Statistical Inference vs. Machine Learning: Why can't we be friends?

Sahir Bhatnagar, Marie Forest, Julyan Keller-Baruch

July 20, 2017

LDI Journal Club

Motivation

This Guy

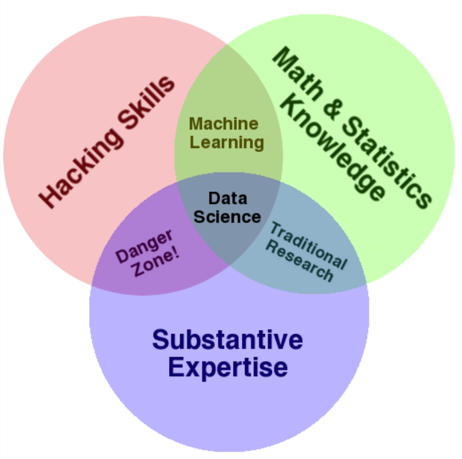
Two Cultures

Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman

Data Science Venn Diagram[1]

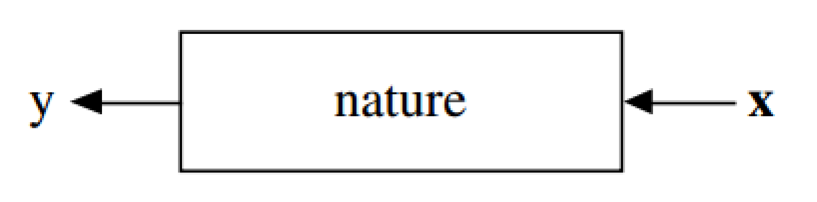


[1] http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Statistical Learning

Nature Functions to Associate X with y[2]

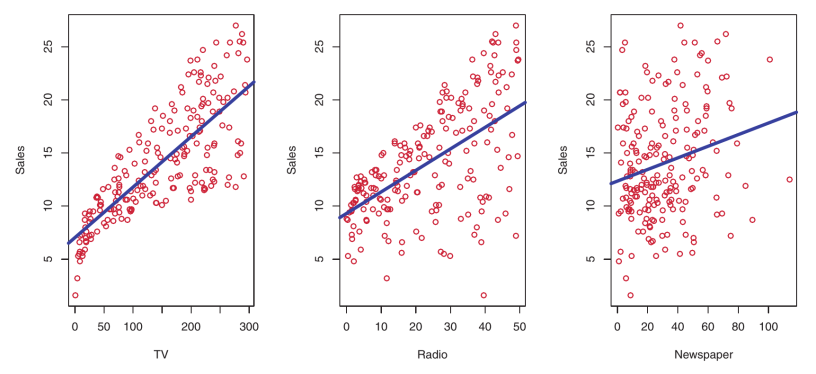
- A matrix of input variables **X** go in one side
- On the other side, response variable **y** comes out



[2] Breiman, Leo. *Statistical modeling: The two cultures*. Statistical science (2001)

Example: Advertising Data Set

- y: `sales` of a product in 200 different markets (*response or dependent variable*)
- $\mathtt{X} = (X_1, X_2, X_3)$: advertising budgets for `TV, radio, newspaper` (*predictors, independent variables, features*)



Model Set-up

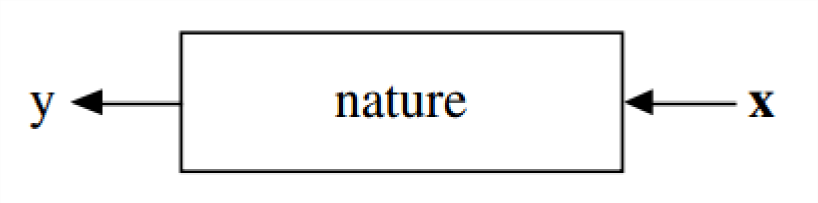
- We observe a quantitative response Y and p different predictors $X = (X_1, \ldots, X_p)$
- ε is an error term independent of X with mean 0
- We assume there is some relationship between Y and X:

Model Set-up

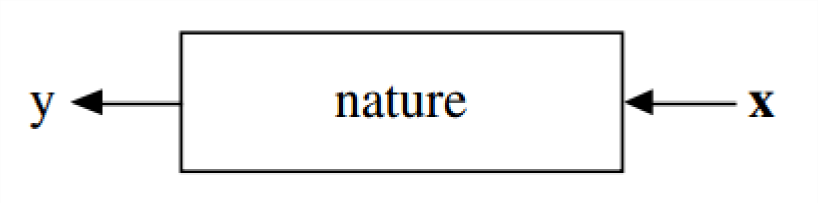
- We observe a quantitative response Y and p different predictors $X = (X_1, \ldots, X_p)$
- ε is an error term independent of X with mean 0
- We assume there is some relationship between Y and X:

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

What is f?



What is f ?



- f is *nature*

What is f

- f is a function that connects X to y and is generally unknown

What is f

- f is a function that connects X to y and is generally unknown
- In this situation, one must estimate f based on observed points

What is f

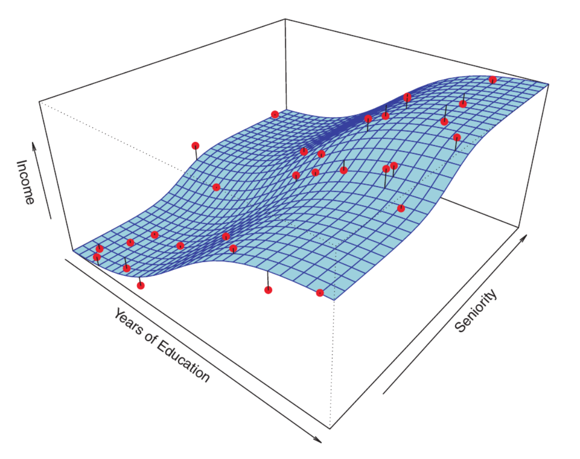
- f is a function that connects X to y and is generally unknown
- In this situation, one must estimate f based on observed points
- $\widehat{f}$ denotes our estimate to f

Exercise: Income dataset

- `income` is the response

Exercise: Income dataset

- `income` is the response
- $\widehat{f}(x) = 1.4 \times \text{Years of Education} + 0.56 \times \text{Seniority}^2$



Statistical Learning refers to a set of approaches for estimating f

Two Cultures: Prediction and Inference

Author of Random Forests:

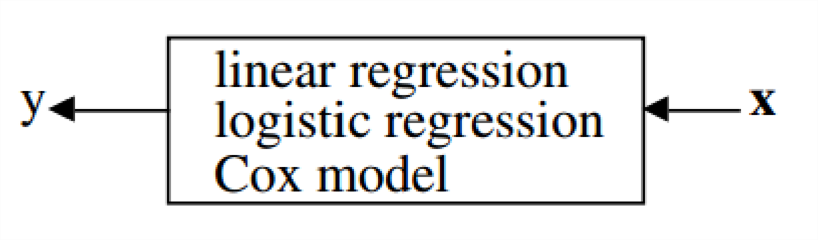
Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman

Data Modelling Culture[3]

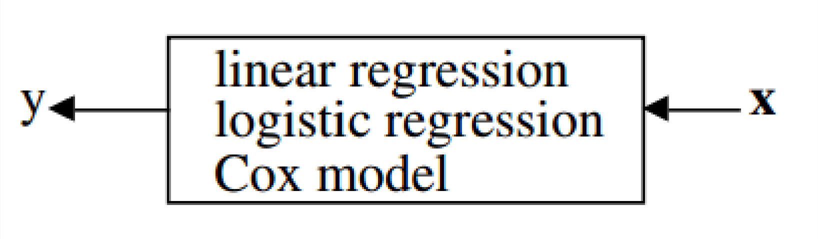
- Starts with assuming a stochastic data model for the inside of the box (e.g. normal, binomial)
- Values of the parameters are estimated from the data
- Information (ORs, HRs)



[3] Breiman, Leo. *Statistical modeling: The two cultures.* Statistical science (2001)

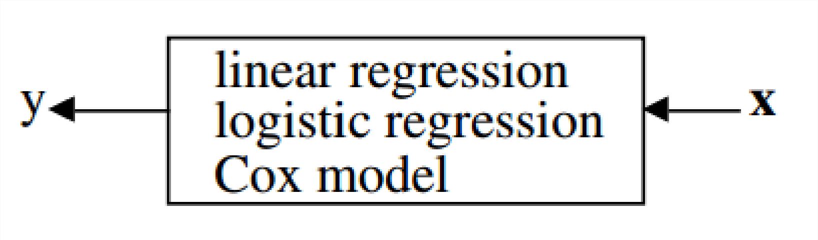
Data Modelling Culture Criticisms

- The belief that by imagination and by looking at the data, can invent a good parametric model for a complex mechanism devised by nature



Data Modelling Culture Criticisms

- The belief that by imagination and by looking at the data, can invent a good parametric model for a complex mechanism devised by nature



- Conclusions are about the model's mechanisms, not nature's
- If the model is a poor emulation of nature, the conclusions maybe wrong

The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: http://amstat.tandfonline.com/loi/utas20

The ASA's Statement on p-Values: Context, Process, and Purpose

Ronald L. Wasserstein & Nicole A. Lazar

The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: http://amstat.tandfonline.com/loi/utas20

The ASA's Statement on p-Values: Context, Process, and Purpose

Ronald L. Wasserstein & Nicole A. Lazar

- *P*-values do not measure the probability that the studied hypothesis is true

Data Modelling Culture Criticisms: Reliance on *P*-values

The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: http://amstat.tandfonline.com/loi/utas20

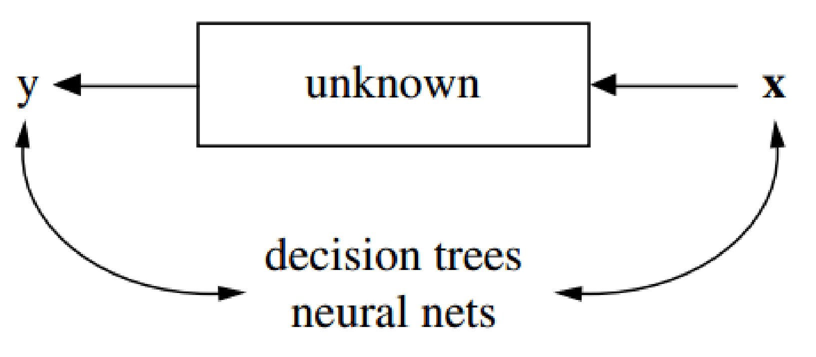
The ASA's Statement on p-Values: Context, Process, and Purpose

Ronald L. Wasserstein & Nicole A. Lazar

- *P*-values do not measure the probability that the studied hypothesis is true - *P*-values can indicate how incompatible the data are with a specified statistical model

Algorithmic Modelling Culture[4]

- Considers the inside of the box **complex** and **unknown**
- Find an algorithm that operates on X to predict y
- Prediction



[4] Breiman, Leo. *Statistical modeling: The two cultures*. Statistical science (2001)

Algorithmic Modelling Culture Criticisms[5]

[5] XKCD comic

Why Estimate f ?

Two Reasons

Prediction

- Machine Learning, Neural Nets, Support Vector Machines, Random Forests

Two Reasons

Prediction

- Machine Learning, Neural Nets, Support Vector Machines, Random Forests

Inference

- Linear, Logistic, Cox Regression

Prediction

Prediction

- In many situations, a set of inputs X are readily available, but the output Y cannot easily be obtained
- We can predict Y using

Prediction

- In many situations, a set of inputs X are readily available, but the output Y cannot easily be obtained
- We can predict Y using

$$\widehat{Y} = \widehat{f}(X)$$

Prediction

- In many situations, a set of inputs X are readily available, but the output Y cannot easily be obtained
- We can predict Y using

$$\widehat{Y} = \widehat{f}(X)$$

- $\widehat{f}$ is often treated as a black box $\rightarrow$ not really concerned with the exact form of $\widehat{f}$ as long as it predicts well.

Prediction

- In many situations, a set of inputs X are readily available, but the output Y cannot easily be obtained
- We can predict Y using

$$\widehat{Y} = \widehat{f}(X)$$

- $\widehat{f}$ is often treated as a black box $\rightarrow$ not really concerned with the exact form of $\widehat{f}$ as long as it predicts well.
- Example: Refer to Brent's talk on deep neural networks and putting pathologists out of work

_computational BIOLOGY

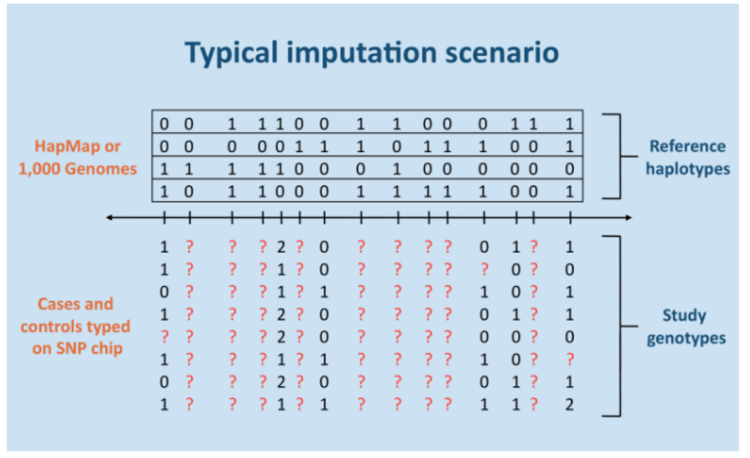
nature biotechnology

ANALYSIS

OPEN

Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues

Jason Ernst[1–5] & Manolis Kellis[6,7]

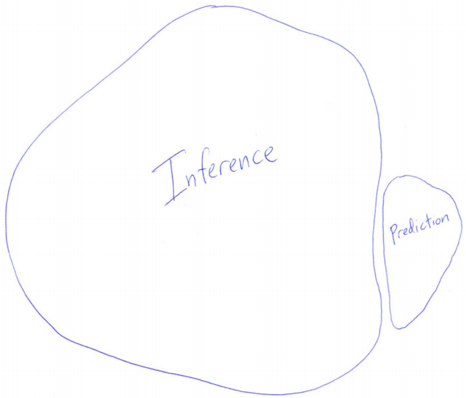

Typical imputation scenario
HapMap or
1,000 Genomes
0 0 1 1 1 0 0 1 1 0 0 0 1 1 1
0 0 0 0 0 1 1 1 0 1 1 1 0 0 1
1 1 1 1 1 0 0 0 1 0 0 0 0 0 0
1 0 1 1 0 0 0 1 1 1 1 1 0 0 1
Reference
haplotypes
Cases and
controls typed
on SNP chip
1 ? ? ? 2 ? 0 ? ? ? ? 0 1 ? 1
1 ? ? ? 1 ? 0 ? ? ? ? ? 0 ? 0
0 ? ? ? 1 ? 1 ? ? ? ? 1 0 ? 1
1 ? ? ? 2 ? 0 ? ? ? ? 0 1 ? 1
? ? ? ? 2 ? 0 ? ? ? ? 0 0 ? 0
1 ? ? ? 1 ? 1 ? ? ? ? 1 0 ? ?
0 ? ? ? 2 ? 0 ? ? ? ? 0 1 ? 1
1 ? ? ? 1 ? 1 ? ? ? ? 1 1 ? 2
Study
genotypes

Inference

Inference

- We are often interested in understanding the way that Y is affected as $X_1, \ldots, X_p$ change
- Which predictors are associated with the response
- $\widehat{f}$ can no longer be treated as a black box
- Examples: GWAS, EWAS

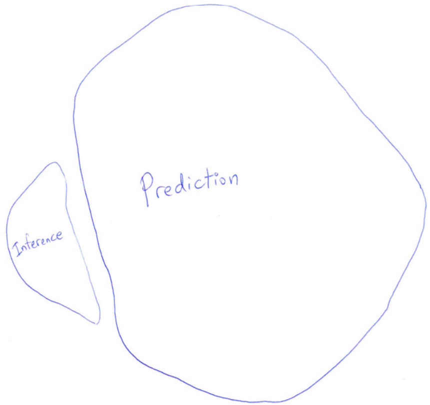
Statistics vs. Machine Learning[6]



How statisticians see the world?

[6] source: http://statweb.stanford.edu/ tibs/ftp/nips2015.pdf

Statistics vs. Machine Learning[7]



How machine learners see the world?

[7] source: http://statweb.stanford.edu/ tibs/ftp/nips2015.pdf

Case Study

UK Biobank

biobank uk

Improving the health of future generations

The Data

- GWAS for Z-Score BMD conducted by John Morris et. al (*Nature Genetics 2017*)
- Covariates: Age, Sex, PC1, PC2, PC3, PC4
- GWAS Hits: 301 Conditionally Independent SNPs
- Control SNPs: 10k SNPs (LD < 0.10 with lead SNPs)
- Dosages were calculated from `BGEN` files using `qctool`
- Training Data: $\sim$ 115k Observations (80%)
- Test Data: $\sim$ 30k Observations (20%)

The Data

- GWAS for Z-Score BMD conducted by John Morris et. al (*Nature Genetics 2017*)
- Covariates: Age, Sex, PC1, PC2, PC3, PC4
- GWAS Hits: 301 Conditionally Independent SNPs
- Control SNPs: 10k SNPs (LD < 0.10 with lead SNPs)
- Dosages were calculated from `BGEN` files using `qctool`
- Training Data: $\sim$ 115k Observations (80%)
- Test Data: $\sim$ 30k Observations (20%)

```
pryr::mem_used()
7.08 GB
```

Finding SNPs in low LD with lead SNPs

- Not looking for tag SNPs.
- Used a plink function to find all genotyped SNPs with LD > 0.1 with at least one of the 301 lead SNPs.
- `–r2 inter-chr –ld-window-r2 0.1 –ld-snp-list leadSNPs`
- Randomly selected 10K SNPs from all available excluding the ones found in the previous step.

The Data

	ID_1	zbmd	PC1	PC2	PC3	PC4	age	sex	rs139603701	rs2708632	rs75077113
1	2610781	-1.255	-7.42628	-3.3967600	-4.6292400	3.19530000	63	0	0.00000000	1.0000000	0.00000000
2	4114347	-0.339	-6.45636	-0.5157170	-0.3762300	-1.28798000	65	0	0.00000000	1.0000000	0.07400510
3	4399930	-0.600	-6.74236	-1.2213700	-2.5755200	4.34368000	66	1	0.00000000	1.0000000	0.00000000
4	2081319	0.809	-6.05389	-0.4911940	-2.0864600	5.37939000	48	0	0.00000000	1.0000000	0.00000000
5	1347380	0.279	-5.52867	-1.2864200	-3.2811600	0.45690000	53	0	0.00000000	2.0000000	1.00000000
6	3262449	-0.421	-6.67005	-0.9253650	-0.4240550	10.36690000	66	1	0.00000000	2.0000000	0.00100708
7	4870063	-0.454	-8.41719	-1.4366200	-1.6072600	-1.29085000	43	1	0.00000000	2.0000000	0.99700900
8	1141212	1.383	-8.73178	-1.6209500	-2.6827300	-2.02636000	65	1	0.00000000	2.0000000	0.02301030
9	2997954	-2.290	-5.55814	-0.4710770	-1.3477500	-0.09621100	68	1	0.00000000	1.0000000	0.00100708
10	5805218	2.289	-9.22547	-3.7621800	-1.2818900	2.37689000	63	1	0.04199220	1.0000000	0.02499390

Methods

Domain	Method	Interpretable	Feature
Stat	`lm`	✓	No selection, p-values
	`ridge`	✓	No selection, shrinks coefficients
	`lasso`	✓	Variable selection and shrinkage
	`enet`	✓	Variable selection and shrinkage
ML	`Random Forest`	✓	Variable importance
	`Neural Net`	✗	Lots of choices to make

Models

1. `zbmd ~ age + sex + PC1 + PC2 + PC3 + PC4`

Models

1. `zbmd ~ age + sex + PC1 + PC2 + PC3 + PC4`
2. `zbmd ~ age + sex + PC1 + PC2 + PC3 + PC4 + 301 lead SNP`

Models

1. `zbmd ~ age + sex + PC1 + PC2 + PC3 + PC4`
2. `zbmd ~ age + sex + PC1 + PC2 + PC3 + PC4 + 301 lead SNP`
3. `zbmd ~ age + sex + PC1 + PC2 + PC3 + PC4 + 301 lead SNP + 300 Control`

Models

1. `zbmd ~ age + sex + PC1 + PC2 + PC3 + PC4`
2. `zbmd ~ age + sex + PC1 + PC2 + PC3 + PC4 + 301 lead SNP`
3. `zbmd ~ age + sex + PC1 + PC2 + PC3 + PC4 + 301 lead SNP + 300 Control`
4. `zbmd ~ age + sex + PC1 + PC2 + PC3 + PC4 + 301 lead SNP + 1k Control`

Models

1. `zbmd ~ age + sex + PC1 + PC2 + PC3 + PC4`
2. `zbmd ~ age + sex + PC1 + PC2 + PC3 + PC4 + 301 lead SNP`
3. `zbmd ~ age + sex + PC1 + PC2 + PC3 + PC4 + 301 lead SNP + 300 Control`
4. `zbmd ~ age + sex + PC1 + PC2 + PC3 + PC4 + 301 lead SNP + 1k Control`
5. `zbmd ~ age + sex + PC1 + PC2 + PC3 + PC4 + 301 lead SNP + 5k Control`

Models

1. `zbmd ~ age + sex + PC1 + PC2 + PC3 + PC4`
2. `zbmd ~ age + sex + PC1 + PC2 + PC3 + PC4 + 301 lead SNP`
3. `zbmd ~ age + sex + PC1 + PC2 + PC3 + PC4 + 301 lead SNP + 300 Control`
4. `zbmd ~ age + sex + PC1 + PC2 + PC3 + PC4 + 301 lead SNP + 1k Control`
5. `zbmd ~ age + sex + PC1 + PC2 + PC3 + PC4 + 301 lead SNP + 5k Control`
6. `zbmd ~ age + sex + PC1 + PC2 + PC3 + PC4 + 301 lead SNP + 10k Control`

```r
lm(lcavol ~ ., data=Prostate)
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -2.260      1.260    -1.8     0.08 .
## lweight         -0.073      0.174    -0.4     0.68
## age              0.023      0.011     2.1     0.04 *
## lbph            -0.087      0.058    -1.5     0.14
## svi             -0.154      0.254    -0.6     0.55
## lcp              0.367      0.082     4.5    2e-05 ***
## gleason          0.191      0.154     1.2     0.22
## pgg45           -0.007      0.004    -1.7     0.10
## lpsa             0.573      0.086     6.7    2e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0
```

Penalized Model (Lasso, Elastic Net, Ridge)

```r
cv.glmnet(X, Y)
```

```
## 9 x 1 sparse Matrix of class "dgCMatrix"
##                     1
## (Intercept) 0.3708615
## lweight     .
## age         .
## lbph        .
## svi         .
## lcp         0.2293733
## gleason     .
## pgg45       .
## lpsa        0.4116749
```

Machine Learning Model: Random Forest

- Used the R package `ranger`, which allows parallel computing.
- Build 1000 trees (using 5,000 crashed with 10,000 control SNPs)
- Used the distribution of variable importance measures to perform variable selection.
- No need to normalize.

```
> pred_rf_mod3
Ranger prediction

Type:                             Regression
Sample size:                      29139
Number of independent variables:  607
> rf_mod3
Ranger result

Call:
 ranger(formula = mod3_fmla, data = ukb_train, importance = "impurity")

Type:                             Regression
Number of trees:                  500
Sample size:                      116571
Number of independent variables:  607
Mtry:                             24
Target node size:                 5
Variable importance mode:         impurity
OOB prediction error (MSE):       0.9209093
R squared (OOB):                  0.08068676
```

Figure 1: Output

Random Forest: Variable Importance

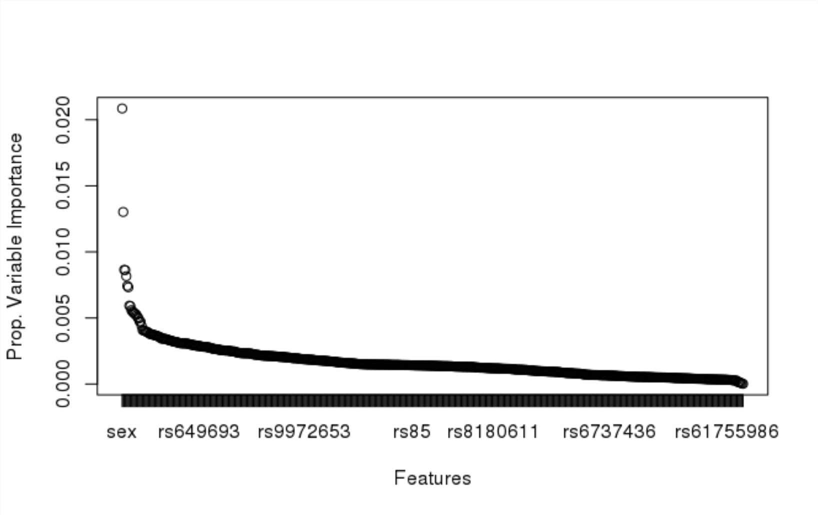


Figure 2: Variable importance

MLM: Support vector machine

- Tried two R packages without success (e1071 with function svm, and parallelSVM).
- It seems that we have too many observations (svm might be more suited when $N << p$)

MLM: Deep Neural Network

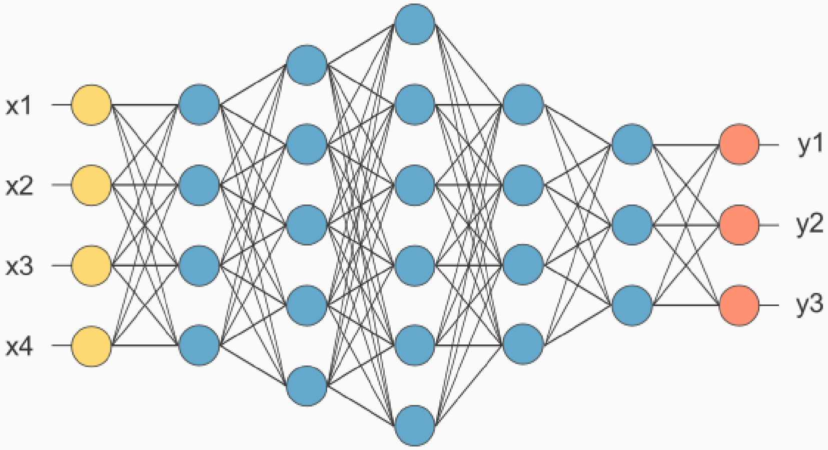
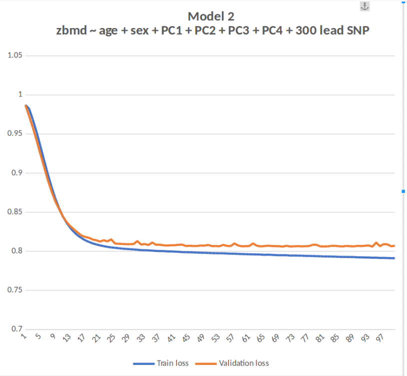
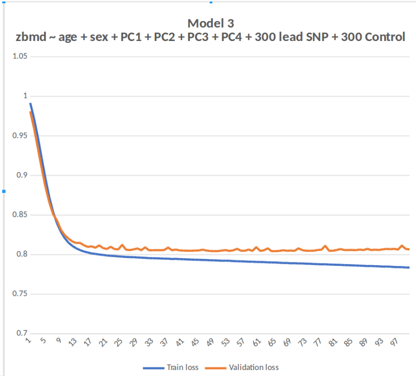


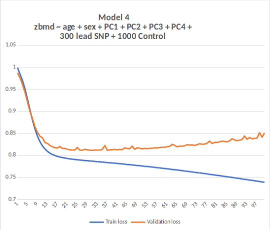
Figure 3: Deep Neural Network

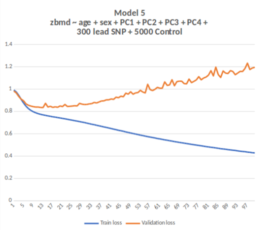
Deep Neural Network (keras + Theano)

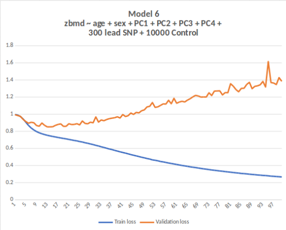




Deep Neural Network (keras + Theano)







Results: R2

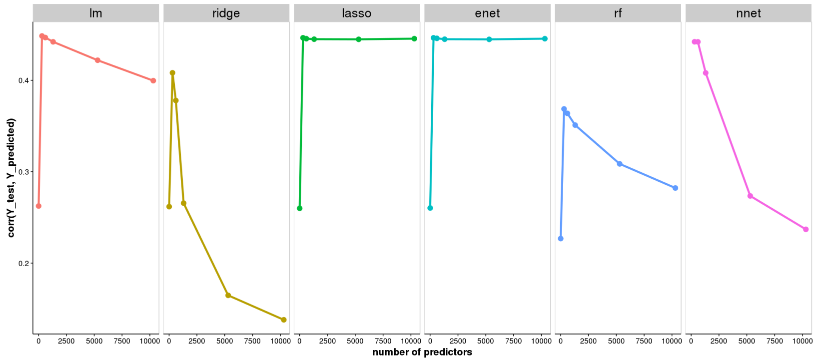


Figure 4: R2 vs. No. of Predictors

Results: Sensitivity vs. Specificity

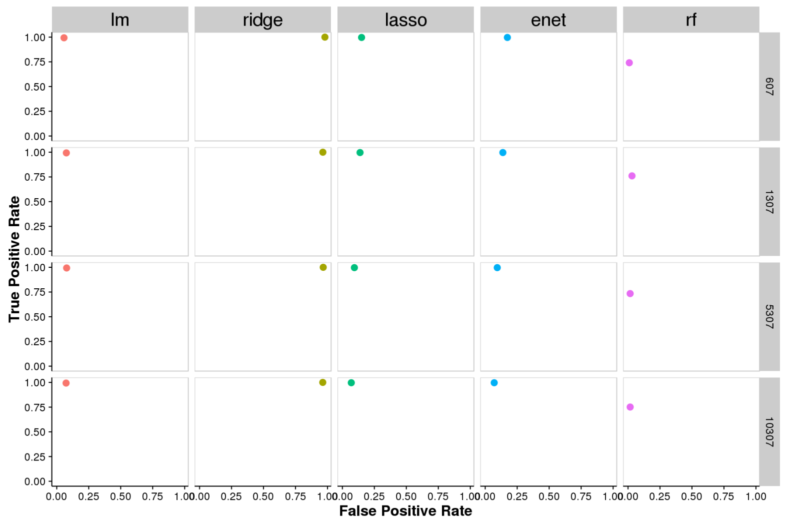


Figure 5: True Positive Rate vs. False Positive Rate

Results: Timings

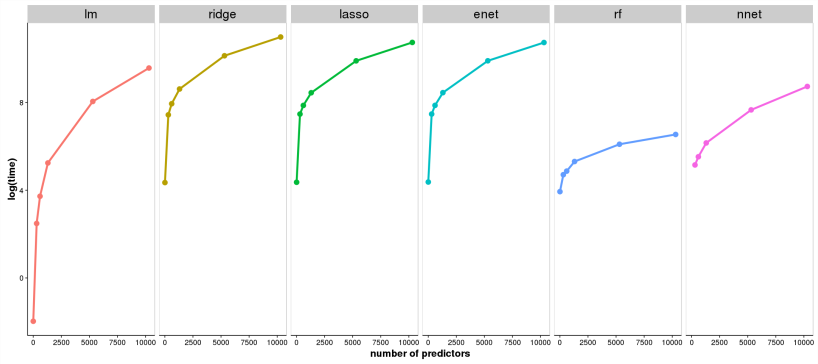


Figure 6: Time vs. No. of Predictors

Three Lessons to be Learned: Rashomon, Occam and Bellman

Rashomon

- Japanese movie (1950) in which four people, from different vantage points, witness a murder and a rape.
- When testifying, they all report the same facts, but their stories of what happened are very different

Rashomon: Many models give same $\hat{y}$, but tell different story

Model 1

$$y = 2.1 + 3.8x_3 - 0.6x_8 + 83.2x_{12}$$

Model 2

$$y = -8.9 + 4.6x_5 + 0.01x_6 + 12x_{15}$$

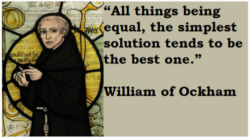
Model 3

$$y = -76.7 + 9.3x_2 + 22x_7 - 13.2x_8$$

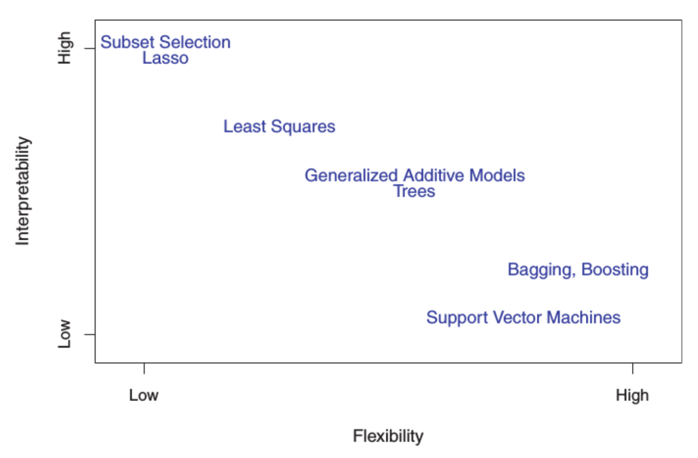
- All produce test set error within 1% of each other
- Which one is better?

Occam

- Unfortunately, in prediction, accuracy and simplicity (interpretability) are in conflict



Trade-off b/w Prediction Accuracy & Model Interpretability[8]



[8] Introduction to Statistical Learning

Bellman: Curse of Dimensionality

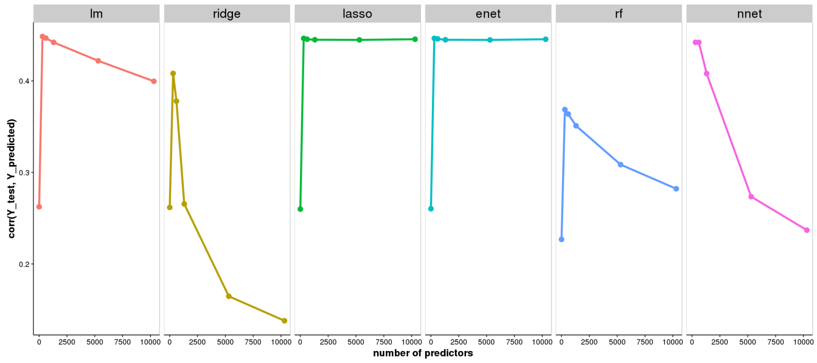


Figure 7: R2 vs. No. of Predictors

Final Thoughts

Take Home Message #1

- Ignore these terms: Machine Learning, Big Data, Statistical Learning

- Ignore these terms: **Machine Learning, Big Data, Statistical Learning**
- Focus on the input and output of a method

Take Home Message #1

- Ignore these terms: Machine Learning, Big Data, Statistical Learning
- Focus on the input and output of a method
- Ignore the arbitrary classification it falls under

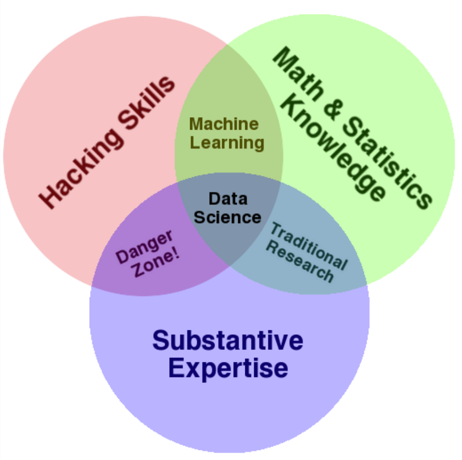
Take Home Message #2

- Focus on the goal of the study
- Inference
- Prediction
- or Both

- Inference and Prediction are both very challenging

Take Home Message #3

- Stay away from the Danger Zone



References

ISL

www-bcf.usc.edu/~gareth/ISL/

An Introduction to Statistical Learning

with Applications in R

Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani

Home

About this Book

R Code for Labs

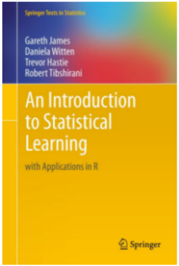
Data Sets and Figures

ISLR Package

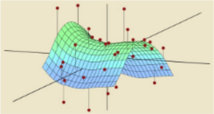
Get the Book

Author Bios

Errata



Download the book PDF
(corrected 6th printing)



Statistical Learning MOOC covering the entire ISL book offered by Trevor Hastie and Rob Tibshirani. Start anytime in self-paced mode.

Leo Breiman (1928 - 2005)

Appendix

How accurate is $\widehat{Y}$

- The accuracy of $\widehat{Y}$ depends on two quantities

$$mean(Y - \widehat{Y})^2 = \underbrace{[f(X) - \widehat{f}(X)]^2}_{reducible} + \underbrace{Variance(\varepsilon)}_{Irreducible}$$

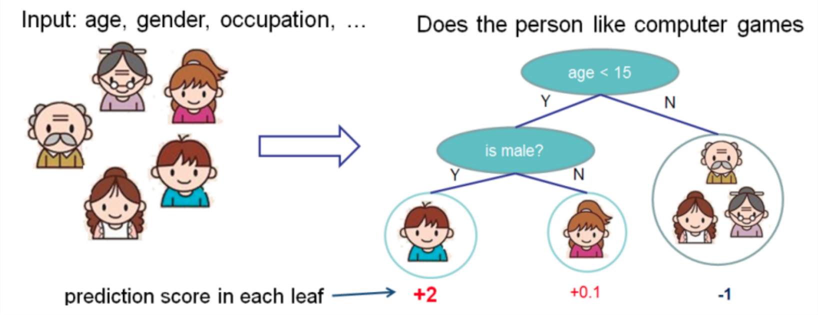
How accurate is $\widehat{Y}$

- The accuracy of $\widehat{Y}$ depends on two quantities

$$mean(Y - \widehat{Y})^2 = \underbrace{[f(X) - \widehat{f}(X)]^2}_{reducible} + \underbrace{Variance(\varepsilon)}_{Irreducible}$$

- Goal of most learning techniques is to estimate f with the aim of minimizing the reducible error

Trees

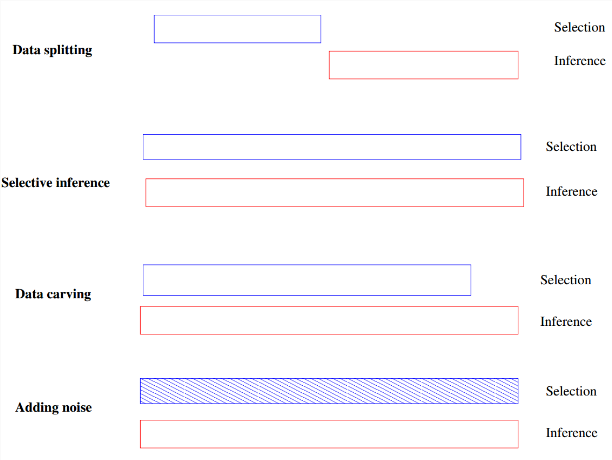

Input: age, gender, occupation, ...
Does the person like computer games
age < 15
Y
N
is male?
Y
N
prediction score in each leaf
+2
+0.1
-1

How do we Estimate f?

- Linear model

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

Inference Techniques[9]



[9] source: http://statweb.stanford.edu/ tibs/ftp/nips2015.pdf