
https://en.wikipedia.org/wiki/Simple_linear_regression

Linear regression

Data structure. Response
Explanatory variables

Y

$(x_1, \dots, x_n)$

Model

$$y_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi}, \quad i = 1, \dots, n$$

$$\varepsilon_i \sim N(0, \sigma^2) \text{ IND.}$$

Likelihood

$$\theta = (\beta_0, \dots, \beta_p, \sigma)$$

$$p(\theta | \underline{x}) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \prod_{i=1}^{n} \exp\left(-\frac{1}{2}\left(\frac{y_i - \beta x_i}{\sigma}\right)^2\right)$$

Log-likelihood

$$\max -\frac{1}{2} \sum_{i=1}^{n} \left(\frac{y_i - \beta x_i}{\sigma}\right)^2 \quad \dots$$

$$\hat{\beta} = (X^t X)^{-1} X^t y$$

MLE

$$s^2 = \frac{1}{(n-p)} (y - X\beta)^t (y - X\beta)$$

Linear regression

If n or n and p large

- COMPUTE $X = QR$ $Q_{n \times p}$ ORTH. COLUMNS, $R_{p \times p}$ UPPER TRIANG.

$$\left[(X^T X)^{-1} = (R^T Q^T Q R)^{-1} = (R^T R)^{-1} = R^{-1} (R^{-1})^T\right]$$

- COMPUTE R^{-1}
- SOLVE $R\hat{\beta} = Q^T y$

$$\left[\hat{\beta} = (X^T X)^{-1} X^T y = (R^T Q^T Q R)^{-1} R^T Q^T y = (R^T R)^{-1} R^T Q^T y = R^{-1} Q^T y\right]$$

Linear regression with regulariser

If p large (much larger than n)

$$\min \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda \sum \beta_i^2$$

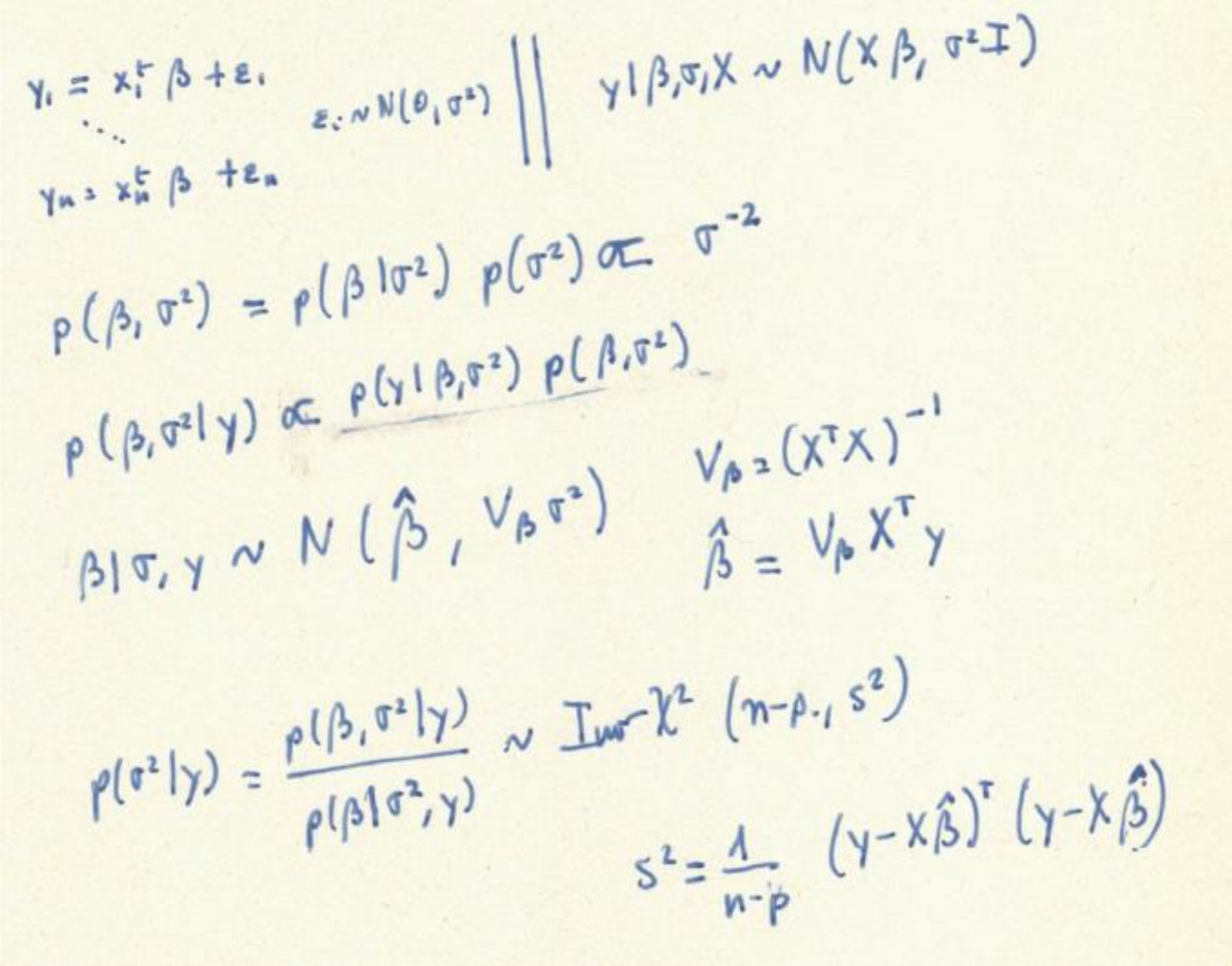
$$\min \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda \left|\sum \beta_i\right|$$

Bayesian inference with linear regression model

Model

Standard noninformative prior

Posterior



$$y_1 = x_1^t \beta + \varepsilon_1 \quad \ldots \quad y_n = x_n^t \beta + \varepsilon_n \qquad \varepsilon_i \sim N(0, \sigma^2) \qquad || \qquad y \mid \beta, \sigma, X \sim N(X\beta, \sigma^2 I)$$

$$p(\beta, \sigma^2) = p(\beta \mid \sigma^2)\, p(\sigma^2) \propto \sigma^{-2}$$

$$p(\beta, \sigma^2 \mid y) \propto p(y \mid \beta, \sigma^2)\, p(\beta, \sigma^2)$$

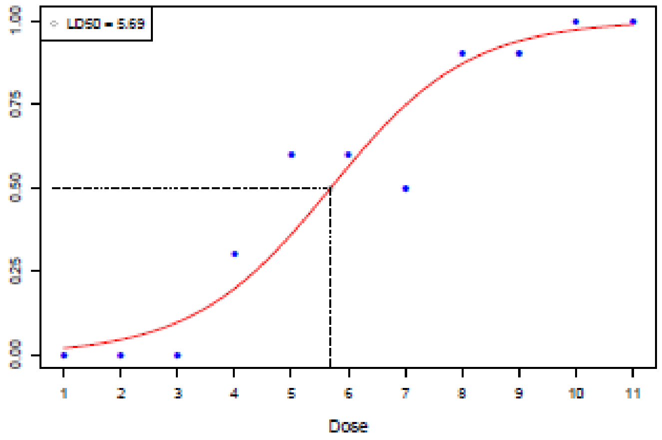
$$\beta \mid \sigma, y \sim N(\hat{\beta}, V_\beta \sigma^2) \qquad V_\beta = (X^T X)^{-1} \qquad \hat{\beta} = V_\beta X^T y$$

$$p(\sigma^2 \mid y) = \frac{p(\beta, \sigma^2 \mid y)}{p(\beta \mid \sigma^2, y)} \sim \text{Inv-}\chi^2(n-p, s^2)$$

$$s^2 = \frac{1}{n-p} (y - X\hat{\beta})^T (y - X\hat{\beta})$$

Logistic regression. A typical example

A new anti-cancer drug is being developed. Before human testing can begin, animal studies are needed to determine safe dosages. A bioassay or dose-response experiment is carried out: 11 groups of 10 mice are treated with an increasing dose of drug and the proportion of deaths are observed.



Check
https://en.wikipedia.org/wiki/Logistic_regression

Intro to Unsupervised Learning

Elements of unsupervised learning

Given

- Input space — $x \in \mathcal{X}$
- Training set — $S = \{x_i\}_{i=1}^{N}$

Objective

- Learn model
- Infer some property — $p(x)$
- Sample

Taxonomy of unsupervised learning algos

- Density estimation
- Manifold learning: PCA, non-linear PCA, …
- Finding modes and groups: cluster analysis, mixture models,…
- Sampling: GANs, Autoencoders, Variational autoencoders,…

Challenges in unsupervised learning

- High dimension of feature space
- Properties of interest more complex than parameter estimation
- No direct error quantification measure

Paradigm: Principal component analysis (PCA)

Two views

- Orthogonal projection to lower dimension space to maximize variance
- Linear projection minimizing average projection cost= average quadratic distance between data and projections

Applications

- Dimension reduction
- Compression
- Visualization
- Extraction of predictor. PC Regression
- ….

PCA: Maximum variance

Given

$$x_i \in \mathbb{R}^D \quad , \quad i = 1, \dots, n$$

Find linear projection to space of smaller dimension maximizing variance of projected data

$$\Pi : \mathbb{R}^D \to \mathbb{R}^M \quad , \quad M < D$$

PCA: Maximum variance

- 1 dimensional projection — $M = 1,\ \mathbb{R}$
- Projection defined by — $u_1 \in \mathbb{R}^D \quad (u_1^t u_1 = 1)$
- Projection is — $u_1^t x$
- Mean of projected data — $\frac{1}{n}\sum_{i=1}^{n} u_1^t x = u_1^t \bar{x}$
- Variance of projected data — $\frac{1}{n}\sum_{i=1}^{n} \left(u_1^t x_i - u_1^t \bar{x}\right)^2 = u_1^t S u_1$

PCA: Maximum variance

- Problem to be solved

$$\max u_1^T S u_1$$
$$\text{s.t. } u_1^T u_1 = 1$$

- Lagrangian formulation

$$\max_{u_1} u_1^T S u_1 + \lambda_1 (u_1^T u_1 - 1)$$

- Solution

$$S u_1 = \lambda_1 u_1$$
$$u_1^T S u_1 = \lambda_1$$

- Projection is eigenvector associated with first eigenvalue!!!

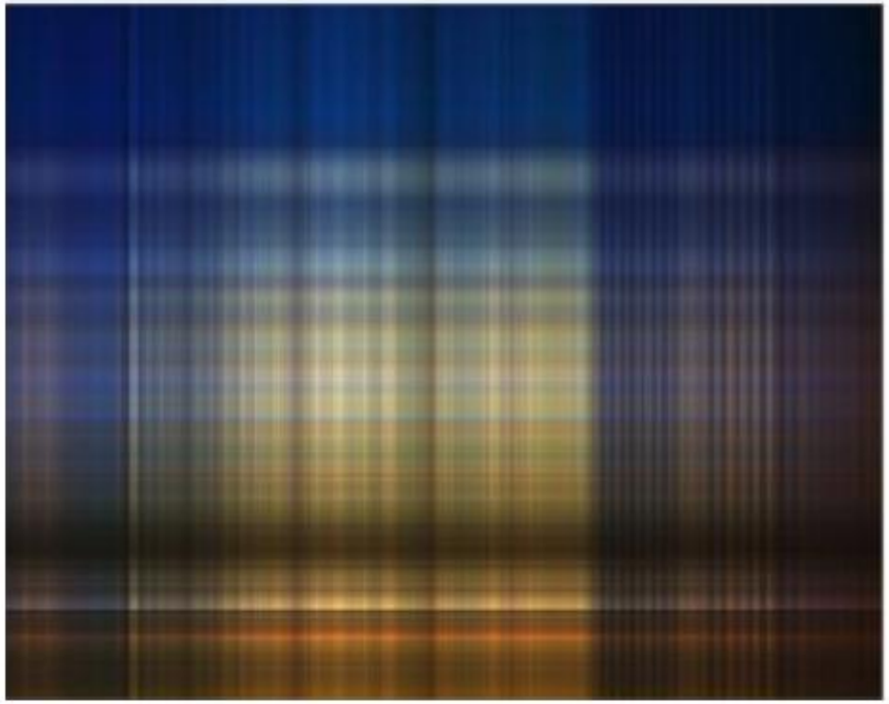
(and so on)

Data compression

Projecting each D-dimension point to M

$$\hat{x}_i = \bar{x} + \sum_{j=1}^{M} (x_i^t - \bar{x} u_j) u_j$$

- M = 1



- M = 3

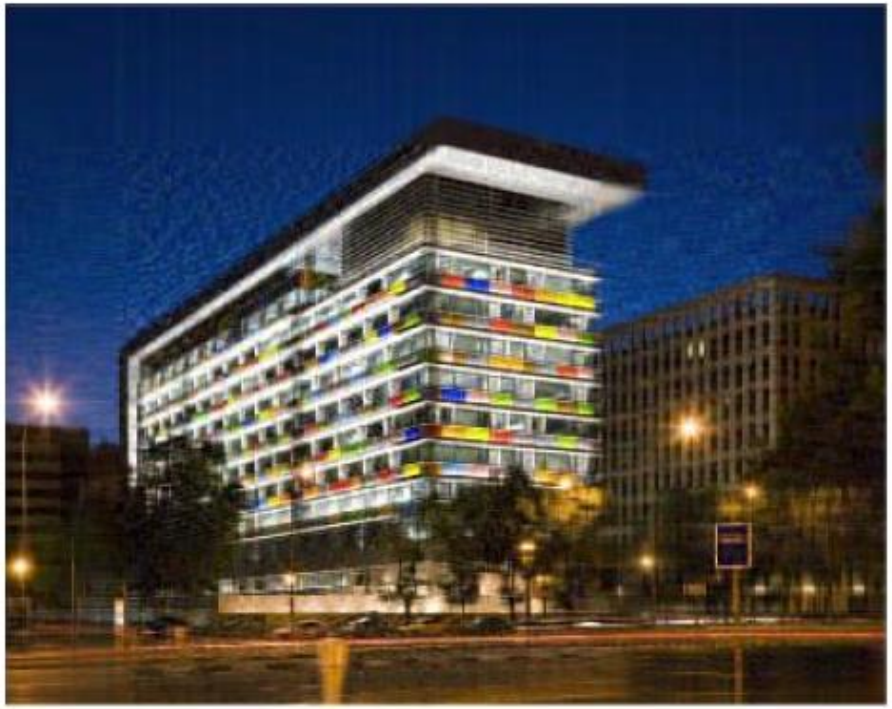
Data compression

- M = 10

- M = 20

Data compression

- M = 50



- M = 200

Implementation challenges

- High dimensionality. What if D>>n
- n points in space of dimension D
- Computational complexity of computing eigenvectors

Reinforcement learning

RL: features

- Learning by interaction with environment
  - ‘Cause-Effect’ relations
  - Consequences of actions
  - What to do to achieve goals
- Goal directed learning: what to do to maximize a reward
  - Discover actions that yield most reward by trying them (trial and error search)
  - Actions affect not only immediate reward but also affect environment (delayed reward)

RL features

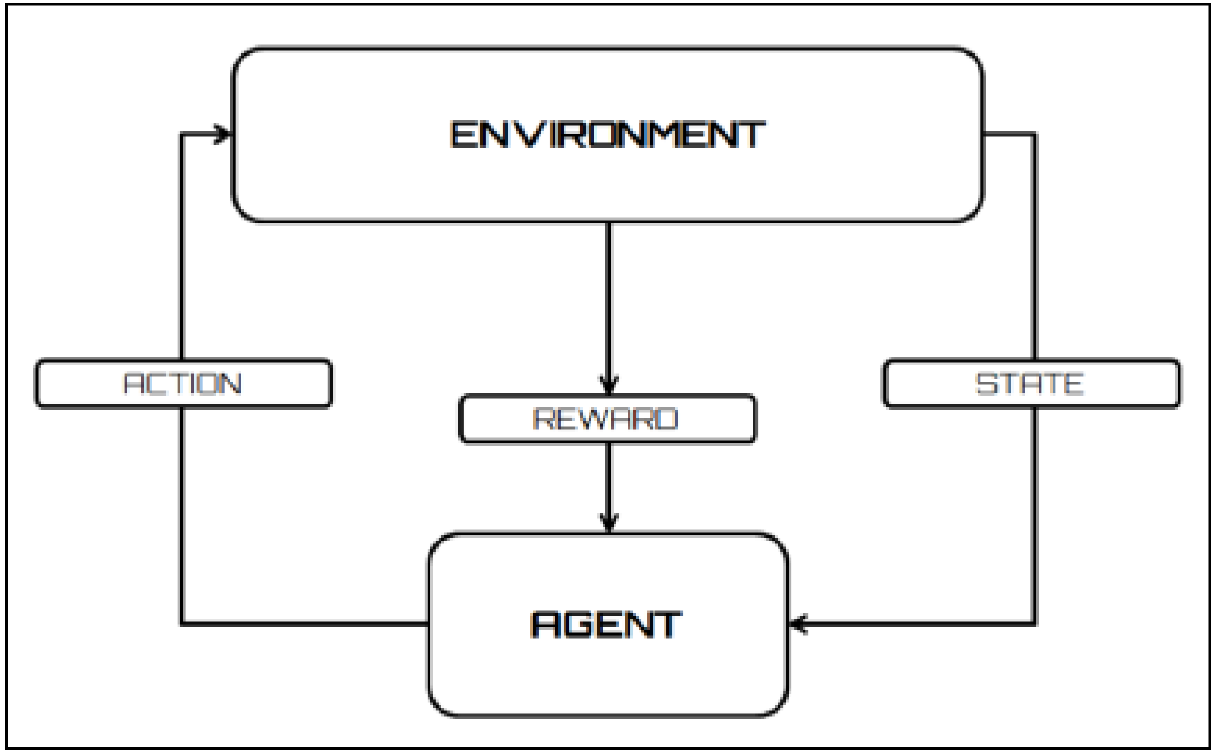
- Optimal control of incompletely known Markov decision processes
  - Schemes for sense-act-respond
  - Exploration (collect more info)-exploitation (best action)
  - Uncertainty about evolution of environment and rewards achieved
  - Sequential learning

RL elements

- Agent
- Environment with states
- Policy
- Reward signal
- Value function
- Model of environment

- Model based methods
- Model free methods

RL elements



RL Elements: MDPs

- States
- Actions
- Transition
- Reward
- History
- LT Expected discounted utility
- Policy

$$s \in S$$

$$a \in A$$

$$T : S \times A \to \Delta(S)$$

$$R : S \times A \to \Delta(R)$$

$$\tau = (s_0, a_0, s_1, a_1, \dots)$$

$$E_\tau \left(\sum_{t=0}^{\infty} \gamma^t R(a_t, s_t) \right)$$

$$\pi : S \to \Delta(A)$$

RL elements: Q-learning

$$Q(s,a) := (1-\alpha)\, Q(s,a) + \alpha\left(r(s,a) + \gamma \max_{a'} Q(s',a')\right)$$

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

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

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

MLE

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

Predictions

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

Recap: BML

Prior $\quad f(\theta)$

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

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

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

Recap: RegML

$$\max_{} h(\theta) + \lambda g(\theta)$$

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

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

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

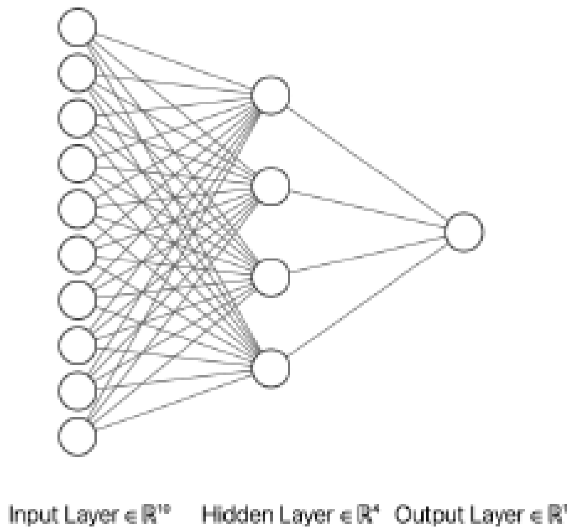
Optimisation. Recall

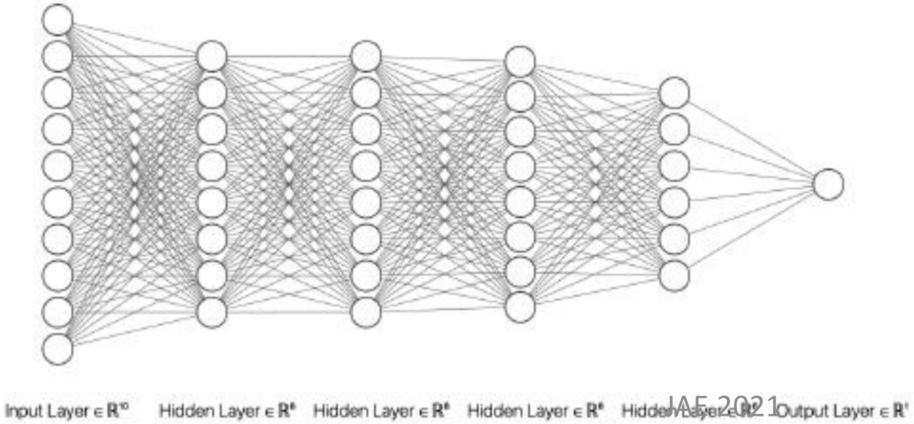
Check e.g. Goodfellow et al Ch. 4 (+8)

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

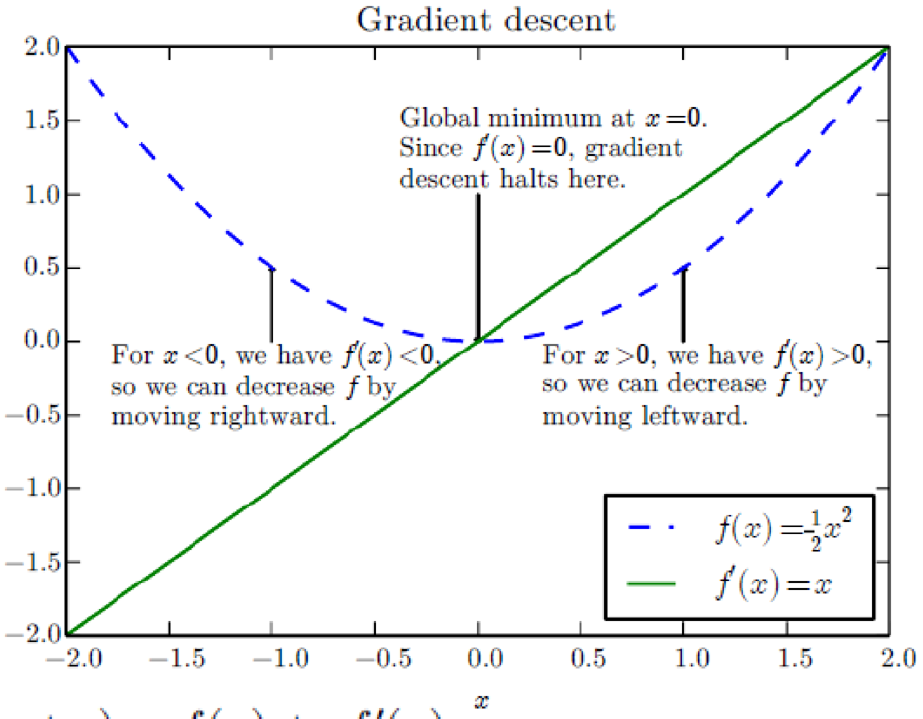
$$y_j = \sum_{i=1}^{m} \beta_i \psi(x_k \omega_i) + \varepsilon_j$$

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





Optimization: Using gradient info



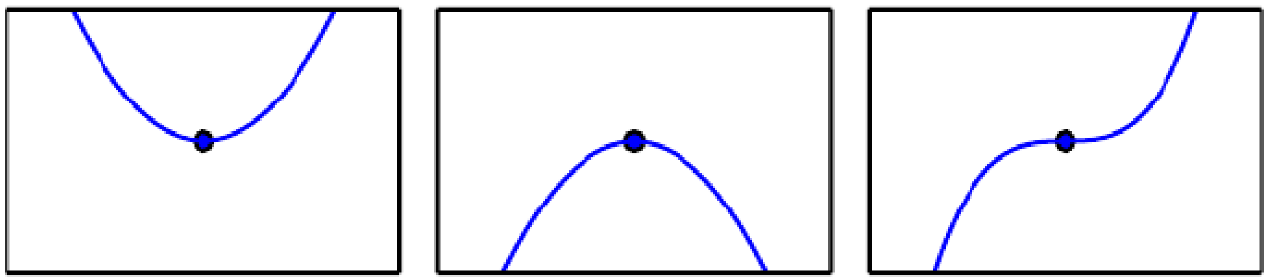
$$f(x+\epsilon) \approx f(x) + \epsilon f'(x)$$

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

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

Learning rate

$f'(x) = 0$ Stationary point

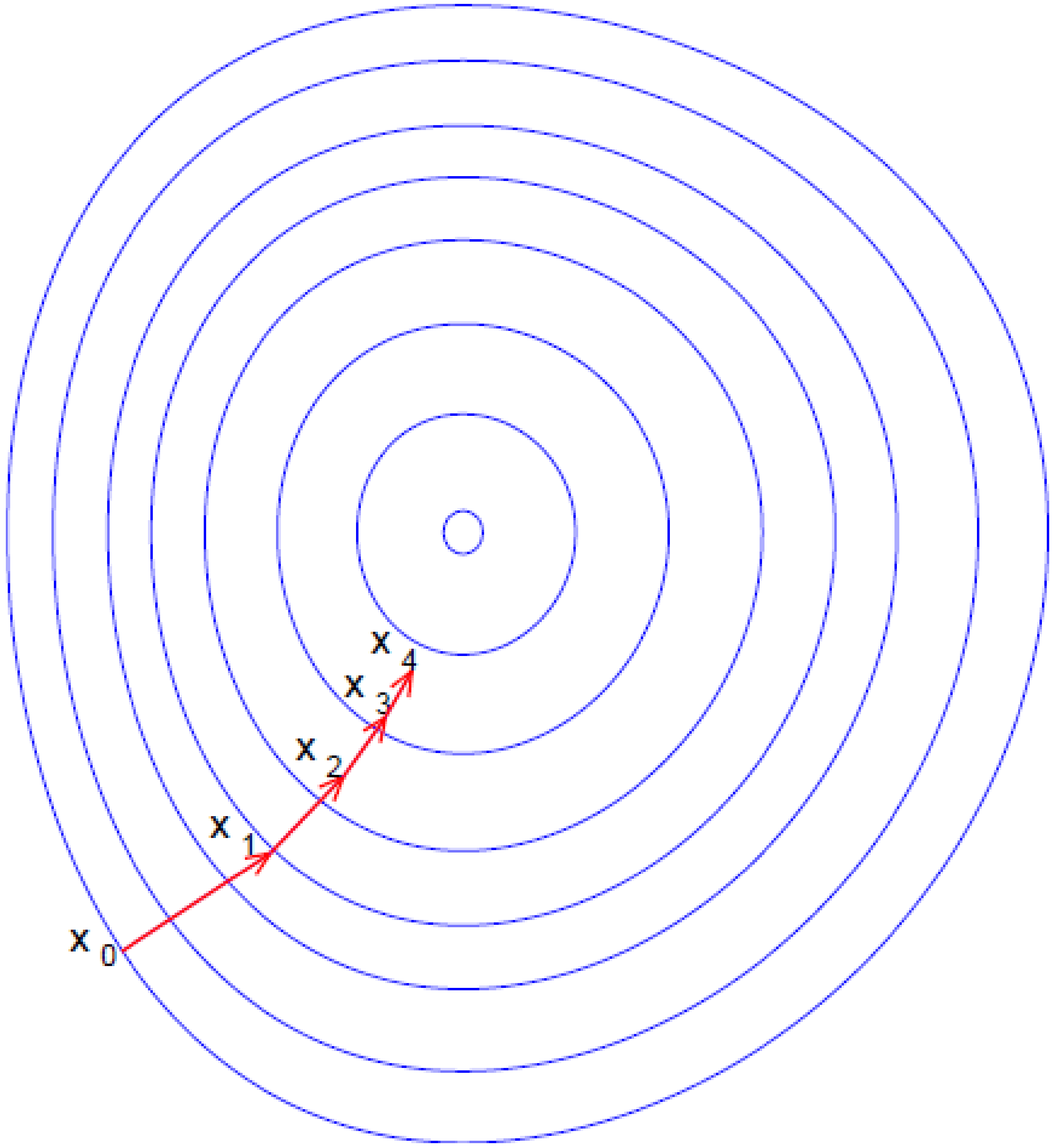


Until stopping condition

Gradient descent

- Fixed and small rate
- Line search

Grad estimation. Backprop for NNs



JAE 2021

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?

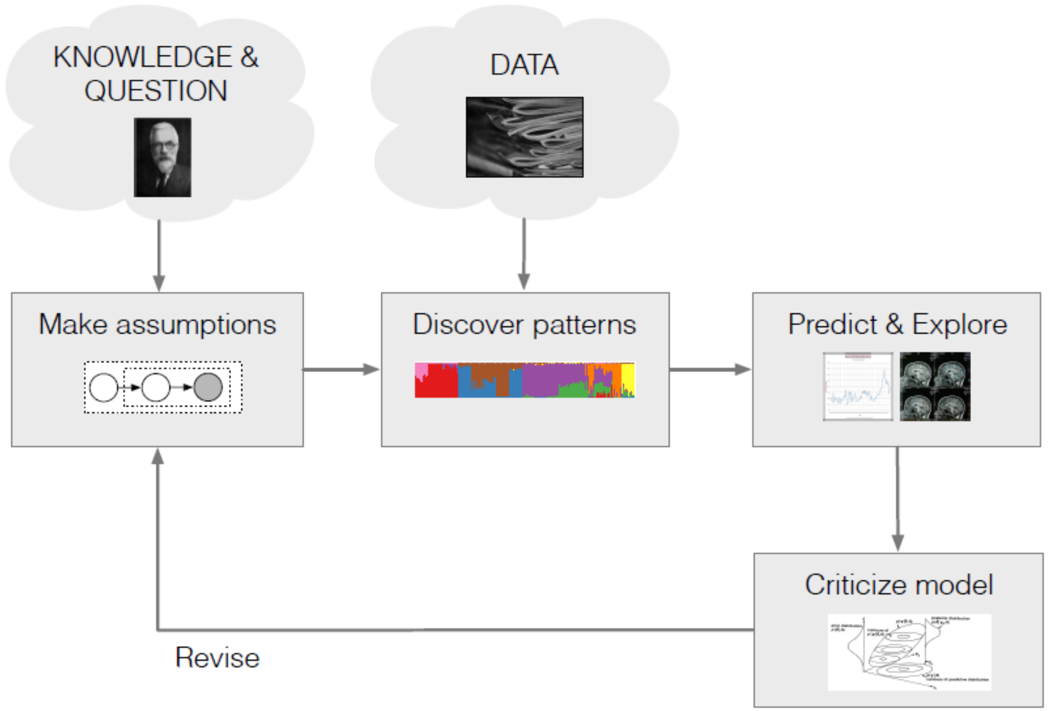
Stochastic gradient descent….

Optimization in ML

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

Data flow in machine learning

Broad learning scheme

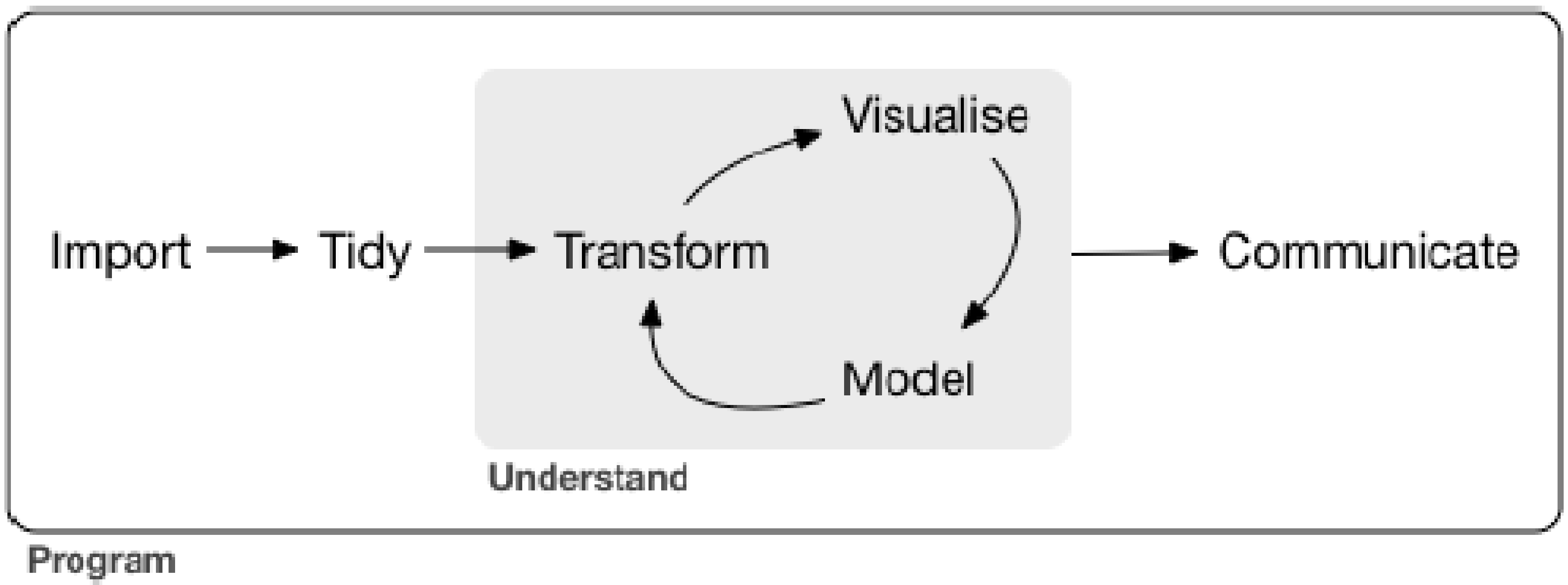


Hypothesis Computing Application

Inference

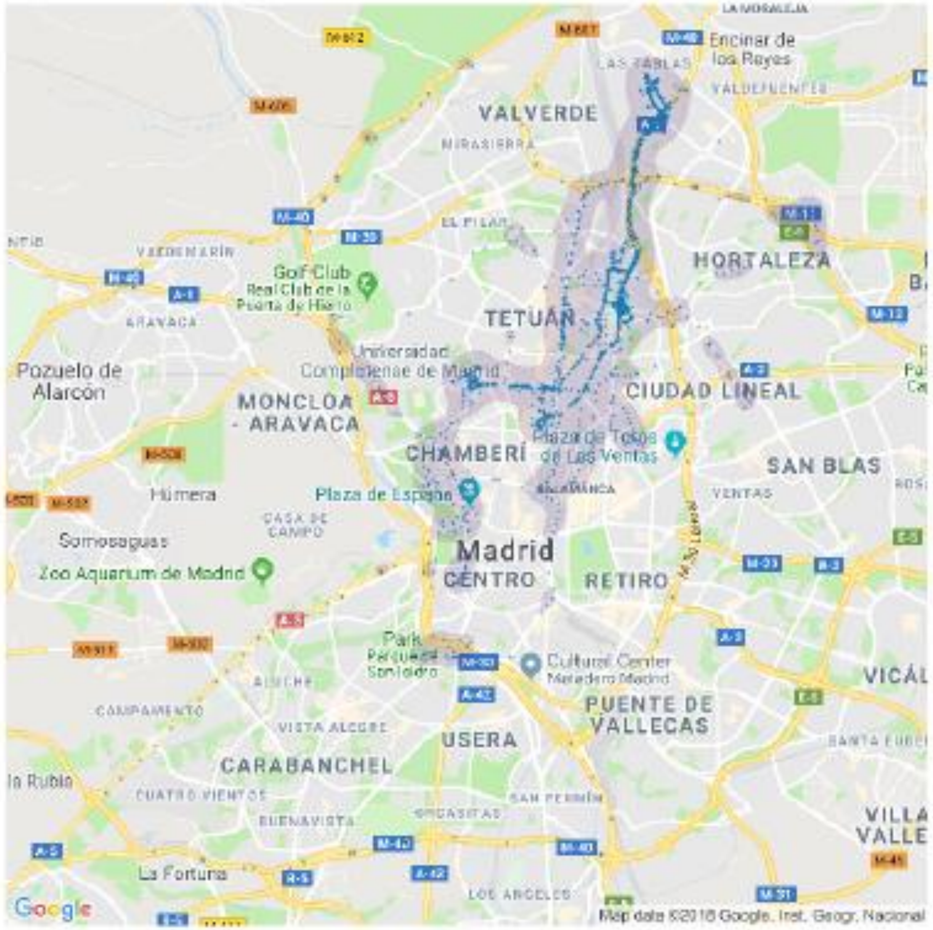
What does my model say about data?

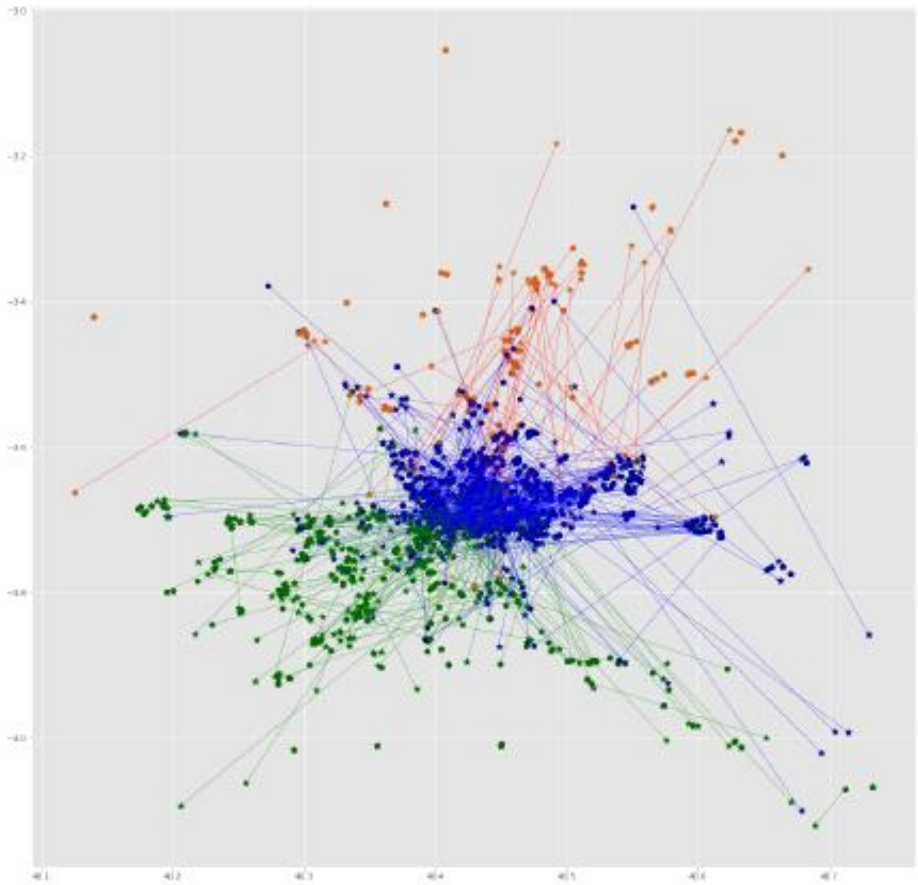
General, Scalable



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

ML and BD


VALVERDE
MIRASIERRA
LAS TABLAS
Encinar de los Reyes
VALDEFUENTES
EL PILAR
HORTALEZA
TETUÁN
Golf Club Real Club de la Puerta de Hierro
ARAVACA
Pozuelo de Alarcón
Universidad Complutense de Madrid
MONCLOA - ARAVACA
CIUDAD LINEAL
CHAMBERÍ
Plaza de Toros de Las Ventas
SAN BLAS
Húmera
Plaza de España
VENTAS
Somosaguas
CASA DE CAMPO
Madrid
CENTRO
RETIRO
Zoo Aquarium de Madrid
Park Parque de San Isidro
Cultural Center Matadero Madrid
PUENTE DE VALLECAS
CAMPAMENTO
VISTA ALEGRE
USERA
CARABANCHEL
la Rubia
CUATRO VIENTOS
BUENAVISTA
SAN FERMÍN
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

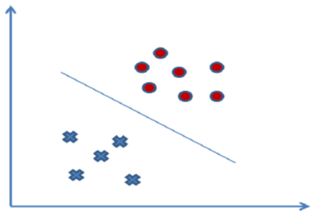
Different types of data

- 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

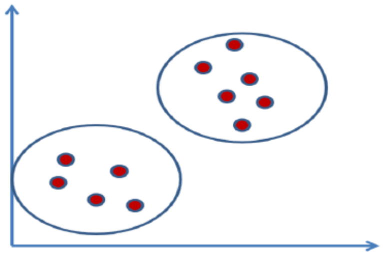
Summing up

Recap

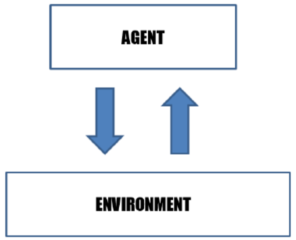
- Supervised



- Unsupervised



- Reinforced



- ML vs Bayes
- Challenges due to BD