

# Use & Value of Unusual Data

Arthur Charpentier

UQAM

Actuarial Summer School 2019

## The Team for the Week

Jean-Philippe Boucher UQAM (Montréal, Canada)  
Professor, holder of The Co-operators Chair in Actuarial Risk Analysis, author of several research articles in actuarial science.

✉@J\_P\_Boucher    Ⓜ jpboucher.uqam.ca



Arthur Charpentier UQAM (Montréal, Canada)  
Professor, editor of *Computational Actuarial Science with R*. Former director of the Data Science for Actuaries program of the French Institute of Actuaries

✉@freakonometrics    Ⓜfreakonometrics    Ⓜ freakonometrics.hypotheses.org



Ewen Galic AMSE (Aix-Marseille, France)  
Assistant professor, teaching computational econometrics, data science and machine learning.

✉@3wen    Ⓜ3wen    Ⓜ egallic.fr



# Practical Issues

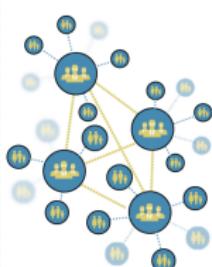
# General Context (in Insurance)

The "insurance company" node at the top holds all the power with one central balance sheet.



Individual risk is transferred to a central entity that has all the power. This one balance sheet gives insurance company immense leverage.

The "insurance company" node still holds the majority of power, but also share risk with each other.



Distributed risk from individuals is transferred to a central entity, but risk is also spread amongst central entities to further reduce risk. Similarly to regular insurance, all the leverage for who receives care and how much is centralized.



## About Friendsurance

In 2010, the founders of the FinTech company, Friendsurance, realized that insurance is expensive and lacks transparency. People pay high premiums each year and get nothing in return.

The vision of Friendsurance is to make insurance easier and more affordable for customers, helping to save customers time and money that can better be spent on the finer side of life. Fifty

## Networks and P2P Insurance (via <https://sharable.life> and source)

# General Context (in Insurance)

FEATURES

## PEER-TO-PEER INSURANCE

### GOING BACK TO BASICS

Innovators have set their sights on simplifying insurance by adopting new peer-to-peer business models. Who are the key players and does the Australian industry need to pay attention?

**DISCUSSION OF** the sharing economy has taken up considerable column space in recent times.

PricewaterhouseCoopers estimates the five main sharing sectors (peer-to-peer finance, online staffing, peer-to-peer accommodation, car sharing and music video streaming) will continue general global revenue of US\$830bn by 2016. Today, PwC says revenue generated is around US\$10bn (A\$15.67bn).

One sharing economy sector still very much in its infancy is peer-to-peer insurance. But Amy Gibbs, digital communications and content strategy manager at ANZIIF, says popularity of the concept is increasing.

"When it takes off, it will likely happen much quicker than we expect," she says.

"P2P insurance is not about a new technology that replaces the old one. It's about people demanding an industry that gives them what they believe they deserve. Consumers expect different things in 2016, and technology now allows them the power to get what they want or go elsewhere."

**The current crop in P2P**  
On 24 March, Germany's high-profile P2P player, Friendinsurance, announced it had collected US\$81.3m (A\$82.60m) from investors in its latest round of financing.

Tim Kundi, Friendinsurance's co-founder and managing director, says the organisation intends to use that fresh capital to grow further in

Germany and expand internationally. He says the first expansion target for 2016 is Australia, and expansion opportunities for other markets are currently being considered.

Friendinsurance is one of the players Gibbs has been keeping a close eye on.

"Friendinsurance and Guevara are ones that I watch closely," she says.

"They have interesting models and appear to have put a lot of thought into them, plus they get the marketing/consumer angle, which is crucial."

Friendinsurance, founded in Germany in 2009, operates as an independent insurance broker. It describes its mission as to make insurance easier and more affordable for customers, and to reduce the number of fraudulent claims.

"Our idea is inspired by insurance in its original form, when people got together in small groups... and supported each other in case of damage with their own resources," says Tim Kundi, friendinsurance's director of insurance business.

"This was easy and efficient but also limited in the extent of coverage. Today, big insurance

companies can carry claims of any amount, but marketing, administration and fraud cause remarkable costs.

"Against this background, we developed an insurance concept that again [creates] smaller groups within bigger insurance societies and rewards remaining claimless within [these groups] with annual cashbacks."

Friendinsurance customers with the same

insurance premium is paid into a cashback pool, and part of it is provided to the group insurer (or reinsurer).

When small claims are made, customers are reimbursed from the general fund. Larger claims still go through the insurer. Groups that have no claims during a year receive a cashback bonus the following January. Friendinsurance says its claims-free bonus is available for private liability, home contents and legal expense insurance.

"So far, more than 80% of users received some of their insurance fees back," says Kundi.

"In the property insurance line, the average cashback was 33% of the paid insurance fees."

Kundi reports that, in 2015, Friendinsurance engaged 75,000 new customers.

"Today, we have a six-digit number of customers, 70 insurance partners, 15 corporates and 60 employees."

Over in the UK, start-up Guevara has received

**"P2P insurance is not about a new technology threatening an industry but about people demanding an industry that gives them what they believe they deserve"**

Amy Gibbs, ANZIIF



## Networks and P2P Insurance, via Insurance Business (2016)

# General Context (in Insurance)

## Telematics Usage-Based Insurance (UBI)

During the first quarter of 2014, LexisNexis commissioned an independent firm to conduct two studies. The first was a study of drivers who were using telematics services to reduce their risk of being involved in a motor vehicle accident. The second was a study of 4000 small fleet managers who were using telematics devices to coordinate fleets of 10 to 20 vehicles.

### Consumer Insights

Overall consumer awareness has plateaued however it's growing among younger drivers



Consumers are now as comfortable sharing UBI driving data as they are sharing many other types of information



And their interest in UBI at lower discount levels is rising



Offering popular value-added services in addition to the discount increases demand



**Decline Dynamics  
IN AUTO  
INSURANCE**

Technology will bring huge changes to automobiles and personal auto insurance over the next decade. Assuming there will be no changes is a risky choice.

By Dr. Michael Rydell

In the U.S., the auto-insurance industry is significant, generating \$20 billion in annual revenue and insuring 270 million vehicles. It makes up more than 10 percent of the total car-industry. It is predicted that by 2020, the number of cars on the road will have substantially changed due to the auto-insurance industry, leading to the eventual decline of the car as we know it.

While this might be true, the more important question is the impact that technology will have on the insurance industry—In decline or rise? Many in the auto-insurance industry have predicted a slow decline of profitability over the next decade. However, the introduction of telematics has caused a major shift in the industry, leading to a much faster decline in the next few years. While the real result is still unknown, there is a need to understand these new dynamics before knowledge of the concept of industrial decline can be used.

**Technology and Turnover Rates**

The lack of concern for money in the insurance industry is the primary driver for the relatively slow introduction of technology in the industry. Creating a \$10 billion automobile driving system—which will compete with losses like the one seen

**TheActuary**  
The magazine of the Institute and Faculty of Actuaries

**INTERVIEW**  
Michael Green  
PwC: Is the big data revolution can benefit society

**INNOVATION**  
Leadership and pricing long-term care

**TECHNOLOGY**  
Tough industrial landscape and insurance landscape

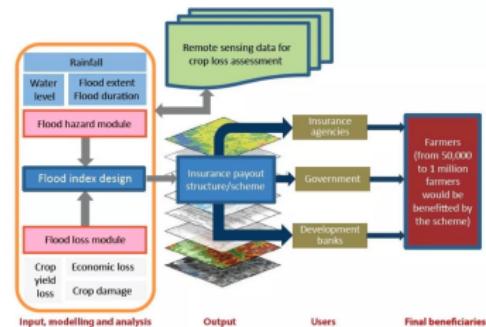
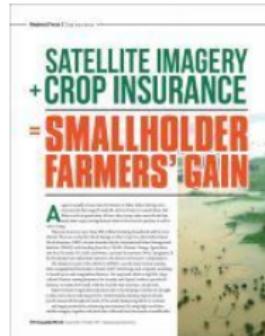
**ENVIRONMENT**  
Emissions performance in sustainable investment

**Who's in the driving seat?**

How telematics could transform motor insurance

## Telematics and Usage-Based-Insurance ([source](#), [source](#), [The Actuary](#))

## General Context (in Insurance)



Satellite pictures ([source](http://ibfi.iwmi.org/) and Index-based Flood Insurance (IBFI), <http://ibfi.iwmi.org/>)

## General Context (in Insurance)



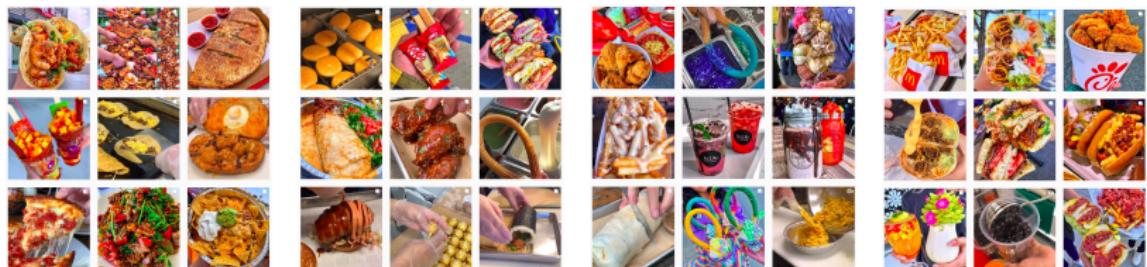
Very small granularity (1 pixel is  $15 \times 15\text{cm}^2$ , see [detail](#))

# More and More (exponentially) Data

1969, (legend says) 5 photographs were taken by Neil Armstrong



More than 800,000,000 pictures are uploaded on every day, on Facebook (2 billion photos per day with [instagram](#), Messenger and Whatsapp)



# Data Complexity

Pictures, drawings, videos, handwriting, text, ECG, medical imagery, satellite pictures, tweets, phone calls, etc.



Hard to distinguish structured and un-structured data...

# Data Complexity



# Insurance Data

From questionnaires  
to data frames  $x_{i,j}$   
row  $i$  : insured, policy, claim  
column  $j$  : variables (features)

HEALTH INSURANCE QUESTIONNAIRE									
<small>Please provide all the information requested and return this form to your eligibility worker. Use and attach a copy of your insurance policy, membership card, or any other aid to help complete this questionnaire. PLEASE TYPE OR PRINT. DO NOT ABBREVIATE. Additional instructions and information collection and access are on the reverse. If you have any questions about completing this form or require Spanish translation, call toll-free 1-800 952-5294 (7:30 a.m. to 5:00 p.m.).</small>									
<small>COMPLETE THIS FORM FOR ANY MEDICAL INSURANCE, INCLUDING MEDICARE SUPPLEMENT, PREPAID HEALTH PLANS/HEALTH MAINTENANCE ORGANIZATIONS, AND OTHER PRIVATE HEALTH INSURANCE. THIS FORM DOES NOT AFFECT YOUR MEDICAL ELIGIBILITY; HOWEVER, FAILURE TO REPORT OTHER HEALTH INSURANCE MAY BE A CAUSE FOR TERMINATION OF YOUR MEDICAL ELIGIBILITY.</small>									
Case name		FOR COUNTY USE ONLY			STATE USE ONLY				
Worker name		Verifier by							
Case address		Date	Date		Initials				
		Worker telephone number (      )	Date		Initials				
<input type="checkbox"/> Initial intake <input type="checkbox"/> Redetermination <input type="checkbox"/> NPI# <small>Optional District number:</small>		Scope			CC number				
<b>SECTION I: Beneficiary Information. LIST ALL PERSONS, INCLUDING UNBORN, ON MEDICAL AND COVERED BY HEALTH INSURANCE POLICY</b>									
<b>14-DIGIT MEDICAL NUMBER</b>									
OHC	Beneficiary Name (First, Middle, Last)	Social Security Number	Sex	Date of Birth	Col. Code	Aid Code	Case Number	FBI No.	Pers. No.
<b>SECTION II: Health Insurance Information</b>									
1. What is the name and address of your health insurance company? Include street number, city, state, and ZIP. Do not use abbreviations. Name: _____ Address: _____ City, State, ZIP: _____									
2. Do you have to obtain medical services from a specific facility or a group of providers? <input type="checkbox"/> Yes <input type="checkbox"/> No									
3. Where do you send your claims? Name: _____ Address: _____ City, State, ZIP: _____									
4. What is the full name, address, phone number, and SSA number of individual, employee, union/member, or person to whom the insurance policy was issued? Name: _____ Address: _____ City, State, ZIP: _____									
Social Security Number: _____ Telephone number: _____ Absent parent? <input type="checkbox"/> Yes <input type="checkbox"/> No									
5. What is the policy number?									
6. When were the dates of your policy? Beginning date: _____ Ending date (if applicable): _____ <input type="checkbox"/> Medical coverage available through employer, but has not been applied for									
7. Premium amount \$: _____ <input type="checkbox"/> Monthly <input type="checkbox"/> Quarterly <input type="checkbox"/> Yearly How are premiums paid? <input type="checkbox"/> By insured <input type="checkbox"/> By employer <input type="checkbox"/> By payroll deduction									
8. Give name, address, and telephone number of plan, employee, group, organization, or school. Name: _____ Address: _____ City, State, ZIP: _____ Telephone number: _____									
9. Does anyone covered beneficiary have an adult, chronic, or pre-existing illness that requires him/her to see a physician? If yes, please specify the illness: <input type="checkbox"/> Hospital outpatient (i.e., lab work/physical therapy) <input type="checkbox"/> Prescription drugs <input type="checkbox"/> Long term care/nursing home <input type="checkbox"/> Hospital stay <input type="checkbox"/> Dental care <input type="checkbox"/> Only specific illnesses (i.e., cancer) <input type="checkbox"/> Doctor visits <input type="checkbox"/> Vision care <input type="checkbox"/> Type of illness									
10. Does your health insurance provide or pay for? (Check all that apply) <input type="checkbox"/> Hospital outpatient (i.e., lab work/physical therapy) <input type="checkbox"/> Prescription drugs <input type="checkbox"/> Long term care/nursing home <input type="checkbox"/> Hospital stay <input type="checkbox"/> Dental care <input type="checkbox"/> Only specific illnesses (i.e., cancer) <input type="checkbox"/> Doctor visits <input type="checkbox"/> Vision care <input type="checkbox"/> Type of illness									
11. Is the policy a Medicare Supplement? <input type="checkbox"/> Yes <input type="checkbox"/> No									
Premium: _____									

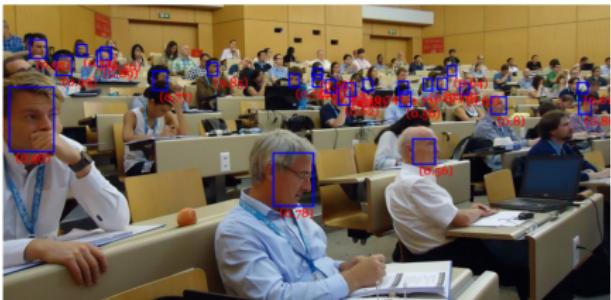
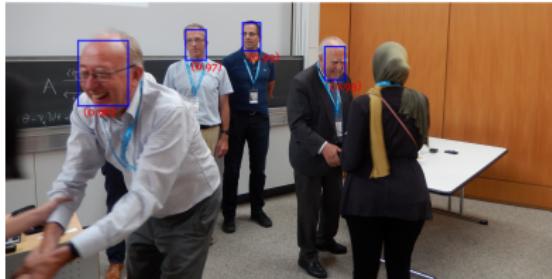
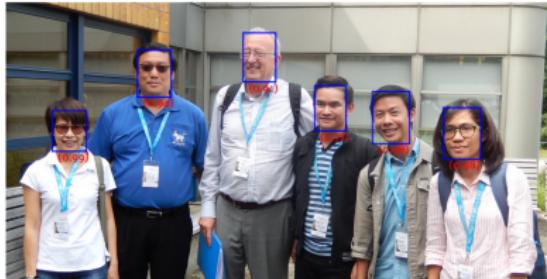
# From Pictures to Data

Consider pictures [saa-iss.ch/testalbum/nggallery/all-old-galleries/](http://saa-iss.ch/testalbum/nggallery/all-old-galleries/)



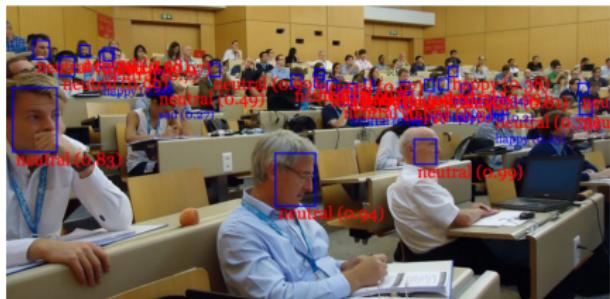
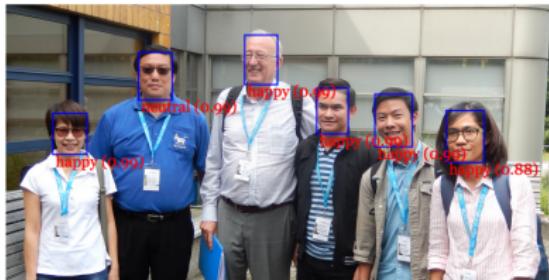
# From Pictures to Data

Face detection face-api.js ([justadudewhohacks.github.io](https://justadudewhohacks.github.io))



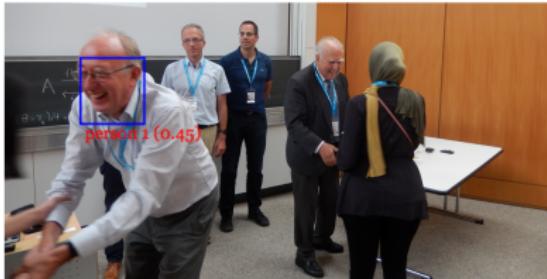
## From Pictures to Data

Expression recognition face-api.js ([justadudewhohacks.github.io](https://justadudewhohacks.github.io))



# From Pictures to Data

Face recognition face-api.js ([justadudewhohacks.github.io](https://justadudewhohacks.github.io))



## Pictures

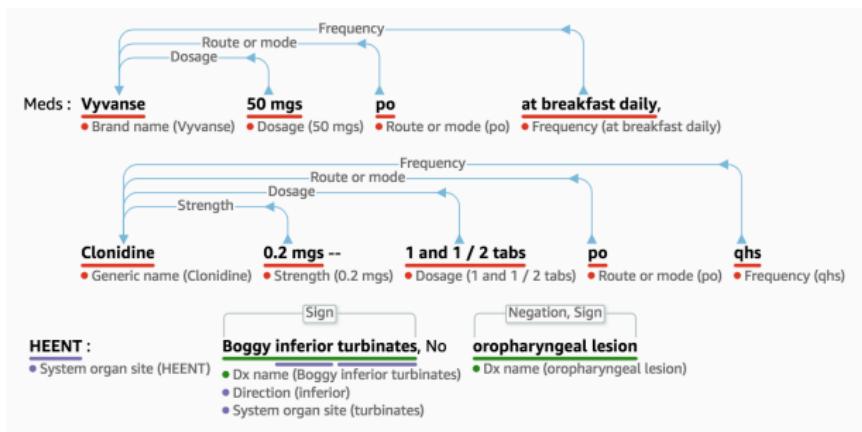
Extracting information from a picture  
What is on the picture ?  
What is the cost of such a claim ?  
High dimensional data :  
 $h \times d$  color picture, e.g.  $850 \times 1280$   
 $\mathbf{x}_i \in \mathbb{R}^d$  with  $d > 3$  million  
Need new statistical tools  
(deep) neural networks perform well  
or reduce dimension

Pictures Odermatts (2003, [Karambolage](#))



# Text

Kaufman *et al.* (2016, Natural Language Processing-Enabled and Conventional Data Capture Methods for Input to Electronic Health Records), Chen *et al.* (2018, A Natural Language Processing System That Links Medical Terms in Electronic Health Record Notes to Lay Definitions) or Simon (2018, Natural Language Processing for Healthcare Customers)

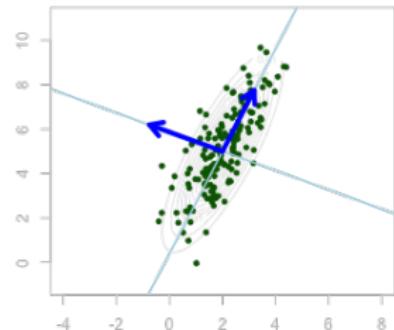


# Statistics & ML Toolbox : PCA

Projection method, used to reduce dimensionality

## Principal Component Analysis (PCA)

find a small number of directions in input space that explain variation in input data represent data by projecting along those directions



$n$  observations in dimension  $d$ , stored in matrix  $\mathbf{X} \in \mathcal{M}_{n \times d}$

Natural ideal: linearly project (multiply by a projection matrix) to much lower dimensional space  $k < d$

Search for orthogonal directions in space with highest variance

Information is stored in the covariance matrix  $\Sigma = \mathbf{X}^\top \mathbf{X}$  when variables are centered

## Statistics & ML Toolbox : PCA

Find the  $k$  largest eigenvectors of  $\Sigma$  : principal components

Assemble these eigenvectors into a  $d \times k$  matrix  $\tilde{\Sigma}$

Express  $d$ -dimensional vectors  $x$  by projecting them to  
 $m$ -dimensional  $\tilde{x}$ ,  $\tilde{x} = \tilde{\Sigma}^\top x$

We have data  $X \in \mathcal{M}_{n \times d}$ , i.e.  $n$  vectors  $x_j \in \mathbb{R}^d$ ,

Let  $\omega_1 \in \mathbb{R}^d$  denote weights.

We want to maximize the variance of  $z = \omega_1^\top x$

$$\max \left\{ \frac{1}{n} \sum_{i=1}^n (\omega_1^\top x_i - \omega_1^\top \bar{x})^2 \right\} = \max \left\{ \omega_1^\top \Sigma \omega_1 \right\} \text{ s.t. } \|\omega_1\| = 1$$

## Statistics & ML Toolbox : PCA

To solve

$$\max \left\{ \frac{1}{n} \sum_{i=1}^n (\omega_1^\top \mathbf{x}_i - \omega_1^\top \bar{\mathbf{x}})^2 \right\} = \max \left\{ \omega_1^\top \boldsymbol{\Sigma} \omega_1 \right\} \text{ s.t. } \|\omega_1\| = 1$$

we can use Lagrange multiplier to solve this constraint optimization problem

$$\max_{\omega_1, \lambda_1} \left\{ \omega_1^\top \boldsymbol{\Sigma} \omega_1 + \lambda (1 - \omega_1^\top \omega_1) \right\}$$

If we differentiate  $\boldsymbol{\Sigma} \omega_1 - \lambda_1 \omega_1 = \mathbf{0}$ , i.e.  $\boldsymbol{\Sigma} \omega_1 = \lambda_1 \omega_1$  i.e.  $\omega_1$  is an eigenvector of  $\boldsymbol{\Sigma}$ , with eigenvalue the Lagrange multiplier. It should be the one with the largest eigenvalue.

## Statistics & ML Toolbox : PCA

$$\begin{aligned} & \max \left\{ \omega_2^\top \Sigma \omega_2 \right\} \\ & \text{subject to } \|\omega_2\| = 1 \\ & \quad \omega_2^\top \omega_1 = 0 \end{aligned}$$

The Lagrangian is  $\omega_2^\top \Sigma \omega_2 + \lambda_2(1 - \omega_2^\top \omega_2) - \lambda_1 \omega_2^\top \omega_1$

when differentiating, we get  $\Sigma \omega_2 = \lambda_2 \omega_2$

$\lambda_2$  is the second largest eigenvalue

Set  $\mathbf{z} = W\mathbf{x}$  and  $\tilde{\mathbf{x}} = M\mathbf{z}$ . We want to solve (where  $\ell$  is the standard  $\ell_2$  norm)

$$\min_{W,M} \left\{ \sum_{i=1}^n \ell(x_i, \tilde{x}_i) \right\}$$

i.e.

$$\min_{W,M} \left\{ \sum_{i=1}^n \ell(x_i, MWx_i) \right\}$$

## Statistics & ML Toolbox : PCA

$$\min_{W,M} \left\{ \sum_{i=1}^n \ell(x_i, MWx_i) \right\}$$

for some  $n \times k$  (for  $W$ ) and  $k \times n$  (for  $M$ ) matrices, with  $k < n$ ,  
see Plaut (2018, [From Principal Subspaces to Principal Components with Linear Autoencoders](#)).

More generally,

$$\min_{W,M} \left\{ \sum_{i=1}^n \ell(x_i, \phi \circ \psi x_i) \right\}$$

for some nonlinear functions  $\phi$  and  $\psi$ .

This is nonlinear PCA (see Monahan (2016, [An Introduction to Nonlinear Principal Component Analysis](#))

See also [autoencoder](#) (neural nets).

# Statistics & ML Toolbox : GLM

## Generalized Linear Model

Flexible generalization of linear regression that allows for  $y$  to have distribution other than a normal distribution

The *classical* regression is (matrix form)

$$\underset{n \times 1}{\mathbf{y}} = \underset{n \times k}{\mathbf{X}} \underset{k \times 1}{\beta} + \underset{n \times 1}{\varepsilon} \text{ or } y_i = \mathbf{x}_i^T \beta + \varepsilon_i$$

$$Y_i \sim \mathcal{N}(\mu_i, \sigma^2) \quad , \text{ with } \mu_i = \mathbb{E}[Y_i | \mathbf{x}_i] = \mathbf{x}_i^T \beta$$

It is possible to consider a more general regression.

For a classification  $\mathbf{y} \in \{0, 1\}^n$ ,

$$Y_i \sim \mathcal{B}(p_i) \quad , \text{ with } p_i = \mathbb{E}[Y_i | \mathbf{x}_i] = \frac{e^{\mathbf{x}_i^T \beta}}{1 + e^{\mathbf{x}_i^T \beta}}$$

Recall that we assumed that  $(Y_i, \mathbf{X}_i)$  were i.i.d. with unknown distribution  $\mathbb{P}$ .

## Statistics & ML Toolbox : GLM

In the (standard) linear model, we assume that  
 $(Y|\mathbf{X} = \mathbf{x}) \sim \mathcal{N}(\mathbf{x}^\top \boldsymbol{\beta}, \sigma^2)$ .

The maximum-likelihood estimator of  $\boldsymbol{\beta}$  is then

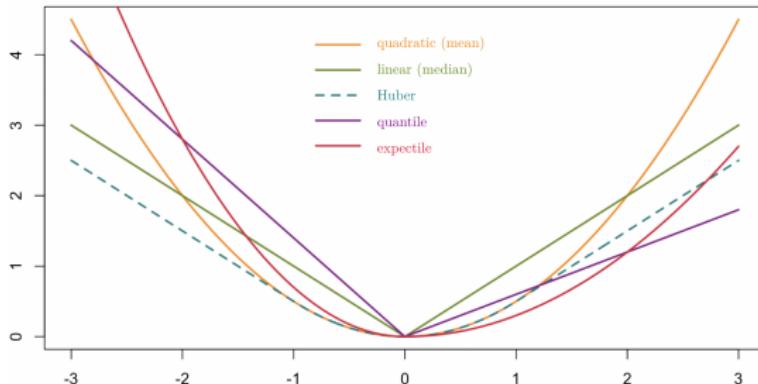
$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^{p+1}}{\operatorname{argmax}} \left\{ - \sum_{i=1}^n (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2 \right\}$$

For the logistic regression,  $(Y|\mathbf{X} = \mathbf{x}) \sim \mathcal{B}(\text{logit}^{-1}(\mathbf{x}^\top \boldsymbol{\beta}))$ .  
The maximum-likelihood estimator of  $\boldsymbol{\beta}$  is then

$$\underset{\boldsymbol{\beta} \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \left\{ \sum_{i=1}^n (y_i \log(\text{logit}^{-1}(\mathbf{x}_i^\top \boldsymbol{\beta})) + (1 - y_i) \log(1 - \text{logit}^{-1}(\mathbf{x}_i^\top \boldsymbol{\beta}))) \right\}$$

# Statistics & ML Toolbox : loss and risk

In a general setting, consider some **loss function**  $\ell : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}_+$ ,  
e.g.  $\ell(y, m) = (y - m)^2$  (quadratic  $\ell_2$  norm)

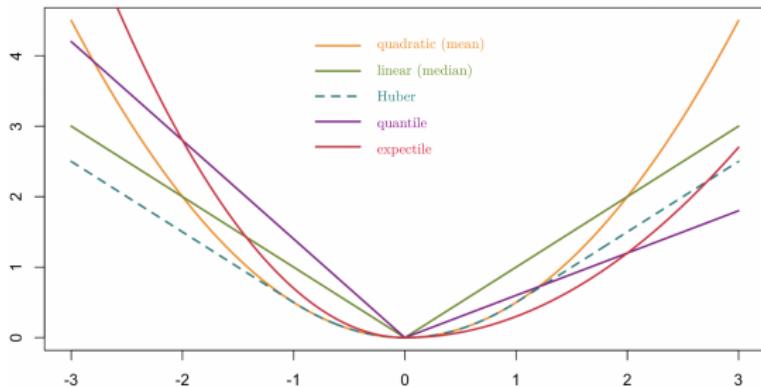


## Average Risk

The **average risk** associated with  $\hat{m}_n$  is  
$$\mathcal{R}_{\mathbb{P}}(\hat{m}_n) = \mathbb{E}_{\mathbb{P}}[\ell(Y, \hat{m}_n(\mathbf{X}))]$$

# Statistics & ML Toolbox : loss and risk

In a general setting, consider some **loss function**  $\ell : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}_+$ ,  
e.g.  $\ell(y, m) = (y - m)^2$  (quadratic  $\ell_2$  norm)



## Empirical Risk

The **empirical risk** associated with  $\hat{m}_n$  is

$$\hat{\mathcal{R}}_n(\hat{m}_n) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, \hat{m}_n(x_i)).$$

# Statistics & ML Toolbox : loss and risk

If we minimize the average risk, we overfit...

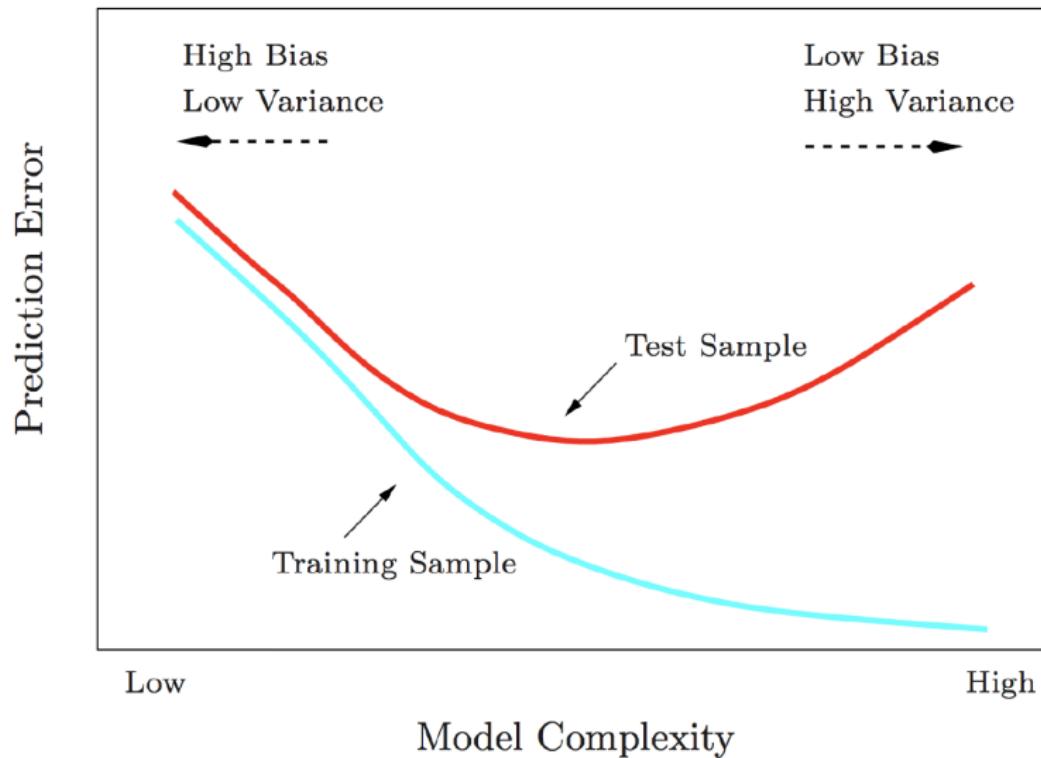
## Hold-Out Cross Validation

1. Split  $\{1, 2, \dots, n\}$  in  $T$  (training) and  $V$  (validation)
- 2 . Estimate  $\hat{m}$  on sample  $(y_i, \mathbf{x}_i), i \in T$  :

$$\hat{m}_T = \operatorname{argmin}_{m \in \mathcal{M}} \frac{1}{|T|} \sum_{i \in V} \ell(y_i, m(x_{1,i}, \dots, x_{p,i}))$$

3. Compute  $\frac{1}{|V|} \sum_{i \in V} \ell(y'_i, \hat{m}_T(x_{1,i''}, \dots, x_{p,i'}))$

# Statistics & ML Toolbox : loss and risk



source: Hastie *et al.* (2009, [The Elements of Statistical Learning](#))

# Statistics & ML Toolbox : loss and risk

## Leave-one-Out Cross Validation

1. Estimate  $n$  models : estimate  $\hat{m}_{-j}$  on sample  $(y_i, \mathbf{x}_i)$ ,  
 $i \in \{1, \dots, n\} \setminus \{j\}$

$$\hat{m}_{-j} = \operatorname{argmin}_{m \in \mathcal{M}} \frac{1}{n-1} \sum_{i \neq j} \ell(y_i, m(x_{1,i}, \dots, x_{p,i}))$$

2. Compute  $\frac{1}{n} \sum_{i=1}^n \ell(y_i, \hat{m}_{-i}(\mathbf{x}_i))$

# Statistics & ML Toolbox : loss and risk

## *K*-Fold Cross Validation

1. Split  $\{1, 2, \dots, n\}$  in  $K$  groups  $V_1, \dots, V_K$
2. Estimate  $K$  models : estimate  $\hat{m}_k$  on sample  $(y_i, \mathbf{x}_i)$ ,  $i \in \{1, \dots, n\} \setminus V_k$
3. Compute  $\frac{1}{K} \sum_{k=1}^K \frac{1}{|V_k|} \sum_{i \in V_k} \ell(y_i, \hat{m}_k(x_{1,i}, \dots, X_{p,i}))$

# Statistics & ML Toolbox : loss and risk

## Bootstrap Cross Validation

1. Generate  $B$  bootstrap samples from  $\{1, 2, \dots, n\}$ ,  $I_1, \dots, I_B$
2. Estimate  $B$  models :  $\hat{m}_b$  on sample  $(y_i, \mathbf{x}_i)$ ,  $i \in I_b$
3. Compute  $\frac{1}{B} \sum_{b=1}^B \frac{1}{n - |I_b|} \sum_{i \notin I_b} \ell(y_i, \hat{m}_b(x_{1,i}, \dots, X_{p,i}))$

The probability that  $i \notin I_b$  is  $\left(1 - \frac{1}{n}\right)^n \sim e^{-1} (= 36.78\%)$

At stage  $b$ , we validate on  $\sim 36.78\%$  of the dataset.

## Statistics & ML Toolbox : loss and risk

$\mathbf{y} \in \mathbb{R}^d$ ,  $\bar{y} = \operatorname{argmin}_{m \in \mathbb{R}} \left\{ \sum_{i=1}^n \frac{1}{n} [\underbrace{y_i - m}_{\varepsilon_i}]^2 \right\}$  is the empirical version of

$$\mathbb{E}[Y] = \operatorname{argmin}_{m \in \mathbb{R}} \left\{ \int [\underbrace{y - m}_{\varepsilon}]^2 dF(y) \right\} = \operatorname{argmin}_{m \in \mathbb{R}} \left\{ \mathbb{E}[\|\underbrace{Y - m}\|_{\ell_2}] \right\}$$

where  $Y$  is a random variable.

Thus,  $\operatorname{argmin}_{m: \mathbb{R}^k \rightarrow \mathbb{R}} \left\{ \sum_{i=1}^n \frac{1}{n} [\underbrace{y_i - m(\mathbf{x}_i)}_{\varepsilon_i}]^2 \right\}$  is the empirical version of  $\mathbb{E}[Y | \mathbf{X} = \mathbf{x}]$ .

See Legendre (1805, *Nouvelles méthodes pour la détermination des orbites des comètes*) and Gauß (1809, *Theoria motus corporum coelestium in sectionibus conicis solem ambientium*).

## Statistics & ML Toolbox : loss and risk

$\text{med}[\mathbf{y}] \in \operatorname{argmin}_{m \in \mathbb{R}} \left\{ \sum_{i=1}^n \frac{1}{n} \underbrace{|y_i - m|}_{\varepsilon_i} \right\}$  is the empirical version of

$$\text{med}[Y] \in \operatorname{argmin}_{m \in \mathbb{R}} \left\{ \int \underbrace{|y - m|}_{\varepsilon} dF(y) \right\} = \operatorname{argmin}_{m \in \mathbb{R}} \left\{ \mathbb{E}[\underbrace{\|Y - m\|_{\ell_1}}_{\varepsilon}] \right\}$$

where  $\mathbb{P}[Y \leq \text{med}[Y]] \geq \frac{1}{2}$  and  $\mathbb{P}[Y \geq \text{med}[Y]] \geq \frac{1}{2}$ .

$\operatorname{argmin}_{m: \mathbb{R}^k \rightarrow \mathbb{R}} \left\{ \sum_{i=1}^n \frac{1}{n} \underbrace{|y_i - m(\mathbf{x}_i)|}_{\varepsilon_i} \right\}$  is the empirical version of  
 $\text{med}[Y | \mathbf{X} = \mathbf{x}]$ .

See Boscovich (1757, *De Litteraria expeditione per pontificiam ditionem ad dimetiendo duos meridiani*) and Laplace (1793, *Sur quelques points du système du monde*).

## Statistics & ML Toolbox : loss and risk

For least-squares ( $\ell_2$  norm), we have a strictly convex problem...  
classical optimization routines can be used (e.g. gradient descent)

More complicated with the  $\ell_1$  norm

Consider a sample  $\{y_1, \dots, y_n\}$ .

To compute the median, solve  $\min_{\mu} \left\{ \sum_{i=1}^n |y_i - \mu| \right\}$  which can be solved using linear programming techniques., see Dantzig (1963, [Linear programming](#)).

More precisely, this problem is equivalent to  $\min_{\mu, \mathbf{a}, \mathbf{b}} \left\{ \sum_{i=1}^n a_i + b_i \right\}$  with  $a_i, b_i \geq 0$  and  $y_i - \mu = a_i - b_i, \forall i = 1, \dots, n$ .

## Statistics & ML Toolbox : loss and risk

For a quantile, the linear program is  $\min_{\mu, \mathbf{a}, \mathbf{b}} \left\{ \sum_{i=1}^n \tau a_i + (1 - \tau) b_i \right\}$

with  $a_i, b_i \geq 0$  and  $y_i - \mu = a_i - b_i, \forall i = 1, \dots, n.$

This is extended to the quantile regression,

$$\min_{\beta^\tau, \mathbf{a}, \mathbf{b}} \left\{ \sum_{i=1}^n \tau a_i + (1 - \tau) b_i \right\}$$

with  $a_i, b_i \geq 0$  and  $y_i - \mathbf{x}_i^\top \beta^\tau = a_i - b_i, \forall i = 1, \dots, n.$

But is the following problem the one we should really care about ?

$$m^* = \operatorname{argmin}_{m \in \mathcal{M}} \left\{ \sum_{i=1}^n \ell(y_i, m(\mathbf{x}_i)) \right\}$$

# Statistics & ML Toolbox : Penalization

Usually, there are **tuning** parameters  $p$ . To avoid overfit,

1. use cross-validation

$$m_{p^*}^* = \operatorname{argmin}_p \left\{ \sum_{i \in \text{validation}} \ell(y_i, m_p^*(x_i)) \right\}$$

where  $m_p^* = \operatorname{argmin}_{m \in \mathcal{M}_p} \left\{ \sum_{i \in \text{training}} \ell(y_i, m(x_i)) \right\}$

2. use penalization

$$m^* = \operatorname{argmin}_{m \in \mathcal{M}} \left\{ \sum_{i=1}^n \ell(y_i, m(x_i)) + \text{penalization}(m) \right\}$$

## Statistics & ML Toolbox : Penalization

Penalization often yield bias... but it might interesting if the risk is the mean squared error...

Consider a parametric model, with true (unknown) parameter  $\theta$ , then

$$\text{mse}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2] = \underbrace{\mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2]}_{\text{variance}} + \underbrace{\mathbb{E}[(\mathbb{E}[\hat{\theta}] - \theta)^2]}_{\text{bias}^2}$$

One can think of a **shrinkage** of an unbiased estimator  $\tilde{\theta}$ ,  $\hat{\theta} = \alpha \cdot \tilde{\theta}$

Let  $\tilde{\theta}$  denote an unbiased estimator of  $\theta$ ,

$$\hat{\theta} = \frac{\theta^2}{\theta^2 + \text{mse}(\tilde{\theta})} \cdot \tilde{\theta} \leq \tilde{\theta}$$

satisfies  $\text{mse}(\hat{\theta}) \leq \text{mse}(\tilde{\theta})$ .

# Statistics & ML Toolbox : Ridge

as in Wieringen (2018 [Lecture notes on ridge regression](#))

## Ridge Estimator (OLS)

$$\hat{\beta}_\lambda^{\text{ridge}} = \operatorname{argmin} \left\{ \sum_{i=1}^n (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^p \beta_j^2 \right\}$$
$$\hat{\beta}_\lambda^{\text{ridge}} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbb{I})^{-1} \mathbf{X}^\top \mathbf{y}$$

## Ridge Estimator (GLM)

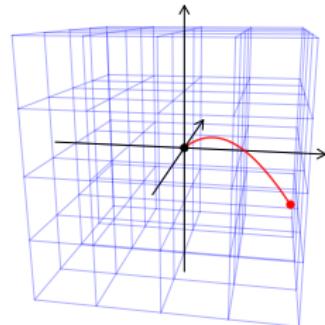
$$\hat{\beta}_\lambda^{\text{ridge}} = \operatorname{argmin} \left\{ - \sum_{i=1}^n \log f(y_i | \mu_i = g^{-1}(\mathbf{x}_i^\top \boldsymbol{\beta})) + \frac{\lambda}{2} \sum_{j=1}^p \beta_j^2 \right\}$$

# Statistics & ML Toolbox : Ridge

Observe that if  $\mathbf{x}_{j_1} \perp \mathbf{x}_{j_2}$ , then

$$\hat{\beta}_\lambda^{\text{ridge}} = [1 + \lambda]^{-1} \hat{\beta}^{\text{ols}}$$

which explain relationship with shrinkage.  
But not in the general case...



\* `chicago.txt` dataset, with three covariates

Smaller mse

There exists  $\lambda$  such that  $\text{mse}[\hat{\beta}_\lambda^{\text{ridge}}] \leq \text{mse}[\hat{\beta}^{\text{ols}}]$

It is important to normalize variables. We assume here that we did.

## Statistics & ML Toolbox : Ridge

From a Bayesian perspective,

$$\underbrace{\mathbb{P}[\theta|y]}_{\text{posterior}} \propto \underbrace{\mathbb{P}[y|\theta]}_{\text{likelihood}} \cdot \underbrace{\mathbb{P}[\theta]}_{\text{prior}} \quad \text{i.e.} \quad \log \mathbb{P}[\theta|y] = \underbrace{\log \mathbb{P}[y|\theta]}_{\text{log likelihood}} + \underbrace{\log \mathbb{P}[\theta]}_{\text{penalty}}$$

If  $\beta$  has a prior  $\mathcal{N}(\mathbf{0}, \tau^2 \mathbb{I})$  distribution, then its posterior distribution has mean

$$\mathbb{E}[\beta|y, X] = \left( X^T X + \frac{\sigma^2}{\tau^2} \mathbb{I} \right)^{-1} X^T y.$$

# Statistics & ML Toolbox : LASSO

## LASSO Estimator (OLS)

$$\hat{\beta}_\lambda^{\text{lasso}} = \operatorname{argmin} \left\{ \sum_{i=1}^n (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^p |\beta_j| \right\}$$

## LASSO Estimator (GLM)

$$\hat{\beta}_\lambda^{\text{lasso}} = \operatorname{argmin} \left\{ - \sum_{i=1}^n \log f(y_i | \mu_i = g^{-1}(\mathbf{x}_i^\top \boldsymbol{\beta})) + \frac{\lambda}{2} \sum_{j=1}^p |\beta_j| \right\}$$

- \* least absolute shrinkage and selection operator  
(pun with *garrote*, Breiman (1995, [Better Subset Regression Using the Nonnegative Garrote](#))  
see Bayesian regression when  $\boldsymbol{\beta}$  has a Laplace prior.

## Statistics & ML Toolbox : sparsity

In several applications,  $k$  can be (very) large, but a lot of features are just noise:  $\beta_j = 0$  for many  $j$ 's. Let  $s$  denote the number of relevant features, with  $s \ll k$ , cf Hastie, Tibshirani & Wainwright (2015, *Statistical Learning with Sparsity*),

$$s = \text{card}\{\mathcal{S}\} \text{ where } \mathcal{S} = \{j; \beta_j \neq 0\}$$

The model is now  $y = \mathbf{X}_{\mathcal{S}}^T \boldsymbol{\beta}_{\mathcal{S}} + \varepsilon$ , where  $\mathbf{X}_{\mathcal{S}}^T \mathbf{X}_{\mathcal{S}}$  is a full rank matrix.

# Statistics & ML Toolbox : sparsity

## $\ell_0$ -regularization Estimator (OLS)

$$\hat{\beta}_\lambda^{\ell_0} = \operatorname{argmin} \left\{ \sum_{i=1}^n (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2 + \lambda \|\boldsymbol{\beta}\|_{\ell_0} \right\}$$

where  $\|\mathbf{a}\|_{\ell_0} = \sum \mathbf{1}(|a_j| > 0) = \dim(\boldsymbol{\beta})$ .

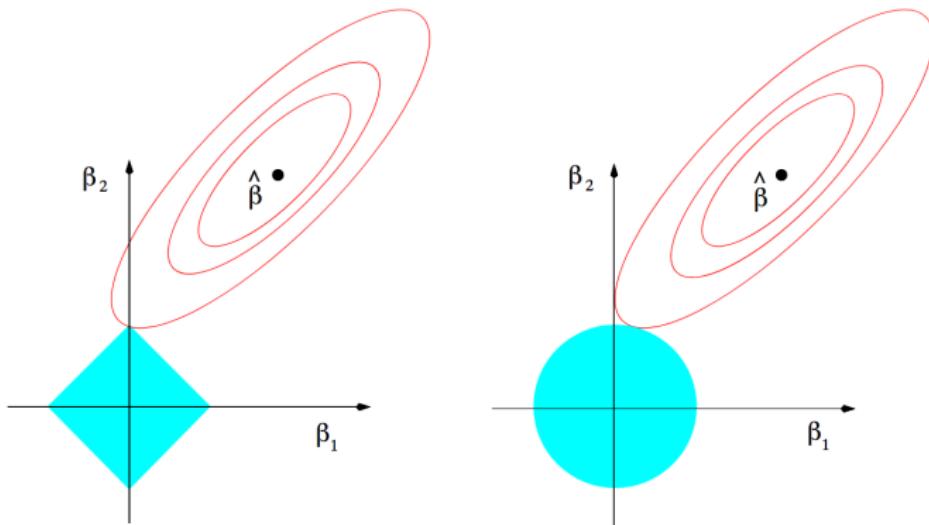
More generally

## $\ell_p$ -regularization Estimator (OLS)

$$\hat{\beta}_\lambda^{\ell_p} = \operatorname{argmin} \left\{ \sum_{i=1}^n (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2 + \lambda \|\boldsymbol{\beta}\|_{\ell_p} \right\}$$

- **sparsity** is obtained when  $p \leq 1$
- **convexity** is obtained when  $p \geq 1$

## Statistics & ML Toolbox : sparcity



**FIGURE 3.11.** Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions  $|\beta_1| + |\beta_2| \leq t$  and  $\beta_1^2 + \beta_2^2 \leq t^2$ , respectively, while the red ellipses are the contours of the least squares error function.

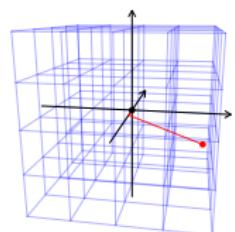
source: Hastie et al. (2009, *The Elements of Statistical Learning*)

# Statistics & ML Toolbox : LASSO

In the orthogonal case,  $\mathbf{X}^T \mathbf{X} = \mathbb{I}$ ,

$$\hat{\beta}_{k,\lambda}^{\text{lasso}} = \text{sign}(\hat{\beta}_k^{\text{ols}}) \left( |\hat{\beta}_k^{\text{ols}}| - \frac{\lambda}{2} \right)_+$$

i.e. the LASSO estimate is related to the soft threshold function

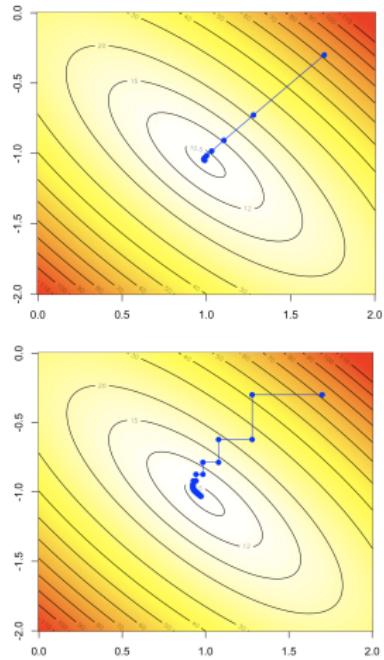


$$S_h(x) = \text{sign}(x)(|x| - h)$$

\* `chicago.txt` dataset, with three covariates

## LASSO Coordinate Descent Algorithm

1. Set  $\beta_0 = \hat{\beta}$
- 2 . For  $k = 1, \dots$   
for  $j = 1, \dots, p$ 
  - (i) compute  $R_j = \mathbf{x}_j^\top (\mathbf{y} - \mathbf{X}_{-j} \beta_{k-1(-j)})$
  - (ii) set  $\beta_{k,j} = R_j \cdot \left(1 - \frac{\lambda}{2|R_j|}\right)_+$
3. The final estimate  $\beta_k$  is  $\hat{\beta}_\lambda$



The **softmax** function  $\log(1 + e^{x-s})$  can be used as a smooth version of  $(x - s)_+$

## Statistics & ML Toolbox : trees

Use recursive binary splitting to grow a tree

### Trees (CART)

Each interior node corresponds to one of the input variables; there are edges to children for each of the possible values of that input variable. Each leaf represents a value of the target variable given the values of the input variables represented by the path from the root to the leaf.

based on some impurity index, with, for leaf  $\ell$ , and  $f(p) = p(1 - p)$

$$I(\ell) = \sum_{y \in \{0,1\}} \frac{n_{y,\ell}}{n_\ell} \cdot f\left(\frac{n_{y,\ell}}{n_\ell}\right) \text{ with no split}$$

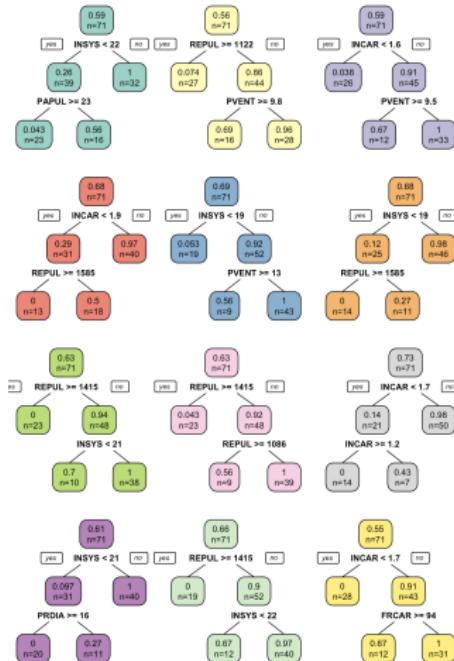
$$I(\ell_L, \ell_R) = \sum_{x \in \{L,R\}} \sum_{y \in \{0,1\}} \frac{n_{y,\ell_x}}{n_{\ell_x}} \cdot f\left(\frac{n_{y,\ell_x}}{n_{\ell_x}}\right) \text{ with split}$$

# Statistics & ML Toolbox : forests

## Bagging (Bootstrap + Aggregation)

1. For  $k = 1, \dots$ 
  - (i) draw a bootstrap sample from  $(y_i, \mathbf{x}_i)$ 's
  - (ii) estimate a model  $\hat{m}_k$  on that sample
2. The final model is

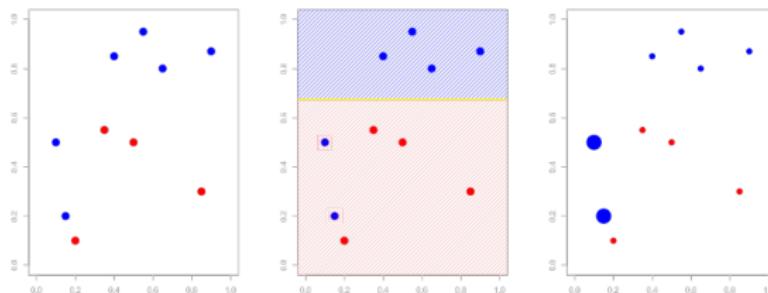
$$m^*(\cdot) = \frac{1}{\kappa} \sum_{k=1}^{\kappa} \hat{m}_k(\cdot)$$



# Statistics & ML Toolbox : boosting

## Bosting & Sequential Learning

$$m_k(\cdot) = m_{k-1}(\cdot) + \operatorname{argmin}_{h \in \mathcal{H}} \left\{ \sum_{i=1}^n \ell(\underbrace{y_i - m_{k-1}(\mathbf{x}_i)}_{\varepsilon_i}, h(\mathbf{x}_i)) \right\}$$



# Statistics & ML : Variable Importance

Parallel bootstrap + trees = random forests (bagging)

Sequential learning + stumps = boosted trees

see Elith *et al.* (2000, [A working guide to boosted regression trees](#))

## Variable Importance

Importance of feature  $k$  calculates each feature importance as the sum over the number of splits (across all trees) that include the feature  $k$ , proportionally to the number of samples it splits.

# Statistics & ML Toolbox : Neural Nets

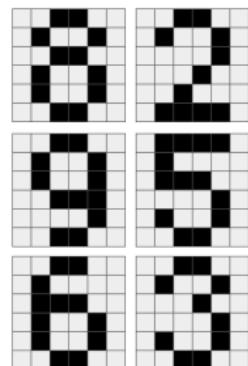
Very popular for pictures...

Picture  $x_i$  is

- a  $n \times n$  matrix in  $\{0, 1\}^{n^2}$  for black & white
- a  $n \times n$  matrix in  $[0, 1]^{n^2}$  for grey-scale
- a  $3 \times n \times n$  array in  $([0, 1]^3)^{n^2}$  for color
- a  $T \times 3 \times n \times n$  tensor in  $(([0, 1]^3)^T)^{n^2}$  for video

$y$  here is the label ("8", "9", "6", etc)

Suppose we want to recognize a "6" on a picture



$$m(x) = \begin{cases} +1 & \text{if } x \text{ is a "6"} \\ -1 & \text{otherwise} \end{cases}$$

# Statistics & ML Toolbox : Neural Nets

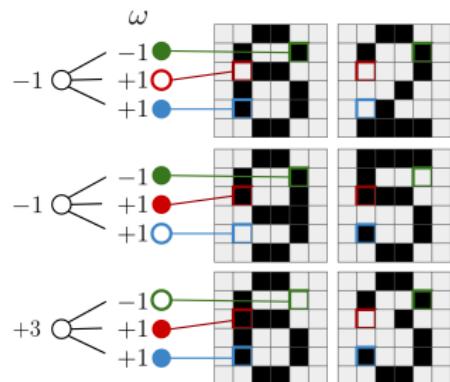
Consider some specific pixels, and associate weights  $\omega$  such that

$$\hat{m}(\mathbf{x}) = \text{sign} \left( \sum_{i,j} \omega_{i,j} x_{i,j} \right)$$

where

$$x_{i,j} = \begin{cases} +1 & \text{if pixel } x_{i,j} \text{ is black} \quad \blacksquare \\ -1 & \text{if pixel } x_{i,j} \text{ is white} \quad \square \end{cases}$$

for some weights  $\omega_{i,j}$  (that can be negative...)



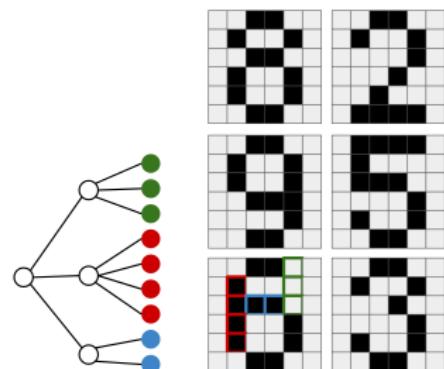
# Statistics & ML Toolbox : Neural Nets

A deep network is a network with a lot of layers

$$\hat{m}(\mathbf{x}) = \text{sign} \left( \sum_i \omega_i \hat{m}_i(\mathbf{x}) \right)$$

where  $\hat{m}_i$ 's are outputs of previous neural nets

Those layers can capture shapes in some areas  
nonlinearities, cross-dependence, etc



## Binary Threshold Neuron & Perceptron

If  $x \in \{0, 1\}^P$ , McCulloch & Pitts (1943, [A logical calculus of the ideas immanent in nervous activity](#)) suggested a simple model, with **threshold**  $b$

$$y_i = f \left( \sum_{j=1}^p x_{j,i} \right) \text{ where } f(x) = \mathbf{1}(x \geq b)$$

where  $\mathbf{1}(x \geq b) = +1$  if  $x \geq b$  and 0 otherwise, or (equivalently)

$$y_i = f \left( \omega + \sum_{j=1}^p x_{j,i} \right) \text{ where } f(x) = \mathbf{1}(x \geq 0)$$

with **weight**  $\omega = -b$ . The trick of adding 1 as an input was very important !

$\omega = -1$  is the **or** logical operator :  $y_i = 1$  if  $\exists j$  such that  $x_{j,i} = 1$   
 $\omega = -p$ , is the **and** logical operator :  $y_i = 1$  if  $\forall j$ ,  $x_{j,i} = 1$

## Binary Threshold Neuron & Perceptron

but not possible for the **xor** logical operator :  $y_i = 1$  if  $x_{1,i} \neq x_{2,i}$

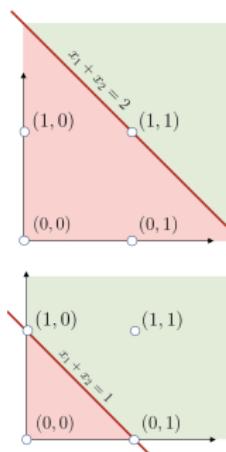
Rosenblatt (1961, *Principles of Neurodynamics: Perceptron & Theory of Brain Mechanism*) considered the extension where  $x$ 's are real-valued, with **weight**  $\omega \in \mathbb{R}^p$

$$y_i = f \left( \sum_{j=1}^p \omega_j x_{j,i} \right) \text{ where } f(x) = \mathbf{1}(x \geq b)$$

where  $\mathbf{1}(x \geq b) = +1$  if  $x \geq b$  and 0 otherwise, or (equivalently)

$$y_i = f \left( \omega_0 + \sum_{j=1}^p \omega_j x_{j,i} \right) \text{ where } f(x) = \mathbf{1}(x \geq 0)$$

with **weights**  $\omega \in \mathbb{R}^{p+1}$



# Binary Threshold Neuron & Perceptron

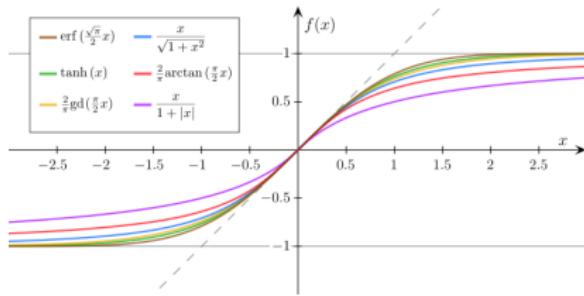
Minsky & Papert (1969, [Perceptrons: an Introduction to Computational Geometry](#)) proved that perceptron were a linear separator, not very powerful

Define the **sigmoid** function  $f(x) = \frac{1}{1 + e^{-x}} = \frac{e^x}{e^x + 1}$  (= **logistic** function).

This function  $f$  is called the **activation function**.

If  $y \in \{-1, +1\}$ , one can consider the hyperbolic tangent

$f(x) = \frac{(e^x - e^{-x})}{(e^x + e^{-x})}$  or the inverse tangent function (see [wikipedia](#)).



# Statistics & ML Toolbox : Neural Nets

So here, for a classification problem,

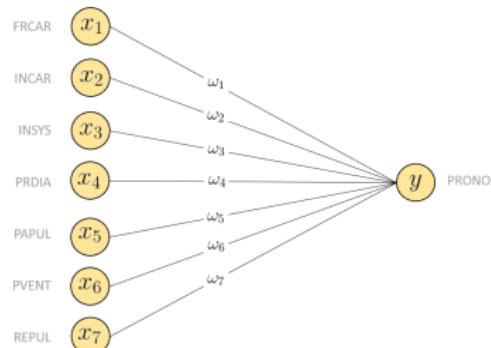
$$y_i = f \left( \omega_0 + \sum_{j=1}^p \omega_j x_{j,i} \right) = p(\mathbf{x}_i)$$

The error of that model can be computed using the quadratic loss function,

$$\sum_{i=1}^n (y_i - p(\mathbf{x}_i))^2$$

or cross-entropy

$$\sum_{i=1}^n (y_i \log p(\mathbf{x}_i) + [1 - y_i] \log [1 - p(\mathbf{x}_i)])$$



# Statistics & ML Toolbox : Neural Nets

Consider a single hidden layer, with 3 different neurons, so that

$$p(\mathbf{x}) = f \left( \omega_0 + \sum_{h=1}^3 \omega_h f_h \left( \omega_{h,0} + \sum_{j=1}^p \omega_{h,j} x_j \right) \right)$$

or

$$p(\mathbf{x}) = f \left( \omega_0 + \sum_{h=1}^3 \omega_h f_h \left( \omega_{h,0} + \mathbf{x}^\top \boldsymbol{\omega}_h \right) \right)$$

