Big Data, Mining and Learning: Perspectives for Social Data

Walter Sosa-Escudero

Seminario CEDLAS/IDRC-Canada, UNLP

May 16, 2015

Out with every theory of human behavior, from linguistics to sociology. Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves.

Chris Anderson, Wired, 23/6/2008

We may one day get to the point where sufficient quantities of big data can be harvested to answer all of the social questions that most concern us. I doubt it though. There will always be digital divides; always be uneven data shadows; and always be biases in how information and technology are used and produced.

Mark Graham, The Guardian, 9/3/2012

Jargon: NoSQL, Hadoop, mining, learning, visualization, fat models, loss, bayes risk, LASSO, CART, GARROTE, overfitting, training data set, supervised/unsupervised learning, cross validation, trees, forest, pruning, nodes, clusters, boosting, bagging, confusion matrix, ROC, curve, regularization, shrinkage, bayesian, model uncertainty, model averaging, reproducible error, out-of-sample prediction, basis, splines, GAM, support vector machines, subset selection, scrapping, networks, phyton, predictive analytics.

Problems

- Decide if an incoming email is spam.
- Identify numbers in a handwritten ZIP code.
- Measure the price of a good.
- Assign cancer patients to treatment.
- Identify factors that might lead to poverty.
- Find the name of a song from just humming of whistling.
- Translate a text.
- Predict preferences for income redistribution before voting.
- Find the middle class.
- Recommend movies or books.

What's in a name

- Learn, mine, predict, find patterns, classify, reduce dimensionality, visualize, summarize, decide.
- Handle, move, organize, store, retrive, explore, massive data.
- Data mining, statistical learning, machine learning, predictive analytics.

A simple predictive paradigm

$$Y = f(X) + u$$

- Goal: predict Y based on X, without observing u and not knowing what $f(.)$ is ('learn' f).
- Success: minimize out-of-sample prective performance.
- Training data: used for estimation.
- Test data: used for validation and evaluation.
- What matters is predictive performance in the *test* data.

- Decide what f can be.
- Learn (estimate) f. Go back to previous step. Iterate.
- Decide the dimension of X.

Econometricians be careful: the goal is out-of-sample prediction.

Measuring (lack of) success

Expected Loss function: $E\left[L(Y,\hat{f}(X))\right]$.

Expected squared loss:

$$E\left[(Y-\hat{f}(X))^2\right]$$

Expectes squared loss in the test data:

$$E\left[(Y-\hat{f}(X))^2 \mid \mathcal{T}\right]$$

Bias-Variance Decomposition

$$\begin{aligned} \text{Err}(x_0) &= E\big[(Y - \hat{f}(x_0))^2 \mid X = x_0\big] \quad (1) \\ &= \text{Bias}^2 \hat{f}(x_0) + \text{Var}\ \hat{f}(x_0) + \sigma_\epsilon^2 \quad (2) \end{aligned}$$

- σ^2 is the *irreductible error*
- Trade-off: the more complex $\hat{f}$ is the lower the squared bias but the higher the variance.
- The unbiased estimation paradigm *does not* solve the problem, it just focuses it.
- Small biases might lead to dramatic reduction in variance. Opens door for biased estimation (remember estimation is a means not an ends).

Cross validation

How to assess $Err(X \mid \mathcal{T})$?. One important stragegy is *k-fold cross-validation*.

Sometimes it is difficult to separate a test data set $\mathcal{T}$

- Split data into K equal sized-parts.
- Fit model leaving out one of the partitions.
- Compute the prediction error for the data left out previously.
- Do it for $k = 1, \dots, K$.

The cross-validation estimate of the squared prediction error is

$$CV(\hat{f}) = \frac{1}{N}\left(y_i - \hat{f}_{-k}(x_i)\right)^2,$$

$\hat{f}_{-k}(x_i)$ is the estimate with the $k-th$ part removed.

- How to choose K?: Standard wisdom: 5, 10. What about N ('leave-one-out)?. Variance-bias trade off again: better estimates, variable prediction error.
- CV for model choice: consider $f(x, \alpha)$ where α parametrizes the complexity of the model (number of variables, knots in splines, bandwitdh in non parametrics, etc.). Then: compute CV for alternative α's and minimize over α

$$CV(\hat{f}, \alpha) = \frac{1}{N}\Big(y_i - \hat{f}_{-k}(x_i, \alpha)\Big)^2,$$

Models for continuous Y

All the popular toys in econometrics

- Linear multiple regression.
- Polynomials, dummys, interactions.
- Non parametric regression. Splines.
- The curse of multidimensionality.

Model choice

Big data: the dimension of X can be ridiculously high. Number of columns might exceed number of rows (fat models).

Rememeber: goal is to measure out of sample performance.
Problem: within sample peformance usually understimates out of sample predction error grossly.

Example: R^2 is non decreasing *within the sample*.

Three approaches.

- Agree and base evaluation on test data (usually difficult).
- Cross validation estimate of mean squared prediction error.
- Modify standard within sample strategies: Adjusted R^2, BIC, Mallows C_p, AIC.

Three strategies

- Subset selection: global, forward, backward
- Shrinkage: ridge, lasso
- Other: principal components, least angle, partial least squares, etc.

Subset selection

- Global: estimate every possible model and choose best based on test data, CV or other criteria. Problem: computationally impossible
- Forward/backward: start simple, increase variables one at a time based on highest improvement in criteria. Choose complexity based on criteria.

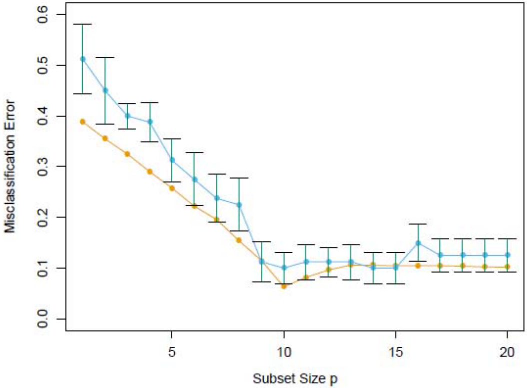


FIGURE 7.9. *Prediction error (orange) and tenfold cross-validation curve (blue) estimated from a single training set, from the scenario in the bottom right panel of Figure 7.3.*

Source: Hastie, Tibshirani and Friedman (2008)

Shrinkage methods

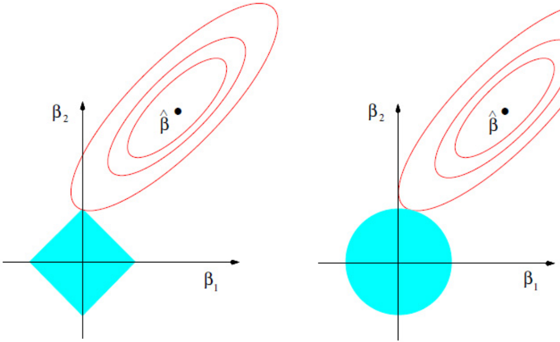
- Ridge

$$\min_{\beta} \sum_{i=1}^{n} e_i^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

- Lasso

$$\min_{\beta} \sum_{i=1}^{n} e_i^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

λ penalizes setting parameters away from zero (including variables?). They look similar but what they end up doing is completely different.



Source: Hastie, Tibshirani and Friedman (2008)

- Ridge was initially invented to deal with multicollinearity. It is a way to bias estimates to produce a variance gain.
- Both shrink towards zero. Lasso sets many at zero: ‘smooth subset seleccion’ (Least Absolute Shrinkage and Selection Operator).
- Both methods have a strong Bayesian feel.
- How to choose λ: cross validation.

Classification

Y is binary (spam or not, poor or not, etc.)

- Econometrics: estimate $Pr(Y = 1 \mid X)$. In particular, marginal effects.
- Data mining: predict Y.

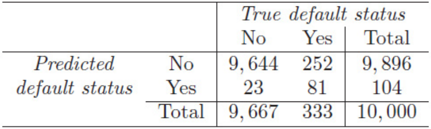
Models:

- LPM, logits, probits (not popular).
- Linear discriminant analysis.
- $K-$ nearest neighboors. Majority voting rule.
- CART's (more later)

Bayes classifier: predict outcome with highest probability (1 if $\hat{p} > 0.5$). Minimizes Bayes risk. Choice of threshold is not obvious. Bayes risk might be irrelevant.

Measuring success

Confusion matrix (example):



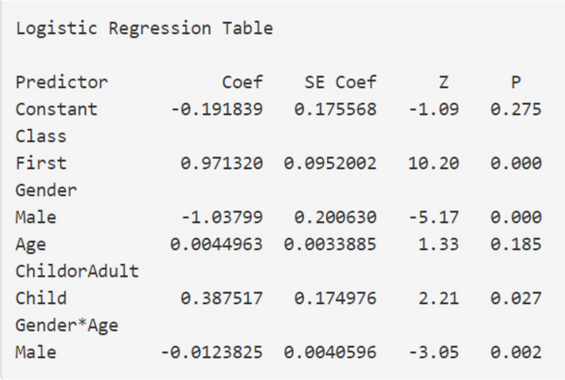
		True default status		
		No	Yes	Total
Predicted *default status*	No	$9,644$	252	$9,896$
	Yes	23	81	104
	Total	$9,667$	333	$10,000$

Source: Hastie, Tibshirani and Friedman (2008)

- Start with $c = 0$ and increase it progressively, one error decreases while the other increases. ROC curve.
- *Criterion:* error rate: $1/N \ \sum I(y_i \neq \hat{y}_i)$
- Model choice: previous strategies work (CV, in particular)

Trees

Titanic victims: logit



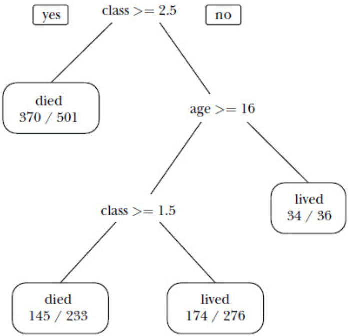
```
Logistic Regression Table

Predictor           Coef    SE Coef      Z      P
Constant       -0.191839   0.175568  -1.09  0.275
Class
First           0.971320  0.0952002  10.20  0.000
Gender
Male            -1.03799   0.200630  -5.17  0.000
Age            0.0044963  0.0033885   1.33  0.185
ChildorAdult
Child           0.387517   0.174976   2.21  0.027
Gender*Age
Male          -0.0123825  0.0040596  -3.05  0.002
```

Titanic victims: CART

Figure 1

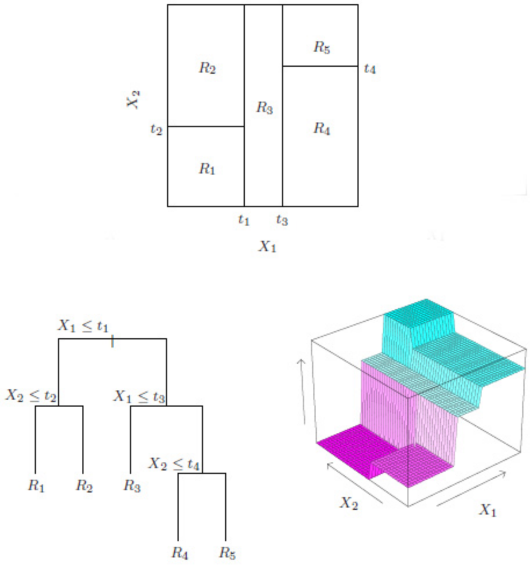
A Classification Tree for Survivors of the *Titanic*



Note: See text for interpretation.

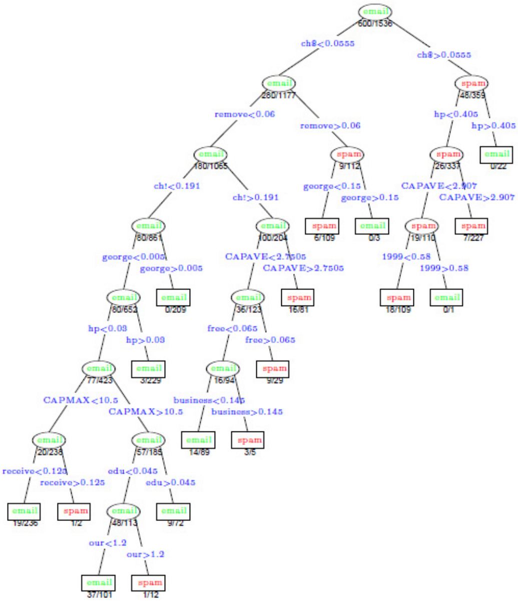
CART algorithm

- Pick the j-th explanatory variable. Split all data according to $X_j > s$. This will split data in two. Within each region, predict: mean, majority rule.
- Problem: how to choose j and s?: to minimize prediction error? How it is done? Try everything (all variables, all possible spliting points! (big data, big computer!).
- Now do the same in each split.
- Keep spliting if prediction error falls.


R_2
R_3
R_5
t_4
R_4
X_2
t_2
R_1
t_1
t_3
X_1
$X_1 \leq t_1$
$X_2 \leq t_2$
$X_1 \leq t_3$
$X_2 \leq t_4$
R_1
R_2
R_3
R_4
R_5
X_2
X_1

Things:

- Widely popular. Very easy to communicate and use.
- Tree branches ‘grow’ in each split. Prunning: how to stop the tree.
- Important variables: appear split at the top.
- Bad performance if underlying model is linear (logit or LPM better).
- Bagging: bootstrap samples and CART in each bootstrapped sample. Then vote.
- Boosting: reweight missclasified observations.
- Random forests: grow multiple trees using bootstrapped predictors at each split. Majority vote.
- Difficult to interpret out of simple trees.


email
600/1536
ch$<0.0555
ch$>0.0555
email
280/1177
spam
48/359
remove<0.06
remove>0.06
hp<0.405
hp>0.405
email
180/1065
spam
9/112
spam
26/337
email
0/22
ch!<0.191
ch!>0.191
george<0.15
george>0.15
CAPAVE<2.907
CAPAVE>2.907
email
80/861
email
100/204
spam
6/109
email
0/3
spam
19/110
spam
7/227
george<0.005
george>0.005
CAPAVE<2.7505
CAPAVE>2.7505
1999<0.58
1999>0.58
email
80/652
email
0/209
email
36/123
spam
16/81
spam
18/109
email
0/1
hp<0.03
hp>0.03
free<0.065
free>0.065
email
77/423
email
3/229
email
16/94
spam
9/29
CAPMAX<10.5
CAPMAX>10.5
business<0.145
business>0.145
email
20/238
email
57/185
email
14/89
spam
3/5
receive<0.125
receive>0.125
edu<0.045
edu>0.045
email
19/236
spam
1/2
email
48/113
email
9/72
our<1.2
our>1.2
email
37/101
spam
1/12

Shadow prices

- Basis expansion
- Neural networks.
- Support vector machines.
- Unsupervised vs. supervised
- Clusters and groups
- Principal components and dimension reduction
- Fat models
- Data handling / programming / scrapping

Perspectives

- People tend to have extreme opinions on this.
- Deep philosophical / methodological discussion.
- The role of induction, theory, modeling.
- Theoryless? Really? Harmless?

Toolkit

Readings

- Hastie, Tibshirani,Friedman (2009)
- James, Witten, Hastie and Tibshirani (2014).
- Murphy (2012, Machine Learning)
- Varian (2014)
- Special issue on big data (JEP, 2014)
- Papers: Keely and Tan (2008, Journal of Public Econommics), Bajari et al. (2015, American Economic Review), Cavallo and Rigobon (2013, Journal of Monetary Economics).
- Mayer-Schonberger y Cukier (Big Data, 2013).

- Tim Harford talk on ‘The Big Data Trap’.
- Nota en Clarin (6/4/2014)
- Super computer intensive
- Forget about Stata. Go R (to start with)
- Online course: Hastie and Tibshirani (Stanford)
- Free books!

Borgesian redux

‘... su antepasado no creia en un tiempo uniforme, absoluto. Creia en infinitas series de tiempos, en una red creciente y vertiginosa de tiempos divergentes, convergentes y paralelos. ... No existimos en la mayora de esos tiempos; en algunos existe usted y no yo; en otros, yo, no usted; en otros, los dos. En este, que un favorable azar me depara, usted ha llegado a mi casa; en otro, usted, al atravezar el jardn, me ha encontrado muerto; en otro, yo digo estas mismas palabras, pero soy un error, un fantasma.’

‘Ireneo tena diecinueve años; haba nacido en 1868; me parecio monumental como el bronce, ms antiguo que Egipto, anterior a las profecias y a las piramides. Pense que cada una de mis palabras (que cada uno de mis gestos) perduraria en su implacable memoria; me entorpecio el temor de multiplicar ademanes inutiles.’