A Gentle Introduction to Data Science

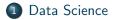
E. Le Pennec



DSE2017 - Brest - 04/07/2017







2 Some Data Science Challenges



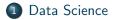


4 Mathematical Insights on Learning



Data Science





2 Some Data Science Challenges

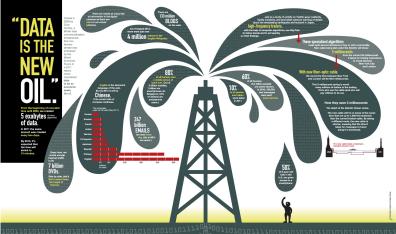




Data Is The New Oil?

Data Science





A New Context



Data everywhere

- Huge volume,
- Huge variety...

Affordable computation units

- Cloud computing
- Graphical Processor Units (GPU)...
- Growing academic and industrial interest!

Data Science



Major Influences

Four major influences act today:

- The formal theories of statistics
- Accelerating developments in computers and display devices
- The challenge, in many fields, of more and ever larger bodies of data
- The emphasis on quantification in an ever wider variety of disciplines

Data Science

Data Science



Major Influences - Tukey (1962)

Four major influences act today:

- The formal theories of statistics
- Accelerating developments in computers and display devices
- The challenge, in many fields, of more and ever larger bodies of data
- The emphasis on quantification in an ever wider variety of disciplines
- He was talking of Data Analysis.
- Data mining, Machine learning, Big Data...

Big Data Is (Quite) Easy

Data Science



Example of off the shelves solution

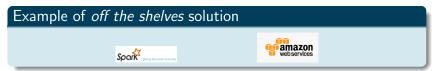




<pre>def run(params: Params) { val conf = new SparkConf() .setApAmed(*BinaryClassification with \$params") val sc = new SparkContext(conf)</pre>											
Logger.getRootLogger.setLevel(Level.WARN)											
<pre>val examples = MLUtils.loadLibSVMFile(sc, params.input).cache()</pre>											
<pre>vml splits = cxmples.randosGplit(Arroy(0.8, 0.2)) vml training = splits().cache() vml test = splits().cache() vml nmfraining = training.comt() vml nmfraining = training.comt() vml nmfraining = training.comt() vmmples.rmpraining.training.comt() vmmples.rmpraining.train</pre>											
examples.unpersist(blocking = Talse)											
<pre>val.updater = params.regType match { case L1 => new L1Updater() case L2 => new SquaredL2Updater() }</pre>											
<pre>via algorithm = now_iopiticRepressionWithSGD() algorithm.cotilatzer .setWaniTerations() .setTrajside(parmas.repSize) .setTrajside(parmas.repSize) .setTrajside(parmas.repSize) .setTrajside(parmas.repSize) .setTrajside(parmas.repSize) .setTrajside(parmas.repSize) </pre>											
<pre>val prediction = model.predict(test.map(features)) val predictionAndLabel = prediction.zip(test.map(label))</pre>											
<pre>val metrics = new BinaryClassificationMetrics(predictionAndLabel) val myMetrics = new MyBinaryClassificationMetrics(predictionAndLabel)</pre>											
<pre>println(s"Empirical CrossEntropy = \${myMetrics.crossEntropy()}.") println(s"Test areaUnderPR = \${metrics.areaUnderPR()}.") println(s"Test areaUnderROC = \${metrics.areaUnderROC()}.")</pre>											
sc.stop() }											

Big Data Is (Quite) Easy





```
export AWS_ACCESS_KEY_ID=<your-access-keyid>
export AWS_SECRET_ACCESS_KEY=<your-access-key-secret>
cellule/spark/ec2/sparl-ec2 -i cellule.pem -k cellule -s <number of machines> launch <cluster-name>
ssh -i cellule.pem root@<your-cluster-master-dns>
spark-ec2/copy-dir ephemeral-hdfs/conf
ephemeral-hdfs/bin/hadoop distcp s3n://celluledecalcul/dataset/raw/train.csv /data/train.csv
scp -i cellule.pem cellule/challenge/target/scala-2.10/target/scala-2.10/challenges_2.10-0.0.jar
cellule/spark/bin/spark-submit \
     --class fr.cc.challenge.Preprocess \
     challenges_2.10-0.0.jar \
     //data/train.csv \
```

```
/data/train2.csv
```

```
cellule/spark/bin/spark-submit \
    --class fr.cc.sparktest.LogisticRegression \
    challenges_2.10-0.0.jar \
    /data/train2.csv
```

\Rightarrow Logistic regression for arbitrary large dataset!

Doing Data Science

Data Science



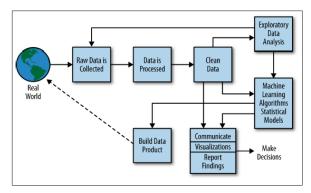


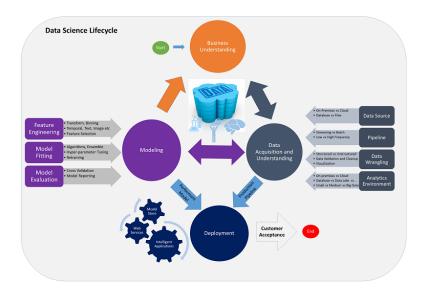
Figure 2-2. The data science process

- Doing Data Science: Straight talk from the frontline.
 - Rachel Schutt, Cathy O'Neil
 - O'Reilly

Data Science Is (Quite) Complex!

Data Science





Data Science Is (Quite) Complex!

Data Science



BIG DATA LANDSCAPE 2017



Last updated 4/5/2017

© Matt Turck (@mattturck), Jim Hao (@jimrhao), & FirstMark (@firstmarkcap) mattturck.com/bigdata2017

FIRSTMARK 📂

Data Science Is (Quite) Complex!

Data Science



The Periodic Table of Data Science

An overview of key companies, resources and tools in data science (as of 4/12/2017)

									Symbol	Dc										_
									Name	DeteCare									Ede	
_	_									Const. and	v								\$12 Nuggets	
Do		- 40																	-	10
DateCamp	General Assembly	Stata + Holoop		Connes		Data			Search & Data Management					Collaboratio			News, Newsleth	ers it Blogs		
				Sect campo		Projects & Challenges, Competitions			Machine Learning & State				Constanty	8 Q8 A	Pedcarb			Moggen	Hartlybes	
8		04		Conferences		Programming Languages & Distributions			Data Visualization & Reporting										Ba	
SpringBoard	Meth	OD9C																	HadarNess	
a a constante a		K Tablees	Pr .	N	13	76	8	Ab	v	Bed		5	Sm	15	OM	She	NI	De	Dec.	Dr
Edit	Data Involutor	Centerence															Demine Date	Data Science	Data Science	
	10		Python.	JanaScript	Visual Basic	Totp:93.	930.N		Weba	RgHL	. KI	inime	Spark MLD	Power \$1	Orade N	Shiny	Lib	Experience	Central	kendap
				G9	N					Der		LI6	Be	5e	AH	2461	N.	RS .	Daw	Q.
Ceations	Insight	UseR			Solu					DataBak	w 18	вум	H2O	Instruction 1	Aberra	Maplotib	News	Tende	Data Science Weekly	
Ola		N																	. Iv	
Udicity	NYC Data Science Academy	PrOsta	5						Max Name of SOL	1.m		Na	n	39	Ser SASYInul	rty -	3o	Ex Booker		
			SQL	Ped						RapidMa	ner Math	wantica	Thoma	Spotlar	Analytics	Plotty	Radao	Netbook	Data Elixár	Vieddy
Ofe		Per-		Mr		349		Acm		Ou		2540	Ani	91	74	26	Sev	и	RH	N
Wang	Galennier	Analytics North	Bub	Microsoft R Open						Cherry			Acture Machine Learning	Otheries	Perufitiat	Marcooft	Sander	Apache Zenevilo	8.99407	Perioly Derivative
ъ	Prg	NM.	1940	open						Chang	6 M.	201	Learning	GENER	MANDANS.	Enod	spyaa	zappeno		
Plankisht	Data faiseur fre Social Good	ACMINISTER Conference	Md	CY .						π		9	D	Co	Gdt	Pe	Data Science	- N	Rds coming a Data	
Pathage	Social Good	Completes	ALISA	Catopy						Tenorfs	w s	iana -	D3	Gogaos	Google Charp	Female	Statio	lapstar	Scientist	
- 10		TPN		44								Du	Mr				- 19	9	Di	DA.
Lynda	Data indicty	Tendota Partners Condension												Adobe			Databricks		Data Starting	
Th.	N	Ref.	Java	Anaconda	spork	Hite	II 54 T 62	Loone	RatiStarth	SikieLa	am Data1	Guptite 1	Microstrategy	Ambrio	Tableau	Robob.	antebook	cates		- and the
	Data Science	IEEE Incentional											_		_		_		La Linear	
TemilineHouse	Daja	Data Mining	Dw.								Re:	50		e q			~		Egration	
144			Dataworkl								Redda	Stack		tost datal Qu		Avites Data 1 Regis Stack I	icionee xchange			
Rig Data Conversity								16			264	Edu								
																		-	1 DataCarto	
			Statistic	Louring Dynakry	World Bank	Terreris	Bandenl	DataSted	DriverData		Monty EDunMining									e seeg



Some Data Science Challenges





2 Some Data Science Challenges

3 Data Scientists



New Interdisciplinary Challenges

Some Data Science Challenges



- Applied math AND Computer science
- Huge importance of domain specific knowledge: physics, signal processing, biology, health, marketing, environmental science...

Some joint math/CS/domain challenges

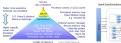
- Data acquisition
- Unstructured data and their representation
- Huge dataset and computation
- Visualization
- Software(s)
- Domain specific issue!

Some Challenges



Data Acquisition

- How to measure new things?
- . How to choose what to measure?
- · How to deal with distributed sensors?
- . How to look for new sources of informations?





Huge Dataset

- . How to take into account the locality of the data?
- · How to construct distributed architectures?
- . How to design adapted algorithms?

Software(s)

- · How to construct a consistent ecosystem?
- . How to construct interoperable systems?

Unstructured Data

- How to store efficiently the data?
- . How to describe (model) them to be able to process them?
- . How to combine data of different nature?



Visualization

- . How to look at the data?
- How to present results?
- . How to help taking better informed decision?



Domain Specific Knowledge

- . How to find the real problem at hand?
- · How to incorporate human expertise?
- . How to measure the performance?

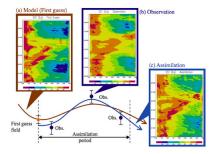




Some Challenges

Some Data Science Challenges





Environment Science

- Data/Model coupling.
- Multiscale modeling / Multimodal modeling.
- Long term/short term prediction.
- Prediction vs understanding.



Data Scientists



1 Data Science

2 Some Data Science Challenges



4 Mathematical Insights on Learning

Skills



MODERN DATA SCIENTIST

Data Scientist, the seeket job of the 21th century, requires a mature of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist i hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

- R packages like ggplot or lattice
 Knowledge of any of visualization tools e.g. Flare, D3 is, Tableau

MarketingDistillery.com is a arrow of practitioners in the area of a commerce marketing. Our fields of expertise include marketing strategy and optimization: customer tracking and on site analytics, predictive analytics and econometrics; data watehousing and big data systems: marketing channel insights in Paid Search, SED, Social, CRM and brand.



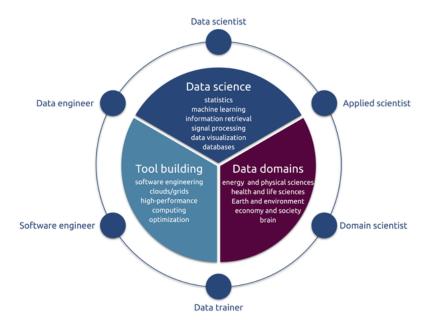
Data Scientist

- Mix of various skills.
- Hard to be an expert of everything!

More Than One Type Of Data Scientist!

Data Scientists

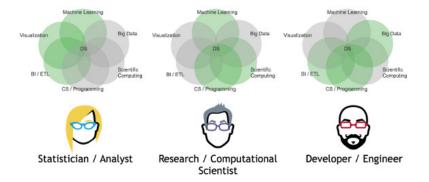




Data Science Team

Data Scientists





• Importance of balanced teams.







2 Some Data Science Challenges



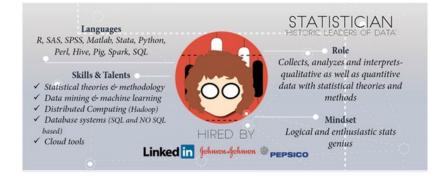


4 Mathematical Insights on Learning

Statistician Point of View

Mathematical Insights on Learning





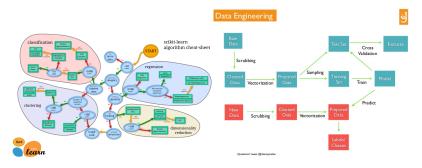
Disclaimer

- I'm a statistician with a signal processing background... posing as a data scientist.
- Not that different in the end...

ML in Practice

Mathematical Insights on Learning





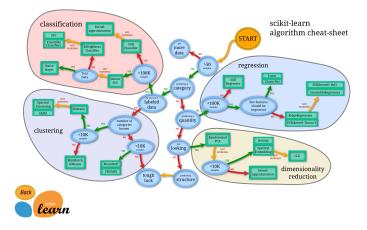
Practical ML

- Build models.
- Test and compare them.
- Use the *best* one...
- No uniformly better methods!
- Mathematical justification...

Machine Learning

Mathematical Insights on Learning





Several methodologies

- Lots of methods...
- Only two main principles...

Supervised Learning

Mathematical Insights on Learning



Experience, Task and Performance measure

- Training data : $\mathcal{D} = \{(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)\}$ (i.i.d. $\sim \mathbf{P}$)
- **Predictor**: $f : \mathcal{X} \to \mathcal{Y}$ measurable
- Cost/Loss function : $\ell(f(\mathbf{X}), Y)$ measure how well $f(\mathbf{X})$ "predicts" Y
- Risk:

 $\mathcal{R}(f) = \mathbb{E}\left[\ell(Y, f(\mathbf{X}))\right] = \mathbb{E}_{X}\left[\mathbb{E}_{Y|\mathbf{X}}\left[\ell(Y, f(\mathbf{X}))\right]\right]$

• Often $\ell(f(\mathbf{X}), Y) = \mathbf{1}_{Y \neq f(\mathbf{X})}$ or $\ell(f(\mathbf{X}), Y) = |f(\mathbf{X}) - Y|^2$

Goal

• Learn a rule to construct a classifier $\hat{f} \in \mathcal{F}$ from the training data \mathcal{D}_n s.t. the risk $\mathcal{R}(\hat{f})$ is small on average or with high probability with respect to \mathcal{D}_n .

Best Solution



X

• The best solution f^* (which is independent of \mathcal{D}_n) is $f^* = \arg\min_{f \in \mathcal{F}} R(f) = \arg\min_{f \in \mathcal{F}} \mathbb{E}\left[\ell(Y, f(\mathbf{X}))\right] = \arg\min_{f \in \mathcal{F}} \mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_{Y|\mathbf{X}}\left[\ell(Y, f(\mathbf{x}))\right]\right]$

Bayes Classifier (explicit solution)

• In binary classification with 0-1 loss:

$$f^*(\mathbf{X}) = \begin{cases} +1 & \text{if } \mathbb{P}\left(Y = +1 | \mathbf{X}\right) \geq \mathbb{P}\left(Y = -1 | \mathbf{X}\right) \\ & \Leftrightarrow \mathbb{P}\left(Y = +1 | \mathbf{X}\right) \geq 1/2 \\ -1 & \text{otherwise} \end{cases}$$

• In regression with the quadratic loss $f^*(\mathbf{Y}) = \mathbb{R}$

 $f^*(\mathbf{X}) = \mathbb{E}\left[Y|\mathbf{X}\right]$

Issue: Explicit solution requires to **know** $\mathbb{E}[Y|\mathbf{X}]$ for all values of \mathbf{X} !



Machine Learning

• Learn a rule to construct a classifier $\hat{f} \in \mathcal{F}$ from the training data \mathcal{D}_n s.t. the risk $\mathcal{R}(\hat{f})$ is small on average or with high probability with respect to \mathcal{D}_n .

Canonical example: Empirical Risk Minimizer

- Restrict f to a subset of functions $S = \{f_{\theta}, \theta \in \Theta\}$
- Replace the minimization of the average loss by the minimization of the empirical loss

$$\widehat{f} = f_{\widehat{\theta}} = \operatorname*{argmin}_{f_{\theta}, \theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f_{\theta}(\mathbf{X}_i))$$

- Examples:
 - Linear regression
 - Linear discrimination with

 $\mathcal{S} = \{\mathbf{x} \mapsto \operatorname{sign}\{\beta^T \mathbf{x} + \beta_0\} \ / \beta \in \mathbb{R}^d, \beta_0 \in \mathbb{R}\}$

Probality vs Optimization? How to find a good function f with a *small* risk $R(f) = \mathbb{E} \left[\ell(Y, f(X)) \right]$? Canonical approach: $\hat{f}_S = \operatorname{argmin}_{f \in S} \frac{1}{n} \sum_{i=1}^n \ell(Y_i, f(\mathbf{X}_i))$

Problems

- How to choose S?
- How to compute the minimization?

A Probabilistic Point of View

Solution: For X, estimate Y|X plug this estimate in the Bayes classifier: (Generalized) Linear Models, Kernel methods, *k*-nn, Naive Bayes, Tree, Bagging...

An Optimization Point of View

Solution: If necessary replace the loss ℓ by an upper bound ℓ' and minimize the empirical loss: SVR, SVM, Neural Network,Tree, Boosting

Probabilistic Approach





• If $Y | \mathbf{X}$ is known, one can compute the best solution f^* $\arg\min_{f \in \mathcal{F}} \mathbb{E}_{\mathbf{X}} \left[\mathbb{E}_{Y | \mathbf{X}} \left[\ell(Y, f(\mathbf{x})) \right] \right]$

Bayes Plugin

• Learning: Estimation of Y|x and pluging of this estimate in the Bayes classifier

• Plugin: a classifier
$$\widehat{f} : \mathcal{X} \to \mathcal{Y}$$

$$\ell^{0/1}$$
 loss: $\widehat{f}(\mathbf{x}) = egin{cases} +1 & ext{if } \widehat{p}_{+1}(\mathbf{x}) \geq \widehat{p}_{-1}(\mathbf{x}) \ -1 & ext{otherwise} \end{cases}$

• Quadratic loss:

$$\hat{f}(\mathbf{x}) = \mathbb{E}\left[Y|\mathbf{x}
ight]$$

• Instantiations:

- Generative Modeling and Bayesian Methods
- Parametric Conditional Models
- Kernel Conditional Density Methods
- Importance of a corresponding efficient numerical scheme!

Optimization Approach





• The best solution f^* is the one minimizing $f^* = \arg \min R(f) = \arg \min \mathbb{E} \left[\ell(Y, f(X)) \right]$

Empirical Risk Minimization

- Restrict f to a subset of functions $S = \{f_{\theta}, \theta \in \Theta\}$
- Replace the minimization of the average loss by the minimization of the empirical loss

$$\widehat{f} = f_{\widehat{\theta}} = \operatorname*{argmin}_{f_{\theta}, \theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f_{\theta}(x_i))$$

- Issue: Minimization may be impossible in practice.
- Solution: Replace ℓ by ℓ' a simpler (convex) majorant and minimize this upper-bound.
- Instantiation: Regression, SVM, Neural Networks...
- Importance of a corresponding efficient numerical scheme!

Probabilistic vs Optimization

Mathematical Insights on Learning



Probabilistic Approach

- **Principle:** estimate the **conditional law** *Y*|*X* and use it to take an **informed** decision.
- Motto: If you know the world, everything is easy!
- Emphasis on Interpretation
- Pro:
 - Interpretable models.
 - Lots of flexibility in the generative model.
 - Simultaneous decision optimization.
- Cons:
 - Computational issue.
 - No need to know the law to take a decision.

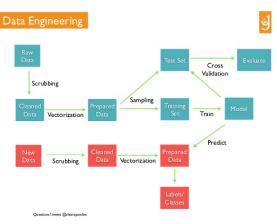
Optimization Approach

- Principle: construct a surrogate decision criterion and use it to take an optimized decision.
- Motto: You should focus on your goal!
- Emphasis on **Prediction**
- Pro:
 - Focus on the true goal!
 - Can use very clever optimization algorithm.
 - No need to obtain the best solution.
- Cons:
 - Black box model.
 - Not robust to a change of decision zone.

Model Validation

Mathematical Insights on Learning





Competition between methods

- Compare methods by their performance...
- on data not used to choose parameters! (Cross Validation)
- Use the best one in the end.

Bias-Variance Dilemna

Mathematical Insights on Learning



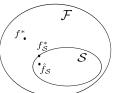
- General setting:
 - $\mathcal{F} = \{ \text{measurable fonctions } \mathcal{X} \to \mathcal{Y} \}$
 - Best solution: $f^* = \operatorname{argmin}_{f \in \mathcal{F}} \mathcal{R}(f)$
 - $\bullet~\mbox{Class}~\mathcal{S}\subset\mathcal{F}~\mbox{of functions}$
 - Ideal target in \mathcal{S} : $f_{\mathcal{S}}^* = \operatorname{argmin}_{f \in \mathcal{S}} \mathcal{R}(f)$
 - Estimate in \mathcal{S} : $\widehat{f}_{\mathcal{S}}$ obtained with a numerical algorithm

Approximation error and estimation error (Bias/Variance)

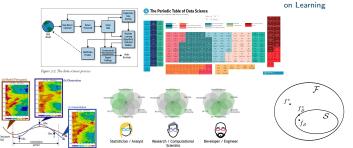
$$\mathcal{R}(\widehat{f}_{\mathcal{S}}) - \mathcal{R}(f^*) = \underbrace{\mathcal{R}(f_{\mathcal{S}}^*) - \mathcal{R}(f^*)}_{\mathcal{R}(f^*)} + \underbrace{\mathcal{R}(\widehat{f}_{\mathcal{S}}) - \mathcal{R}(f_{\mathcal{S}}^*)}_{\mathcal{R}(f^*)}$$

Approximation error Estimation error
 Different behavior for different model complexity

- Low complexity model are easily learned but the approximation error ("bias") may be large (Under-fit).
- High complexity model may contains a good ideal target but the estimation error ("variance") can be large (**Over-fit**)



Conclusion



- Data Science is **not** a new thing.
- Big Data: easier and easier ability to deal with large dataset.
- Environment science: coupling complex modeling and data is the key!
- Importance of collaboration (and team) in Data Science.
- Practical insights can be learned from theory.

Mathematical Insights on Learning