ESPCI PARIS

EDUCATION SCIENCE INNOVATION

Introduction to Big Data

Statistical & Machine Learning

Julien Barrier — class of 2018

Introduction

I have spartialed writing this handout from the notes I have taken from Olivier Rivoire's course on Big Data and Statistical learning at ESPCI Paris from May to April 2018. Please be considerate if some mistakes crop up in this work.

Julien

Some book reading is advised during the course, partialicularly:

- *The Elements of Statistical Learning*, T. Hastie, R. Tibshirani and J. Friedman, Springer Series in Statistics, 2008;
- *Information Theory, Inference, and Learning Algorithms*, D.J.C. MacKay, Cambridge University Press, 2003.

Dr Olivier Rivoire
Center for Interdisciplinary Research in Biology (CIRB)
Collège de France
olivier.rivoire@college-de-france.fr

Applications

There are plenty of applications for Big Data problems. A few examples may be given:

Post learn + identify digits on enveloppes

Biology DNA sequencing

IT Face recognition

etc.

Big Data is an issue of growing importance. As engineers, we may be familiar with such concepts.

Idea of marchine learning

The main idea of machine learning is to find models to give prediction of input data. In facts, Big Data models are deduced from a

training batch of N input-output data, on which programs train to generalise models. The deduced model *input* $i \rightarrow$ *output* i can then be generalised to give prediction from a random input, as long as it relates to the training batch.

Analyticall, let's spartial with a collection of x and y data, where x stands for the input data and y is the vector of the output data. Each sample is going to have multiple dimensions, therefore we may use an algebraic model. Let N be the number of samples used and p the dimension of each x data. We may write x as an N, p matrix and y as a vector of p dimensions.

We now N samples of p dimensions x_{ij} associated with the N output data y_i.

From now on there are two possible cases: y_i can be known or unknown. In the fist case (y_i known), the problem is said to be *supervised*. Hence we may work with a finite discrete set of data: $y_i = 1, \cdots, K$. This problem is called categorical, and we can solve it with *classification*. We may also work with an infinite set of numbers: $y_i \in \mathbb{R}$. This problem is called quantitative, and we can solve it with *regression*.

The second case (y_i unknown) is said to be *unsupervised* and can be solved via *clustering* or *dimension reduction* methods.

Deep learning

In the past few years, there have been huge progress in the *deep learning* approach. It is based on so-called neurol networks, that are models inspired by the brain operation.

People are trying to understand how to train these networks. It has had remarkable outcomes in image recognition, social network filtering, medical diagnoses, etc.

Deep learning is based on hidden layers, placed inbetween input and output layers , that are trained to find correlations and mathematical models.

The goal of this course is to explain whate these objects are, how do they work, and put it in relation with state of thepartial research.

What kind of open problems are there? How do neural networks operate? What are their unsuperised learning behaviour?

Contents

Least square regression, from small to big data

Linear Regression at One Dimension

Let $p = 1$. If we work with N points, then $i = 1, \cdots, N$, and we work with a set of data (x_i, y_i).

The goal here is to make a prediction of what the y data should be when x is given.

The simplest possible model is the linear regression given by the equation **??**.

$$y = \alpha + \beta x. \tag{1}$$

Here, the main issue is to get the best α and β for a partialicular set of data. To know what the best choice is, we may define a cost function, that returns a number representing how well the regression permorms. In neural network problems, the cost fuction return number is associated with how well the neural network performs in mapping training examples to the correct output.

There are several choices that can be made to define the cost function. At one dimension, the simplest choice is the sum of squared residuals, defined in equation **??**, usually shortened as SSR.

$$l(\alpha, \beta) = \frac{1}{N} \sum_{i=1}^{N} (y_i - \alpha - \beta x_i)^2 \tag{2}$$

Figure **??** illustrate a simple geometrical interpretation of what the SSR is. Actually, the lower ϵ_i, the better the fit.

TODO!
Figure 1: geometrical intepretation of the SSR, where ϵ_i is given by the relation: $\epsilon_i^2 = (y_i - \alpha - \beta x_i)^2$

For all we have done up to now, we never have never worked with big data. We need p large enough to consider this as a real big data issue.

If we're looking at a hundreds or thousands pixels picture composed of hundreds, p will be large in comparison with N. That is a full statistics problem.

Currently, $p = 1$ is small data, but all we did there has been a correct introduction to clearly understand big data problems.

Striking a good fit necessitates finding the best α and the best β. For this, we may look at the optimum, defined as the points where the derivative of l versus α and β vanishes. This is given by equations **??** and **??**.

$$\frac{\partial l}{\partial \alpha} = -\frac{1}{N}\sum_{i=1}^{N}(y_i - \alpha - \beta x_i) = 0 \quad (3)$$

$$\frac{\partial l}{\partial \beta} = -\frac{1}{N}\sum_{i=1}^{N} x_i(y_i - \alpha - \beta x_i) = 0 \quad (4)$$

To solve this set of equations, we require a substitution for x and y.

Let's define the mean values[1]:

[1] We must keep in mind that $\overline{x^2} \neq \bar{x}^2$.

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \quad (5)$$

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i \quad (6)$$

$$\overline{xy} = \frac{1}{N}\sum_{i=1}^{N} x_i y_i \quad (7)$$

$$\overline{x^2} = \frac{1}{N}\sum_{i=1}^{N} x_i^2 \quad (8)$$

Thus, equations **??** and **??** can be reduced as:

$$\frac{\partial l}{\partial \alpha} = -(\bar{y} - \alpha - \beta\bar{x}) \quad (9)$$

$$\frac{\partial l}{\partial \beta} = -(\overline{xy} - \alpha\bar{x} - \beta\overline{x^2}) \quad (10)$$

This yields to:

$$\alpha = \bar{y} + \beta\bar{x} \quad (11)$$

$$\overline{xy} = -\bar{y}\bar{x} - \beta\bar{x}^2 - \beta\overline{x^2} \quad (12)$$

There we may substitute α and β:

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

$$\begin{aligned}\hat{\beta} &= \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} \\ &= \frac{\mathrm{Cov}(x,y)}{\mathrm{Cov}(x,x)} \\ &= \frac{\mathrm{Cov}(x,y)}{\mathrm{Var}(x)}\end{aligned}$$

We use the given notations:

$$\begin{aligned}\mathrm{Cov}(x,y) &= \bar{xy} - \bar{x}\bar{y} \\ &= \overline{(x-\bar{x})}(y-\bar{y}) \\ \mathrm{Var}(x) &= \mathrm{Cov}(x,x) \\ \sigma(x) &= \sqrt{\mathrm{Var}(x)}\end{aligned}$$

Let us define the Pearson coefficient $\mathcal{R}$ by the relation **??**.

$$\mathcal{R} = \frac{\mathrm{Cov}(x,y)}{\sigma(x)\sigma(y)} \quad (13)$$

The Pearson coefficient $\mathcal{R}$ is always comprised between 0 and 1. Thus, we can define the quantity $\mathcal{R}^2$, that relates to the quality of the fit:

$$\mathcal{R}^2 = 1 - \frac{\hat{l}}{\mathrm{Var}(y)} \quad (14)$$

In the general case, we look at models where $\beta = 0$. In this case, the cost function l would be the sum of the square distance to the line. Figure **??** depicts two linear regressions with different parameters. The right figure shows a much better linear regression, with much lower square distances between the points and the line.

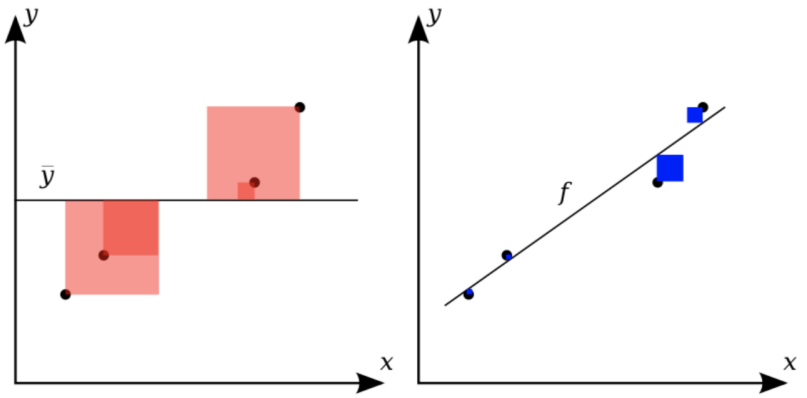

Figure 2: Two linear regression taken from two datasets. The left one shows higher square distances than the right one.

We may note that in some cases, it would be better to rescale the axis to fit the data with a linear regression. Logarithm axis is the most popular way of rescaling an axis to have a correct assumption. It is usually more valuable to rescale the axis and perform a linear regression, than trying to find an higher order fit.

Linear Regression at Higher Dimensions

We are now considering higher dimensions data ($p > 1$), that are full, meaning that $p << N$. It means that, for an example of data, we might add different parameters. If we take the example (given in class) of correlations between the velocity of people versus the size of towns, we might add other relevant parameters, like the average heigh of people, their ages, etc. We might then examine many potential predictors. Thus we need to generalise the same things, where each input now becomes a vector of p dimensions, as presented in equation **??**

$$(x_{i,1}, \cdots, x_{i,p}), \qquad \forall i \tag{15}$$

Let now $x_{i,j}$ be the matrix of the input data, where i is the number of samples, varying from 1 to N, and j is the dimension, ranged between 1 to p .

We may generalise the relation $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$ in the new **??** equation:

$$\hat{y}_i = \hat{\alpha} + \sum_{j=1}^{p} \hat{\beta}_j x_{ij} \tag{16}$$

If we have this key figure, we can always add 1 in the x vector, as the $p+1$ coordinate. We can thus assume that α vanishes. In fact, we

can always redefine the data so that α vanishes. We can also rescale the variable, by removing the mean:

$$x' = x - \bar{x} \tag{17}$$
$$y' = y - \bar{y} \tag{18}$$

Therefore, the output coordinate $\hat{y}_i$ can be written as the product $\hat{y}_i = X\hat{\beta}$, that is a much more convenient way to write it.

That are just restrictions of the problems that help us to compute it.

Let $l(\beta)$ be the cross-function, define with equation **??**.

$$l(\beta) = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 \tag{19}$$

Let Z be a vector whose components z_i are defined as follows:

$$\sum_{i=1}^{N} z_i^2 = ||Z||^2 = Z^T Z \tag{20}$$

Therefore we can write the cross-function as:

$$l(\beta) = \frac{1}{N}(Y - X\beta)^T(Y - X\beta) \tag{21}$$

This form can easily be differentiated with β, and the retrieved derivative vanishes at the extremum (eq. **??**).

$$\frac{\partial l}{\partial \beta} = -\frac{Z}{N}X^T(Y - X\beta) = 0 \tag{22}$$

The equation **??** can be reduced as $X^T Y = X^T X\beta$, which can be solved by introducing the matrix $C = X^T X$ (eq. **??**)[2]

$$\hat{\beta} = (X^T X)^{-1}X^T Y = C^{-1}X^T Y \tag{23}$$

At higher dimension, the geometry consists in fitting with an hyperplane, as shown on figure **??**.

Here, we are essentially solving a system of equations. We must consider the number of variables adapted to the number of equations that we get. When there is not enough equations (when p is too small for example), the system is undetermined. We cannot reduce it and do not have a single solution.

Actually, when $p > N$, we can solve this problem with the condition $\hat{l} = 0$. This is a situation where there are more parameters than there are equaltions. It is easy to solve. The solutions consists in overfitting.

For instance if we have 100 parameters and 10 equations, we can never manage to get any result. However, we can find easy solutions, but this will overfit. At this stage, the system cannot be inverted.

[T] denotes the transpose matrix, defined by the relation: $[\mathbf{A}^{\mathrm{T}}]_{ij} = [\mathbf{A}]_{ji}$

[2] Note that

$$C_{ij} = \sum_{k=1}^{N} x_{kj}x_{ki}$$

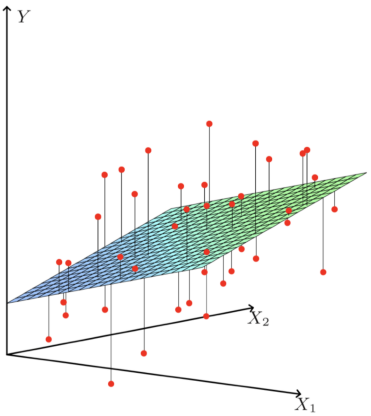

Figure 3: Linear least squares fitting with $X \in \mathbb{R}^2$. In this problem, we are looking for the linear function of X that minimises the sum of squared residuals from Y, which is an plane (hyperplane in dim 3)

When p is large, even if it is of the same order of magnitude as N, we are working with big matrixes, that can be tricky to invert with both proper mathematical accuracy and computation performances. In such cases, we should handle the data carefuly.

Large p are typical way of using statistical learning methods, by discriminating between datasets without having the aforesaid issues.

General principles

Models, bias vs variance, cross-validation, maximum likelihood, Bayes, etc.

So, what do we do now? In general, we want to know what are the most interesting parameters. At the end, we can predict the issue with one or two parameters. Even if we spartial with a lot of data, we want to find what are the parameterts, the dimensions important in the problem we are working on.

Two advantages:

Interpretation With less parameters, we can estimate them with much more accuracy than if we have more. In general, it's easier and more accurate to estimate things on a condensed set of parameters than on a large set. The issue is: how do you compromise: having enough parameters to have a good enough estimation of the problem, without having too much and loosing precision. Enough information VS enough precision.

Let $\tilde{l}(\beta)$ be defined as:

$$\tilde{l}(\beta) = l(\beta) + \lambda ||\beta||_q \tag{24}$$

With $||\beta||_0 = \#(\beta_j \neq 0)$ (cardinal)

We look for $min_\beta l(\beta)$ given $||\beta||_0 \leq C$

With $\beta = [0, \cdots, \beta_i, 0, \cdots, 0]$ There ain't any good numerical solution.

There's always a compromise between what we are able to optimise efficiently, and what is possible to optimise. A version of the problem that is easy to solve is:

$$min_\beta l(\beta) given ||\beta||_2 \leq C_1 \tag{25}$$

Ridge regression: there's a way to get to this problem, using the constraints that can be solved efficiently numerically.

Or $\min_\beta l(\beta)$, given $||\beta||_1 \leq C_2$. This is called the Lasso regression.

In the ridge problem, we assume that the problem is sparse: we only need a few parameters to capture the relationship.

Why did we spartial to write it this way?

$$\tilde{l}(\beta) = (Y - X\beta)^T (Y - X\beta) + \lambda \beta^T \beta \tag{26}$$

$$\frac{\partial \tilde{l}(\beta)}{\partial \beta} = 2(-X^T(Y - X\beta) + \lambda\beta) = 2(-X^T Y + (X^T X + \lambda\infty)\beta \quad (27)$$

∞ is the identity matrix.

What is doing is putting constraints. You add constraints and then you can solve mathematically the problem.

On wenesday, we will compare those two regression, and define the framework for machine learning.

$$C_{jk} p > N \quad (28)$$

$$C_{jk} = \frac{1}{N} \sum_{i=1}^{N} x_{ij} x_{ik} \quad (29)$$

$$\bar{x} = 0; C = X^T X \quad (30)$$

$$X_{ij} N x p \quad (31)$$

$p > N => C$ is non invertible.
$N = 1, C_{jk} = x_{1j} x_{1k}$
$C = XX^T$
here, C is of rank 1.
NB : remind that $Z^T Z = ||Z||^2$

$$Z^T y =< Z, y >= \vec{Z} \cdot \vec{y} \quad (32)$$

Mathematically, $rank(C) \leq N$. I have too many parametetrs, not enough samples. When I try to solve this linear regression problem, I have too many solutions.

There are issues when p>N, but also when they are of the same order of magnitude.

Example: financial data.

We want to get information from this data, but there's no label, no y data. In general, we don't take the raw data, but try to find something more adapted to the problem. Here, we want to get rid of the α parameter; for this reason, we use the r_{ti} data instead of the $s_i(t)$ parameter.

Then we define the x_{ti}, by substracting the mean and normalising with the standard deviation.

Therefore, the x_{ti} value has a null mean, and a standard deviation rescaled to 1. Therefore each data vary within the same range.

If we move to the data C_{ij}

When we have some data, an important step is to watch it by eye, and try to find correlations. If something is obvious to the eye, we'll try to interpret it with the math. Example: Exxon&Chevron are strongly correlated, JP Morgan and Bank of America are also strongly correlated. Thus, we may suppose that $C_{Ex,Ch} > C_{Ex,JP}$.

In order to analyse the data, we may compute the spectrum. Or see clearly from the definition that the matrix is symmetric, and has all the properties to be diagonalised.

$$C_{jk}, C_{jj} = 1; C_{ij} = C_{ji} \tag{33}$$

Thus we get $\lambda_1, \ldots, \lambda_p$ eigenvalues, and $v_1, \ldots, v_p$ eigenvectors.

TODO!
Figure 4: dispersion of the eigenvalues

There's no way that I can have a good estimation of these metrics.

Bottom = control. They just shuffle the data, make permutation of the values. It is the same stocks. Randomy shuffle the data, to remove all the interesting information (correlation between the different stocks).

It's a way to see what kind of correlation we can get just from randomness, in a case where there's no corraltion from the data.

We can quantify the quantity of noise.

98% of the eigenvalues are contained in the 1st partial of the data. That is 98% of noise.

Therefore I write:

$$C = \sum_{j=1}^{p} \lambda_j v_j v_j^T \tag{34}$$

$$v_j^T v_k = \delta_j^k \tag{35}$$

Random metrics theory: branch of statustical physics, we can compute analytically the shape of the control series. RMT.

$min_x l(x)$ given $g(x) \leq C$

$min_x l(x) + \lambda g(x)$; $\lambda \geq 0$.

TODO!
Figure 5: scheme

I assume everything is differentiable.

$$\nabla_x l(x) = -\lambda \nabla_x g(x) \tag{36}$$

The conditions tells that the two gradients should be aligned, and directed in opposite directions. Generally, it would depend on the value C. Minimum value., can be a local minimum if the problem isn't perfectly shaped.

If I have $l(\beta)$, and am considering the norm of β to be less that C value:

$$||\beta||_q \leq C_q \tag{37}$$

We can do it with l_0, l_2, sometimes also with the l_1 norm.

Why are we interested in that kind of things, what does it give us?

With a not too big dataset, taken from a book. The goal here is to try to predict the crime rate, and to what it is correlated.

(x_{ij}, y_i) $i = 1..N = 50$ cities $j = 1 \ldots 5$

The idea is to consider naively a simple problem.

Here we may find linear combination of all different problems. For a physicist, it seems we're not allowed to to so, because not homogeneous. But this helps finding correlations.

$x'_{ij} = x_{ij} - \bar{x_j}$. In this case, we have zero mean, We may also want try to divide by the standard deviation. Not the case here.

$$y = \sum_{j=1}^{p} \beta_j x_j \tag{38}$$

Therefore we can write $l(\beta) = \frac{1}{N} \sum_{i=1}^{N} (y_i - \sum_{j=1}^{p} \beta_j x_{ij})^2$

The result of this optimisation may be given as a function.

Graph on the right, ridge regression. What is plotted is the value of the β along the x axis. This has to do with the cost (c_q). We can repeat for different values of the cost. If we do it for large C, we do not put any constraints, and therefore get the same β.

if we put a very strong cost, like o, the only solution is $\beta = 0$. Hence we're looking at different solutions, constrained, and then we relax it to a state where there's no constraint anymore.

The lasso graph is the same, but performed with l_1 norm.

There's a way to understand this, by giving an illustration.

Fig 2.2 slides.

$$y = F(x, \theta). \tag{39}$$

Linear models:

$$F(x, \theta) = \sum_{j=1}^{p} \theta_j x_j \tag{40}$$

The principle is to have the results of p data, and then once we get another dataset, similar to the previous one, we're able to fit it and to find the solution.

$$x_1, \cdots x_p, x_1^2, \cdots, x_p^2, \cdots cosx_1, \cdots \tag{41}$$

$$F(x, \theta) = \sum \theta_j h_j(x) \tag{42}$$

Later on, we'lll see neural networks. There, the function h people are generally using is: $h_j(x) = \frac{1}{1+\exp(-\omega_j^T x)} \omega_j$ is the weight. We'll see this later.

-> There's a relation between x and y. There's no limit to the complexity of the problem we can take.

Loss function $L(y, F(x, \theta)) = (y - F(x, \theta))^2$.

Training error: $err_{training}$ given a model and given a loss function:

$$err_{training} = \frac{1}{N} \sum_{i=1}^{N} L(y_i, F(x_i, \theta)) \tag{43}$$

This is not the only quantity we want to consider. test/generalisation error. This one would be the error we get when we are using these datapoint that have not been used in the training of the problem.

There's the training set, used to learn the parameters, and the additional data, on which we're going to apply the model, and to try to generalise the data that have been used as an input for the fit.

$$Err_{test} = \frac{1}{N'} \sum_{i=1}^{N'} L(y'_i, F(x'_i, \theta)) \tag{44}$$

Plot in slides: training vs test errors.

The objective is not to get the best fit, but to generalise the given data.

procedure: K-fold cross-validation. We would divide the data in 5 datasets, and then, take 1 out to use as the test, and all the other one as training test. And then we repeat for all the other combinations. Divide data in K=5 or 10, use most of the data to train, and use one set to test.

Data set is splitted in:

- training set (for the fit)
- test set (for model selection)
- Validation set (for assessment)

Typical number would be : 50%,25%,25%.

We use this to understand ho to find the best parameters.

$y = F(x)$

Training set: $\hat{\theta}$

Model $y = F(x, \theta)$

Thus we have $y = \hat{F}(x) = F(x, \theta)$

Another dataset:

x_0

$$\mathrm{E}[(F(x_0) - \hat{F}(x_0))^2] \tag{45}$$

Where $\hat{(}F)$ is the prediction.

This is considered over different training sets. It says how far I am from the value I want to get the prediction.

$$\mathrm{E}[] = F(x_0)^2 - 2F(x_0)\mathrm{E}(\hat{F}(x_0)) + \mathrm{E}(\hat{F}(x_0)^2) \tag{46}$$

Where $\mathrm{E}(\hat{F}(x_0)^2)$ is $\mathrm{Var}(\hat{F}(x_0) + (\mathrm{E}[\hat{F}(x_0)])^2$

i.e. $\mathrm{E}[] = (F(x_0) - \mathrm{E}[\hat{F}(x_0)]^2 + \mathrm{Var}(\hat{F}(x_0))$.

Test error = $(bias)^2$ + variance

We can generalise this when there is some noise ϵ (random variable) :

$$y = F(x) + \epsilon \tag{47}$$

There we will add another parameter : $var(y)$ that is irreducible error.

In the context of linear regression, there's the Gauss-Markov theorem. It tells us that in linear regression, all the estimators that have no bias, the best one is the one that is minimising the loss function

$$L(y, F(x)) = (y - F(x))^2 \tag{48}$$

This theorem is telling us that if we're interesting in minimising the bias in the context of linear regression we should take $min_\beta l(\beta)$

No bias + min var => $min_\beta l(\beta)$. In general, the best solution is not the solution that has no bias. I will estimate better the parameters that I have with a constrained set of data.

Supervised learning: classificaton, regression, nearest neighbours

last time, we learnt unsupervised learning. $(x, y)_{i=1\cdots N}$

Goal : learn $x \rightarrow y$. function: theta so that when given new x_i

$x_i \rightarrow^{predict} \hat{y_i} \sim y_i$ function: $\hat{\theta}$

function with two components : err_{test} wich is composed of the bias + variance. if large amount of data, high variance, very hard to learn, and we get very different parameters.

we may want to compromise this with a model with less parameters with a lower variance problem. not only we want to pick the best variable in the model, but also the model itself. we consider a class of problems, described with a parameter λ. we introduce λ as a parameter of regularisation.

we are difining, for example, the error as:

$$l(\beta, \lambda) = \frac{1}{N} \sum_{i=1}^{N} (y_i - \beta x_i)^2 - \lambda ||\beta||_0$$

- training -> $\hat{\theta}$ given λ
- validation -> $\hat{\lambda}$
- test -> Err_{test}

At the begining, we divide the dataset into multiple datasetts.

K-fold cross validation : the idea is that we have one dataset, that we divide into K subsets. We can get it with statistics. two weeks ago, we've seen an example of this, in the context of linear regression. minimising the mean square error

planning of the day:

- Bayesian approach
- how we do the approximation in practice.
- example of CLASSIFICATION.

First, let's discuss about maximum likelihood estimation. This is not a big-data specific approach. general in statistics. In general, I would have a model of the form $y = f(x, \theta) + \epsilon$ where ϵ stands for the noise. it is a random variable. We suppose ϵ is a normal variable: $\epsilon(0, \sigma^2)$

The probability to see x with theta:

$$P(y|x,\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp - \frac{(y - f(x,\theta))^2}{2^2} \tag{49}$$

TODO!
Figure 6: P(y | x,theta) representation. with the standard deviation σ^2 and centered on f(x,theta)

Now, let's imagine we are given a set of values (x_i, y_i). We want to find the best parameter $\hat{\theta}$. We look at the parameter that make the data the most likeable.

$$\begin{aligned}
\mathcal{L}(\theta|_1, \cdots, Z_n) &= \log P(Z^N|\theta) && (50)\\
&= \sum_{i=1}^{N} \log P(Z_i|\theta) && (51)\\
&= \sum_{i=1}^{N} \left[-\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y_i - f(x_i,\theta))^2 \right] && (52)\\
&= -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1} N(y_i - f(x_i,\theta))^2 && (53)\\
&= -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2} l(\theta) && (54)
\end{aligned}$$

MSE
MLE: $max_\theta \mathcal{L}(\theta, Z^N) \rightarrow \hat{\theta}$ theorem: if $y = f(x,\theta_0) + \epsilon$

$$\mathrm{E}[\hat{\theta}] \rightarrow_{N\rightarrow\infty} \theta_0$$

$$\hat{\theta}(\theta_0, F(\theta_0)^{-1})$$

$F(\theta) = \mathrm{E}\mathcal{L}(\theta)$ this is called the Fisher information.
$I(\theta)_{ij} = -\frac{\partial^2 \mathcal{L}}{\partial\theta_j \partial\theta_k}$
Therefore,
$\hat{\theta} \sim (approx) N(\hat{\theta}, I(\hat{\theta})^{-1})$

TODO! figure parabolle inversée, maximum : courbature $-\partial\mathcal{L}/\partial\theta^2$, absisse max : θ_0. ordonnée : L.

We want to find the best parameter, *i.e.* the one that is the more likelihood to be ...

L is called the log-likelihood. P is colled the likelihood. In a sense, we want to find the most likely model.

We can mention, that there are two difficulties :

- we need to find the maximum
- problem of validation: if we have a more complicated model, we would increase the log likelihood, and no way to do the validation.

Now, another approach: the Bayesian approach. Again, this is not specific to big-data. We may know that there are lot of debates in the meaning of probabilities. There are two schools: the frequentists: probabilities have a meaning only when the event is reapeating many times. Like a coin tail or head. Fundamentally, if I do it a large number of time, this is converging. If we take an event that can happen only once: the probability that there's lifeon the moon: for a frequentist, there's no meaning. the Bayesian view is different in nature: probability is not about counting, but about biliefs. This represent how I believe the event to be actually the case.

Concretely, the idea of the bayesian approach is: elementary probability: when I have two random variables X, Y. The joint probability of (X,Y): P(X,Y). we can also define the marginals : $P(X)$, $P(Y)$, that are only the probability $P(X) = \sum_y P(X, Y)$. $P(X|Y)$: conditional probability. $P(X, Y) = P(X|Y)P(Y)$ therefore $P(X|Y) = \frac{P(X,Y)}{P(Y)}$.

$P(Z, \theta) = \frac{P(Z,\theta)}{P(Z)} = \frac{P(Z|)P(\theta)}{P(Z)}$. This is called the Bayes formula.

$P(Z = 1) = \theta$ $P(Z = 0) = 1 - \theta$, for example, for a binomial problem.

$P(Z|\theta)$ is the likelihood I had before. In this approach, there's something new, that is $P(\theta)$ wich is called the prior. For a bayesian, we always have some *a priori* bieliefs about the distribution of the parameters. When I see the data, I have to update my beliefs. $P(\theta|Z)$ is called the posterior. $P(Z)$ is called the evidence.

Here we have to do the inference, that is the general model. Usually, when we look at probabilites, we look at $P(Z|\theta)$. reverse approach. We look at the model, that also incorporates the prior.

On the slides, one partialicular example of the prior to solve a praticular problem.

Here, there's nothing to do with big data. example taken from the book of MacKay example: Jo has a test for a disease. a = state of health. a=1 if sick, 0 otherwise. the test is giving another variable b what is known about this test is that it can be positive even if there's no disease. $P(b = 1|a = 1) = 95\%$. same for zero. it means the test gives the right result in 95% of the cases. we need to know the case when somebody of HIS AGE has the disease. this is gonna be the prior. $P(a = 1) = 1\%$. The exercise is to find what is the probability $P(a = 1|b = 1)$. In the bayesian approach, we do not have a theta, but a distribution of the theta we always have a probability to have different values, especially the maximum *a posteriori* estimate (MAP) which is given by taking the max of this: $\max_\theta P(\theta|Z)$.

It will give the same results if we are assuming that this is not depending on theta. If we have a flat prior. Then, this is equivalent to the MLE.

When I'm maximising the posterior, it means I'm maximising the likelihood. Thus we can see the likelihood as a bayesian approach, with a partialicular prior value.

Let's say we have the same model as before. This time, we assume the prior is a one dimensional variable, with a gaussian distribution.

$f(\theta) = \sqrt{\frac{\lambda}{2\pi}} e^{-\lambda \frac{\theta^2}{2}}$

Then, this will be very similar to the l validation, if we take the log of this. Then, this is multiplied by lambda theta square over two.

When we take a prior on these parameters, we want to give more probability to the small values of the parameters. width of the gaussian: 1/sqrt(lambda). control the probability of the parameter to have a large value. restricting the range of the parameter that we

are considering, as we saw before

The goal here is to present this partialicular approach, and recover the maximum likelihood. An interesting aspect of the bayesian approach. Again, this is very general. Let's say that, in general, we have the probability of the data, given some hypotheses:

$P(D|H_1)$ Dis the data, H_1 thehypothesis.

we want to compare with another hypothesis: $P(D|H_2)$

What the bayesian approach is telling us is that we have to consider the probability : $\frac{P(H_1|D)}{P(H_2|D)} = \frac{P(D|H_1)P(H_1)}{P(D|H_2)P(H_2)}$

it depends on the prior we put there. we do not here have no belief????

One example of the maximum likelihood and the general approach.

Problem of generalization. Let's say we have the data: $x_{i=1}$. two approaches are possible: MLE that gives $\hat{\theta}$ bayesian that gives $P(\theta|X)$.

what is the probability of a partialicular value?

$P(x_{N+1})$ =(MLE) $P(x_{N+1}|\hat{\theta})$. if we need a best value of approximation, let's replace the parameter.

= (bayesian)$P(x_{N+1}|X^N) = \sum_\theta P(x_{N+1}|\theta P(\theta|x^N)$

here, $P(\theta|x^N)$ should be used as the new prior. We are constantely updating our believes. this is the prior we get before knowing the value, depending on what we saw before. This has to be equal to: $\frac{P(x^N|\theta)P(\theta)}{P(x^N)}$

There's a famous problem that has to do with that: the rule of succession. The sun is rising every morning, what is the probability it will rise tomorrow? Laplace discussed this issue.

With different approaches, we may get different results. One of the simplest example. Let's assume there's some probability θ the sun is rising in the morning. We hae a binomial model. This is the same issue as a coin that always ends up in the same edge: tail for example. Maximum likelihood estimation: the probability would be 1.

$x_i = 0/1$.

$x^N = (x_1, \ldots, x_N) = (1, \ldots, 1)$

$P(x_{N+1} = 1)$ =? If we do the maximum likelihood approximation, this would be 1 everytime.

X^N : $N_1 times 1; N - N_1 times 0$. (times : frequency it happens over time)

$P(x_i|theta) = \theta$. One parameter model, binomial model. Obviously, the probability $P(x_i = 0|\theta) = 1 - \theta$. this example cannot be simpler than this.

theta can be anything between 0 and 1. $0 \leq \theta \leq 1$.

I'm giving the same probability for every θ. $P(\theta)\Gamma uniform$.

$P(x_{N+1} = 1|x^N) = \int d\theta P(x_{N+1} = 1|\theta)P(\theta|x^N) = \int d\theta\theta \frac{P(x^N|theta)P(\theta)}{P(x^N)} where P(\theta) = 1 = \frac{N!}{N_1!(N-N_1)!}\theta^{N_1}(1-\theta)^{N-N_1}$

The difficulty lies in the $P(x^N)$ that does not depend on theta.

$P(\theta|x^N) = C(N, N_1)\theta^{N_1}(1-\theta)^{N-N_1}$

$\int d\theta P(\theta|x^N) = 1$

$\int_0^1 d\theta\theta^a(1-\theta)^b = \frac{a!b!}{(a+b+1)!}$ this exists also for a and b that are not integer values, and this is called the gamma function. If we use this formula, we would find this to be:

$c(N, N_1) = \frac{(N+1)!}{N_1!(N-N_1)!}$ with the proper normalisation. I need just one line to compute the stuff. If I compute this:

$P(x_{N+1} = 1|x^N) = \int_0^1 d\theta\theta \frac{(N+1)!}{N_1!(N-N_1)!}\theta^{N_1}(1-\theta)^{N-N_1} = \frac{(N+1)!}{N_1!(N-N_1)!}\frac{(N_1+1)!(N-N_1)!}{(N+2)!} = \frac{N_1+1}{N+2} \neq \frac{N_1}{N}(MLE)$

this is called as the laplacian formula.

we have a non-zero probability to observe something that we have never observed before.

$N = 3.X^N = (1,1,1).canwebetthatx_4 = 1$? maybe it is not very wise to say this. this rule takes this into account. here, the probability to be 0 will be $1/5$, not zero.

people that carry out statistics use pseudo-counts. we are adding one zero and one one. It is a way to regularise the variation. In this case, a frequentist would say that we have too few points and that we must give up the problem. for a bayesian, the calculation would be very dependant on the prior. Prediction on the next outcomes.

commentary: $I(\theta) = -\frac{\partial^2\mathcal{L}}{\partial\theta_i\partial\theta_j}$

$\mathcal{F}(\theta) = \mathrm{E}[I(\theta)]$ MLE: $\hat{\theta}(\dots)$ For each dataset, we can use another expectation. It is mathematical consideration, considering all the possibilities of my data.If we arepartialificially generating data when we want to prove all mathematical results analytically, we need all the possible datasets that we can get. variance about everything we can get when generating different datasets.

second hour.

Now I want to discuss the computational issues.

$\hat{\theta}$ that I want to maximise. log likelihood: $\mathcal{L}(\theta|x^N)$. we would have, in general, to get these data numerically, and not analytically, with optimised function. compromises to be done.

Very simple, but the problem can be complicated if the function has several minimum. Let's assume the problem is convex: the function is convex, as well as the set of data. Any global minimum is a global minimum. It can be generate, but,...

In all these problems, we can consider a gradient descent.

If we want to minimise the function,

scheme fig 2. If we look at the gradient, and spartial from a point (random). we look at the grandient, and move in the direction of it. we usually take a very small displacement. spartial iteratively.

$\theta_{t+1} = \theta_t - \eta\left(\frac{\partial L}{\partial\theta}\right)_{\theta_t}$

scheme: cf notes.

cf learning rate, etc.

$L(\theta) = \sum_{i=1}^N (y_i - f(x_i,\theta))^2 = \sum L_i(\theta)$. for each calculation, we have to recompute the data. a version of this algorithm is used very often, and is called stochastic gradient. What we do is: compute $L_i(\theta)$ for a subset of the points. So we take a subset of samples to

estimate $L(\theta) = \sum_{i \in subset}$ and we change at each iteration. The sample is called mini-batch.

There's one version of this algorithm, where everytime we take a single value as a subset: subset =1. this is called on-line learning. in this case, what we are doing is:

$\theta_{t+1} = \theta_t - \eta \nabla_\theta L_i(\theta_t)$.

The issue is to get faster. In general, we have to take the sum. At the end, it is equivalent to do a move at everytime than getting the sum from the very beginning. Actually ,this algorithm is very powerful. In neural networks, backpropagation is nothing more than this algorithm put in application.

This is really a local method. If I spartial from one point, I can end somewhere different. I can be trapped in local minima. It is really difficult to find the correct minimum. That's why we usually work with convex functions, where local minima do not exist elswhere than the global minimum.

l_0 norm, l_1 norm. $||\beta||_0 = \#nonzero\beta_j; ||\beta||_1 = \sum_j |\beta|_j$

For this reason the closest problem to the first is with l_1, and it is convex so that it can be solved with an iterative method. Generally, these are considerations we want to take into account. We will see examples of doing this next time.

Next week: exercise as homework. practical. that will be the grade.

$min_\beta L(\beta) -> \hat{\beta}$

Lasso regression. this is one in which we are going to impose a condition: $||\beta|| \leq t$ this is equivalent to the fact that we want to minimize : $min_\beta L(\beta) + \lambda ||\beta|| with some parameter \lambda$.

ridge regression: $||\beta|| = ||\beta||_2^2 = \sum \beta_i^2$

if we take the l_1 norm:

lasso regression: $||\beta|| = ||beta||_1 = \sum |\beta_i|$

At this stage, we can take it as an exercise.

Le'ts spartial with the function I want to minimize. $\mathcal{L}(\beta) = \frac{1}{N} \sum_{i=1}^N (y_i - \beta x_i)^2 + \lambda |\beta|$

differentiation: we must be careful.

$\frac{\partial \mathcal{L}}{\partial \beta} = 0$

the maximum, if β is positive ,let's say, we can get the value: $(\beta > 0)$

$\frac{\partial \mathcal{L}}{\partial \beta} = -\frac{2}{N} \sum_{i=1}^N (y_i - \beta x_i) x_i + \lambda$.

I would take $\bar{x} = 0, \bar{y} = 0$ and $\overline{x^2} = 1$. normalise all the data.

In general, everything can have completely different units. It makes sense here to normalise, so that each data has the same range of variation. So

$dL/dbeta = -2(\frac{X^T y}{N} - \beta) + \lambda = 0$

thus $\hat{\beta} = \frac{X^T y}{N} - \frac{\lambda}{2}$

If $\frac{X^T y}{N} > \frac{\lambda}{2}$, thus $\hat{\beta} = \frac{X^T y}{N} - \frac{\lambda}{2} > 0$

if $\beta < 0$, then we can use the same approach.

$\hat{\beta} = S_{\lambda/2}\left(\frac{X^T y}{N}\right)$

$S_a(u) = (u) \max(O, |u| - a)$. soft-thresholding operator. cf figure

notes.

This is a figure for p=1.

cf on the website a slide with the formulas.

for $p > 1$,

$L(\beta) = \frac{1}{N} \sum_i (y_i - \sum_k \beta_k x_{ik})^2 + \lambda \sum_k |\beta_k|$

$= \frac{1}{N} \sum_i (y_i - \beta_j x_{ij} - \sum_{k \neq j} \beta_k x_{ik})^2 + \lambda |\beta_j| + \lambda \sum_{k \neq j} |\beta_k|$

let's define $r_{ij} = y_j - \sum_{k \neq j} \beta_k x_{ik}$ $\hat{\beta}_j < -S_{lambda/2}(\frac{1}{N} x_j^T r_j)$ cyclical coordinate descent. we do this for j, and then repeat for j + 1 until convergenc. we get an interative algorithm. β_1, then β_2, ...

Because the truc is convex, this is going to converge to the minimum of the function $L(\beta)$.

cf note on this algorithm. If we understand the case for p=1, then we repeat, and because the problem is convex, we're going to converge to the single minimum.

On wednesday, we will see single classification.

Unsupervised learning: dimensionality reduction, PCA, SVD

Unsupervised learning: clustering, K-means, hierarchical

Neural networks, from single neuron to multilayer networks

Physics of machine learning, statistical mechanics of machine learning, applications