Machine Learning for Data Analysis

Jesús Fernández-Villaverde[1] and Galo Nuño[2]

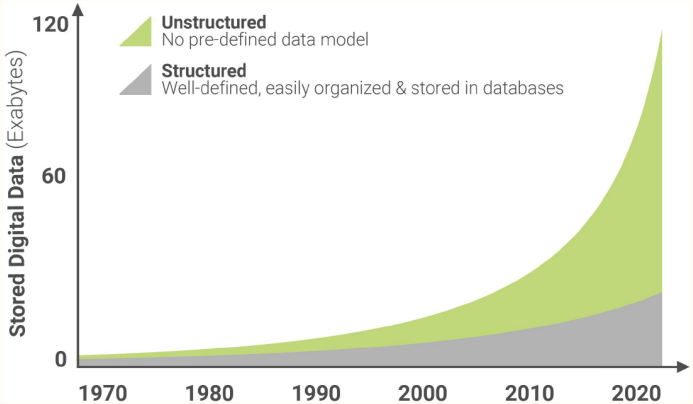
August 7, 2023

[1]University of Pennsylvania

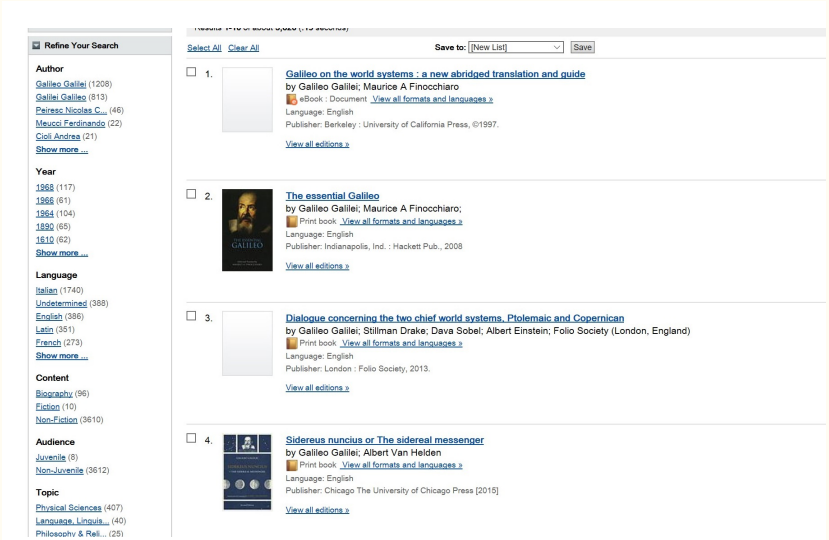
[2]Banco de España

New data

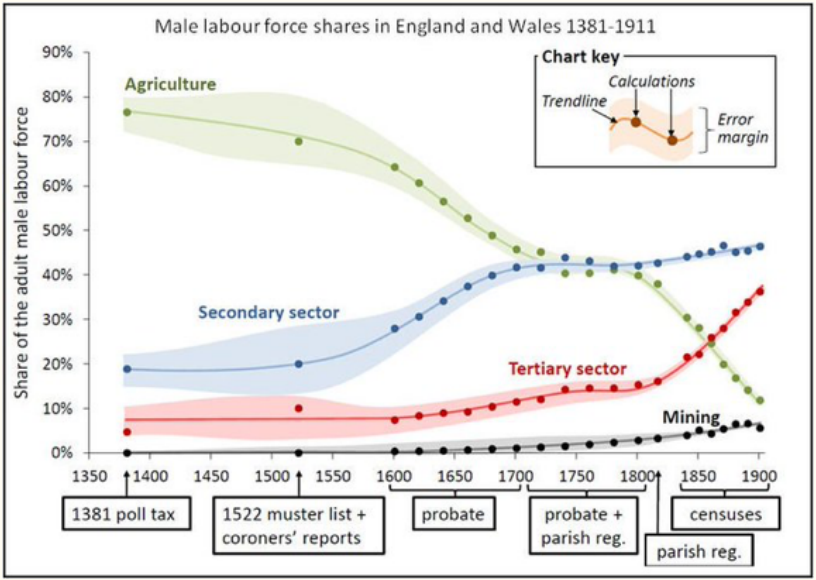
- Most important lesson for economists from data science: Everything is data.
- Unstructured data: Newspaper articles, business reports, congressional speeches, FOMC meetings transcripts, satellite data, photographs, audio, mobility, ...



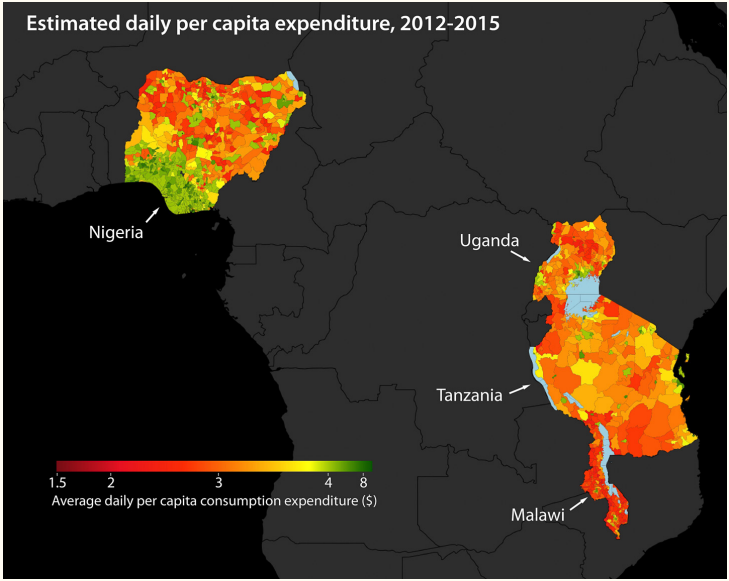
Library data



Parish and probate data



Satellite imagery



Luminosity

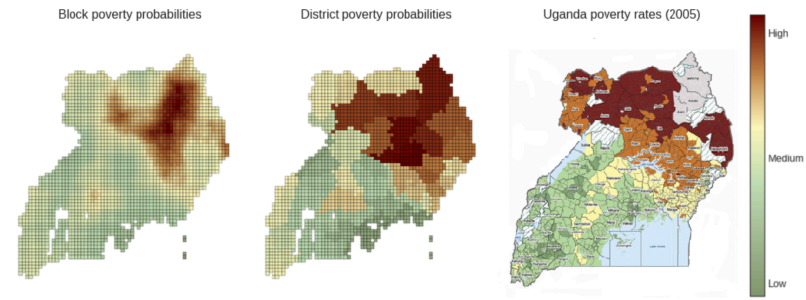


Figure 3: **Left:** Predicted poverty probabilities at a fine-grained 10km × 10km block level. **Middle:** Predicted poverty probabilities aggregated at the district-level. **Right:** 2005 survey results for comparison (World Resources Institute 2009).

Xie et. al. (2016)

Cell use

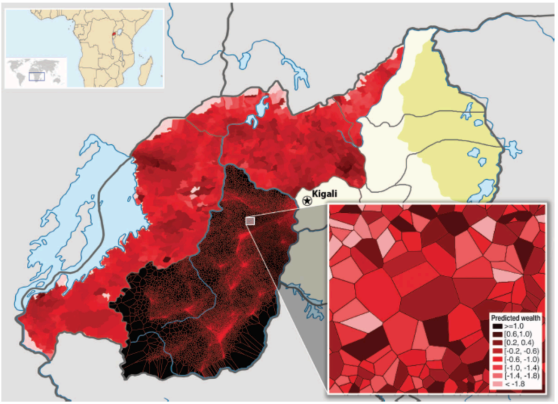


Fig. 2. Construction of high-resolution maps of poverty and wealth from call records. Information derived from the call records of 1.5 million subscribers is overlaid on a map of Rwanda. The northern and western provinces are divided into cells (the smallest administrative unit of the country), and the cell is shaded according to the average (predicted) wealth of all mobile subscribers in that cell. The southern province is overlaid with a Voronoi division that uses geographic identifiers in the call data to segment the region into several hundred thousand small partitions. (**Bottom right inset**) Enlargement of a 1-km^2 region near Kiyonza, with Voronoi cells shaded by the predicted wealth of small groups (5 to 15 subscribers) who live in each region.

Blumenstock et. al. (2015)

TABLE I

MOST PARTISAN PHRASES FROM THE 2005 *CONGRESSIONAL RECORD*[a]

Panel A: Phrases Used More Often by Democrats

Two-Word Phrases		
private accounts	Rosa Parks	workers rights
trade agreement	President budget	poor people
American people	Republican party	Republican leader
tax breaks	change the rules	Arctic refuge
trade deficit	minimum wage	cut funding
oil companies	budget deficit	American workers
credit card	Republican senators	living in poverty
nuclear option	privatization plan	Senate Republicans
war in Iraq	wildlife refuge	fuel efficiency
middle class	card companies	national wildlife
Three-Word Phrases		
veterans health care	corporation for public broadcasting	cut health care
congressional black caucus	additional tax cuts	civil rights movement
VA health care	pay for tax cuts	cuts to child support
billion in tax cuts	tax cuts for people	drilling in the Arctic National
credit card companies	oil and gas companies	victims of gun violence
security trust fund	prescription drug bill	solvency of social security
social security trust	caliber sniper rifles	Voting Rights Act
privatize social security	increase in the minimum wage	war in Iraq and Afghanistan
American free trade	system of checks and balances	civil rights protections
central American free	middle class families	credit card debt

TABLE I—*Continued*

Panel B: Phrases Used More Often by Republicans

Two-Word Phrases		
stem cell	personal accounts	retirement accounts
natural gas	Saddam Hussein	government spending
death tax	pass the bill	national forest
illegal aliens	private property	minority leader
class action	border security	urge support
war on terror	President announces	cell lines
embryonic stem	human life	cord blood
tax relief	Chief Justice	action lawsuits
illegal immigration	human embryos	economic growth
date the time	increase taxes	food program
Three-Word Phrases		
embryonic stem cell	Circuit Court of Appeals	Tongass national forest
hate crimes legislation	death tax repeal	pluripotent stem cells
adult stem cells	housing and urban affairs	Supreme Court of Texas
oil for food program	million jobs created	Justice Priscilla Owen
personal retirement accounts	national flood insurance	Justice Janice Rogers
energy and natural resources	oil for food scandal	American Bar Association
global war on terror	private property rights	growth and job creation
hate crimes law	temporary worker program	natural gas natural
change hearts and minds	class action reform	Grand Ole Opry
global war on terrorism	Chief Justice Rehnquist	reform social security

[a]The top 60 Democratic and Republican phrases, respectively, are shown ranked by χ^2_{pl}. The phrases are classified as two or three word after dropping common "stopwords" such as "for" and "the." See Section 3 for details and see Appendix B (online) for a more extensive phrase list.

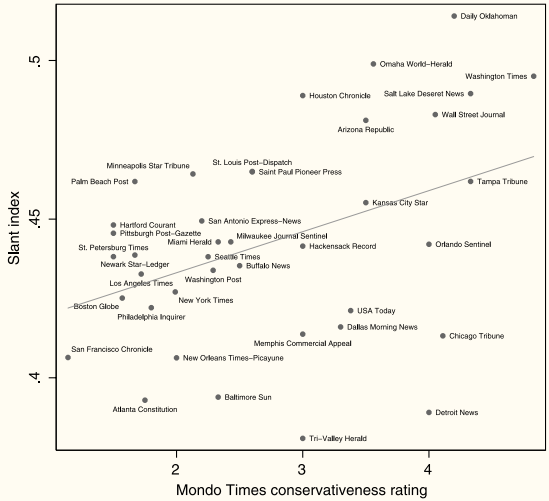
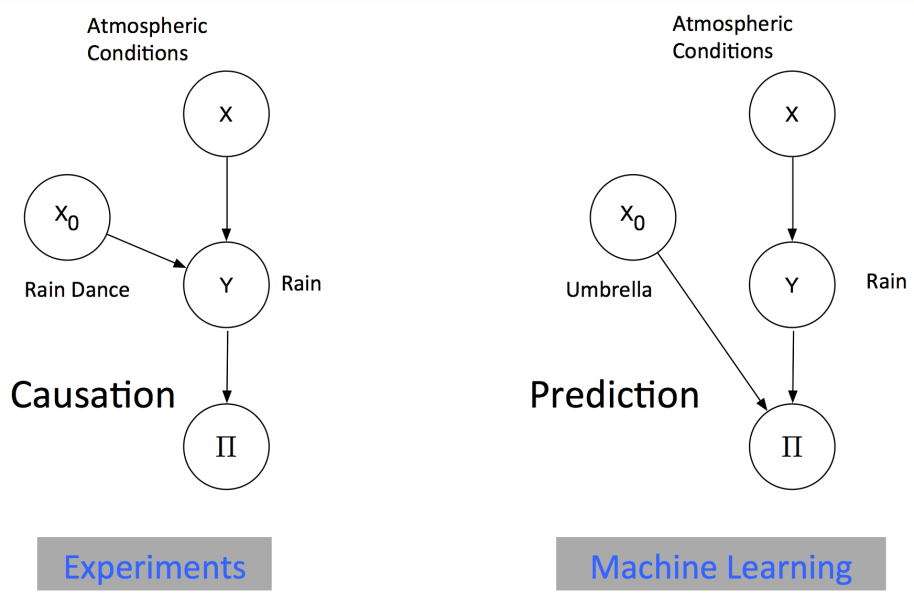


FIGURE 1.—Language-based and reader-submitted ratings of slant. The slant index (y axis) is shown against the average Mondo Times user rating of newspaper conservativeness (x axis), which ranges from 1 (liberal) to 5 (conservative). Included are all papers rated by at least two users on Mondo Times, with at least 25,000 mentions of our 1000 phrases in 2005. The line is predicted slant from an OLS regression of slant on Mondo Times rating. The correlation coefficient is 0.40 ($p = 0.0114$).

Economics and machine learning II

- A more general point $\Rightarrow$ role of causality in economics:
  1. Counterfactuals.
  2. Welfare.
  3. General equilibrium effects.
  4. New changes.
  5. Less data.
- Another example by Athey (2017): hotel prices and occupancy rates. In the data, prices and occupancy rates are strongly positively correlated, but what is the expected impact of a hotel raising its prices on a given day?


Atmospheric Conditions
X
X_0
Rain Dance
Y
Rain
Causation
Π
Experiments
Atmospheric Conditions
X
X_0
Umbrella
Y
Rain
Prediction
Π
Machine Learning

Unsupervised learning

Unsupervised learning

- Use a sample:

$$\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^{N}$$

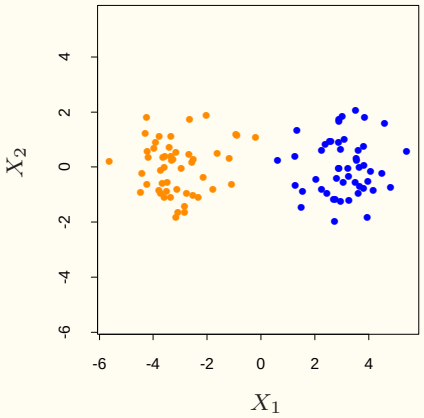
to:

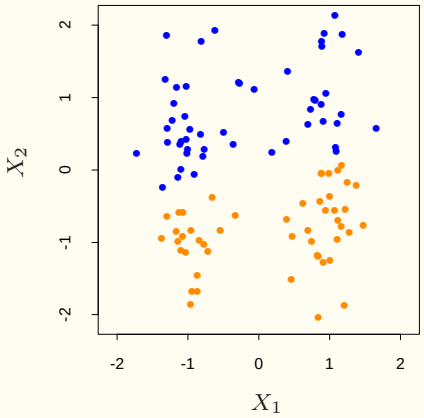
1. Group observations in interesting patterns.
2. Describe most important sources of variation in the data.
3. Dimensionality reduction.

- Example: what can we learn about the loan book of a bank without imposing too much a priori structure?
- More concretely, we search for:

$$p\left(\mathbf{x}_i|\theta\right)$$

- Clustering and association rules.


X_2
4
2
0
-2
-4
-6
-6
-4
-2
0
2
4
X_1


X_2
2
1
0
-1
-2
-2
-1
0
1
2
X_1

Cluster discovering

1. Select K clusters

$$K^* = \underset{K}{\operatorname{argmax}}\, p\left(K|\mathcal{D}\right)$$

2. Assign each observation to a cluster

$$z_t^* = \underset{k}{\operatorname{argmax}}\, p\left(z_i = k|\mathbf{x}_i, \mathcal{D}\right)$$

3. A common method to pin down K is the silhouette. For each observation i, we compute:

$$s_i = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

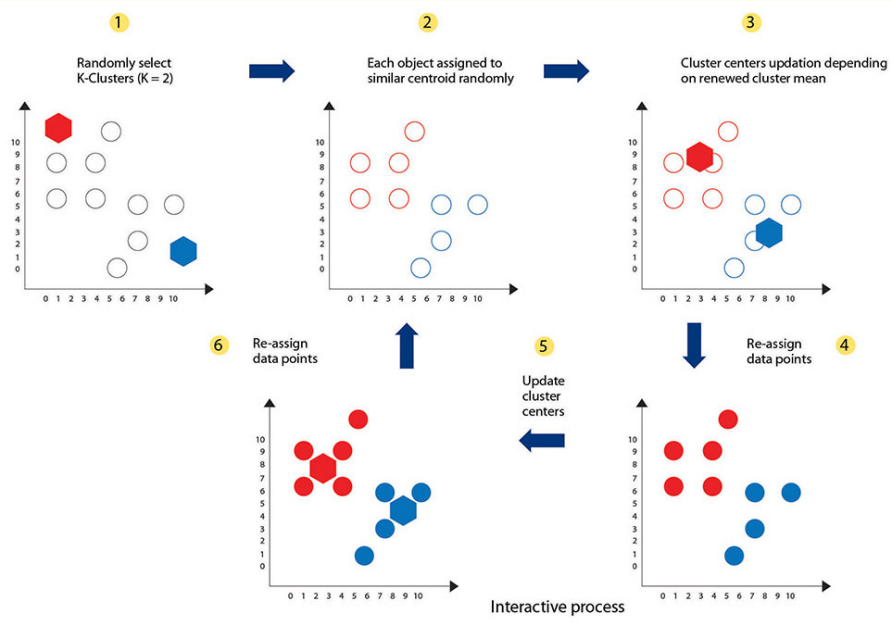
where $a(i)$ is the average distance between i and all other members of the cluster while $b(i)$ is the minimum distance between i and all other members of another cluster.

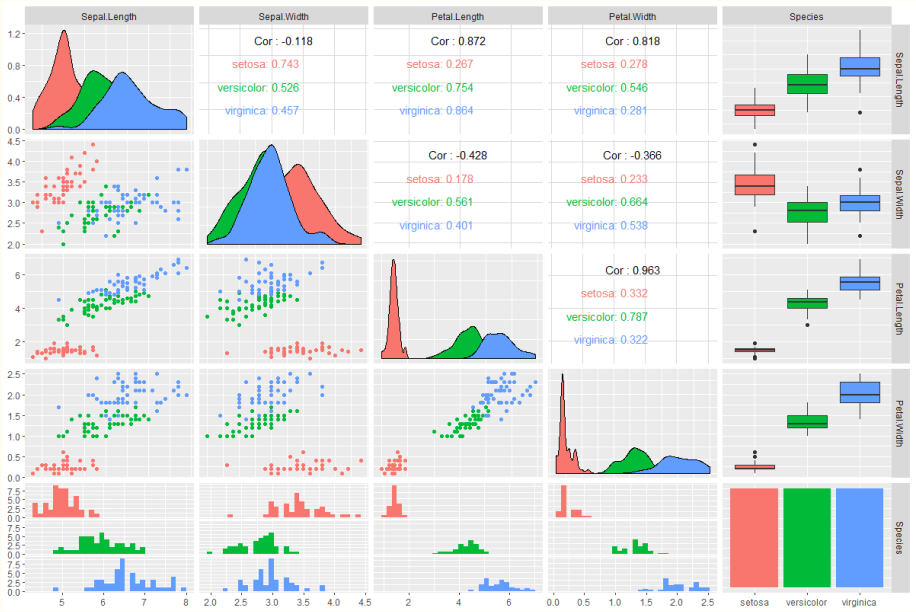
K-means

- K-means clustering by Steinhaus (1957)

$$\underset{S}{\operatorname{argmax}} \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2$$

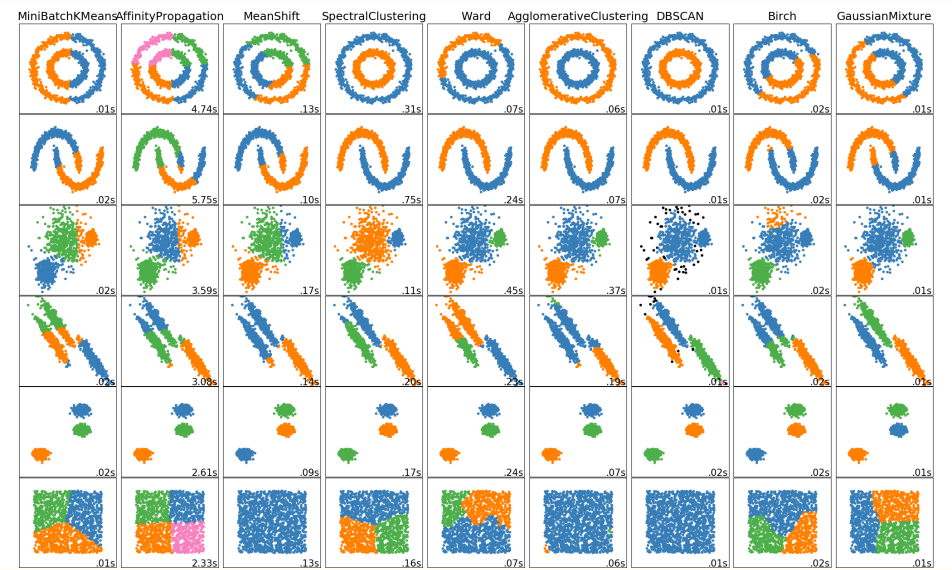
- It requires an iterative algorithm for implementation Lloyd (1957).
- Related variations:
  1. k-medians $\Rightarrow$ uses medians computed through the Taxicab geometry.
  2. k-medoids $\Rightarrow$minimizes a sum of pairwise dissimilarities.
  3. k-SVD.

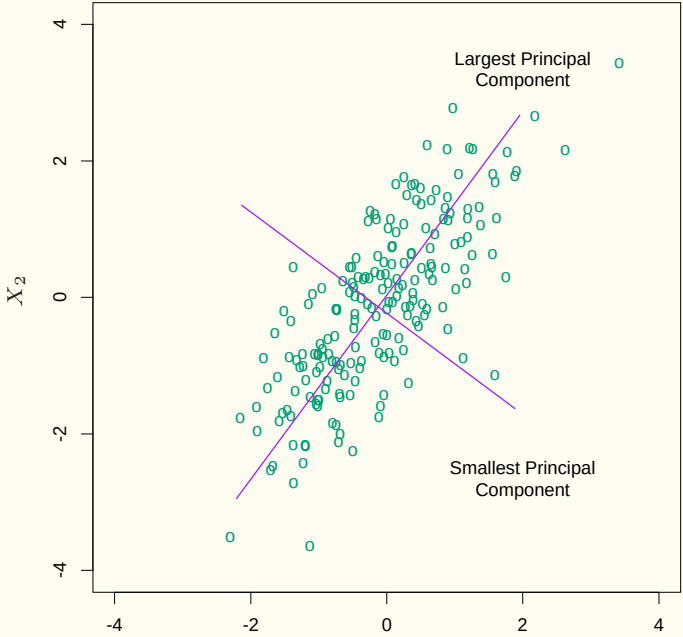

1
Randomly select
K-Clusters (K = 2)
2
Each object assigned to
similar centroid randomly
3
Cluster centers updation depending
on renewed cluster mean
4
Re-assign
data points
5
Update
cluster
centers
6
Re-assign
data points
Interactive process


Sepal.Length
Sepal.Width
Petal.Length
Petal.Width
Species
Cor : -0.118
setosa: 0.743
versicolor: 0.526
virginica: 0.457
Cor : 0.872
setosa: 0.267
versicolor: 0.754
virginica: 0.864
Cor : 0.818
setosa: 0.278
versicolor: 0.546
virginica: 0.281
Cor : -0.428
setosa: 0.178
versicolor: 0.561
virginica: 0.401
Cor : -0.366
setosa: 0.233
versicolor: 0.664
virginica: 0.538
Cor : 0.963
setosa: 0.332
versicolor: 0.787
virginica: 0.322
setosa
versicolor
virginica

Other algorithms

- Other clustering methods:
  1. Agglomerative clustering.
  2. DBSCAN.
  3. Birch.
- Principal component analysis.
- Density estimation.
- Gaussian mixture models.
- Association rules and the Apriori algorithm (Agrawal and Srikant, 1994).


MiniBatchKMeans
AffinityPropagation
MeanShift
SpectralClustering
Ward
AgglomerativeClustering
DBSCAN
Birch
GaussianMixture
.01s
4.74s
.13s
.31s
.07s
.06s
.01s
.02s
.01s
.02s
5.75s
.10s
.75s
.24s
.07s
.01s
.02s
.01s
.02s
3.59s
.17s
.11s
.45s
.37s
.01s
.02s
.01s
.02s
3.08s
.14s
.20s
.23s
.19s
.01s
.02s
.01s
.02s
2.61s
.09s
.17s
.24s
.07s
.02s
.02s
.01s
.01s
2.33s
.13s
.16s
.07s
.06s
.01s
.02s
.01s


Largest Principal
Component
Smallest Principal
Component
X_2
4
2
0
-2
-4
-4
-2
0
2
4

Supervised learning

Supervised learning, I

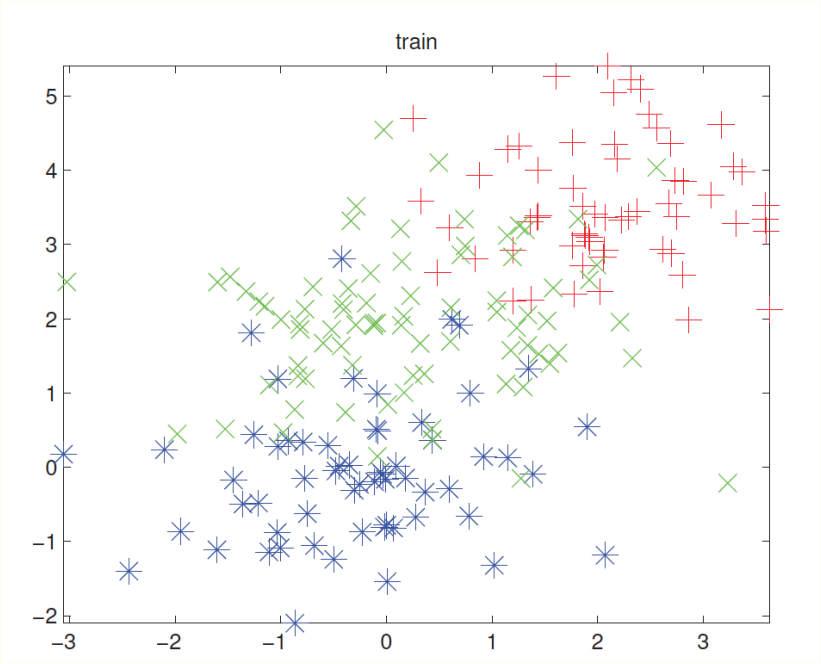
- Use a training sample

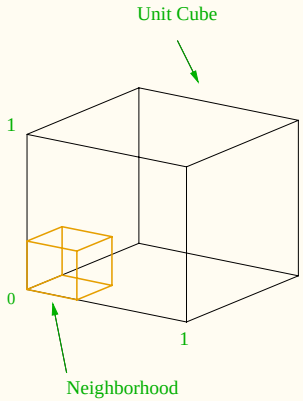
$$\mathcal{D} = \{y_i, \mathbf{x}_i\}_{i=1}^{N}$$

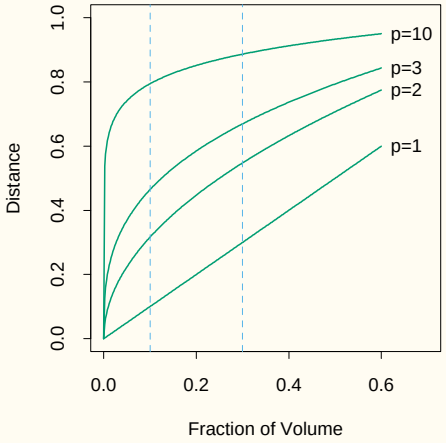
  to classify new observations

$$\mathcal{F} = \{y_i, \mathbf{x}_i\}_{i=N+1}^{M}$$

- $\mathbf{x}_i$ may potentially belong to a high-dimensional space.
- Standard regression in econometrics is a particular example when y_i is a continuous variable.
- But often, in machine learning, y_i is categorical (classification problem).
- Example: classify loans of a bank into good and bad risks.
- Caveat: it can be highly labor intensive!


train
5
4
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
3


Unit Cube
1
0
1
Neighborhood


Distance
0.0
0.2
0.4
0.6
0.8
1.0
p=10
p=3
p=2
p=1
0.0
0.2
0.4
0.6
Fraction of Volume

Supervised learning, II

- Problem of function approximation:

$$y = f(\mathbf{x})$$

  when we know next to nothing about the function $f(\cdot)$.

- Linked with projection and perturbation methods.

- Also, linked with traditional econometrics.

- Then:

$$\widehat{y} = \widehat{f}(\mathbf{x}) = \arg\max_{c \in C} p(y = c|\mathbf{x}, \mathcal{D})$$

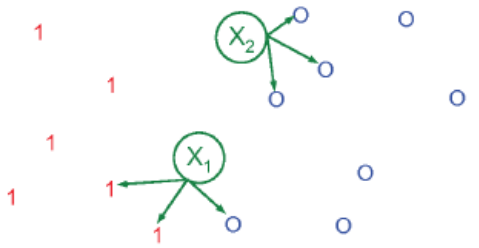
- In Bayesian, this is the mode of the posterior, but in machine learning it is called a MAP estimate (maximum a posteriori).

Example I: K-nearest algorithm

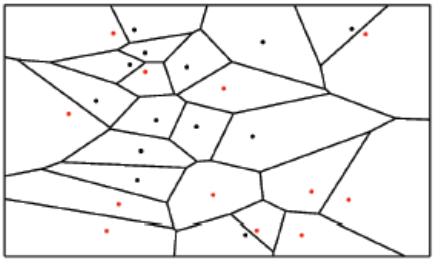
- We look at the $K-$nearest points to x_i in the training sample.
- Formally:

$$p\left(y=c|\mathbf{x},\mathcal{D},K\right)=\frac{1}{K}\sum_{i\in N_K(\mathbf{x},K)}\mathbb{I}\left(y_i=c\right)$$

- It generates a Voronoi tessellation.
- Simple and intuitive, but it suffers from curse of dimensionality.



(a)



(b)

Bias-variance trade-off

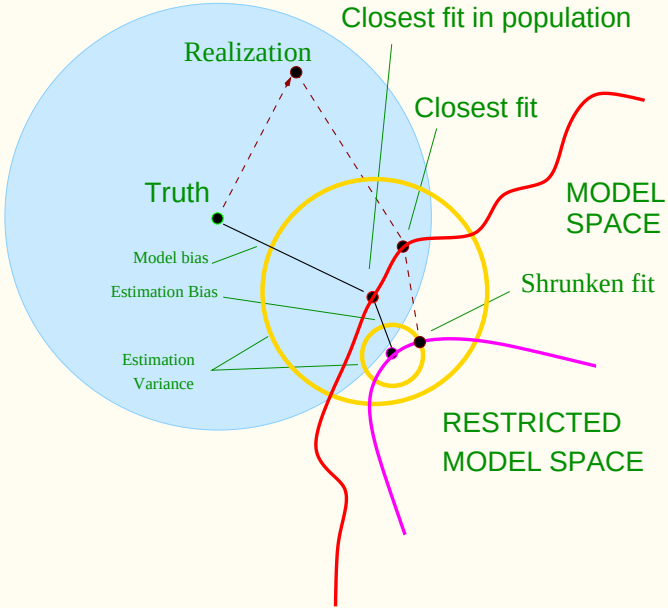
- How do we select K? Minimize misclassification rate on a validation set.
- Bias-variance trade-off is about underfitting vs. overfitting.
- We have:

$$y = f(\mathbf{x}) = \widehat{f}(\mathbf{x}) + \varepsilon,\ \varepsilon \sim ID\left(0, \sigma^2\right)$$

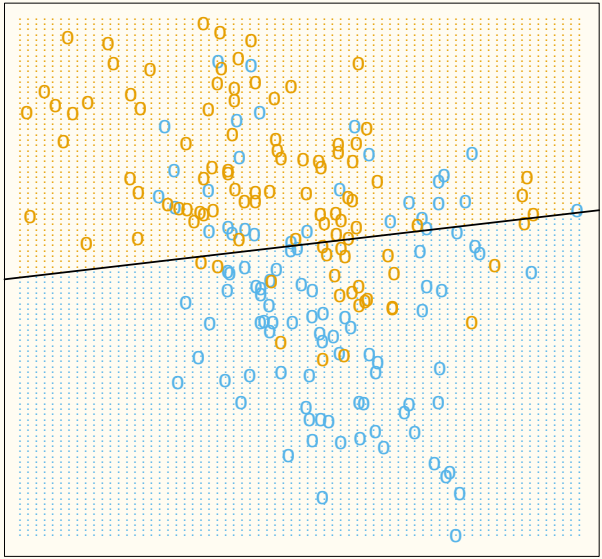
- Then:

$$\begin{aligned}
\mathbb{E}\left(y - \widehat{f}(\mathbf{x})\right)^2 &= \mathbb{E}\left(\widehat{f}(\mathbf{x}) - f(\mathbf{x})\right)^2 \\
&= \left(\mathbb{E}\widehat{f}(\mathbf{x}) - f(\mathbf{x})\right)^2 + \mathbb{E}\left(\widehat{f}(\mathbf{x}) - \mathbb{E}\widehat{f}(\mathbf{x})\right)^2 \\
&= Bias^2 + var
\end{aligned}$$

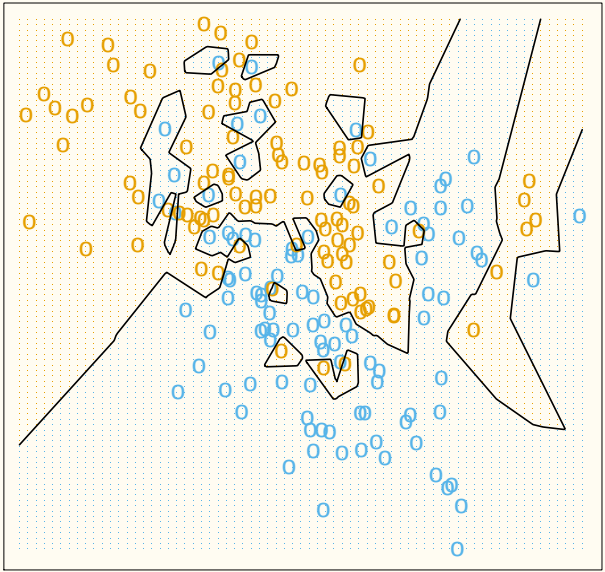
- Traditional econometrics obsession with BLUE is mysterious.


Closest fit in population
Realization
Closest fit
Truth
MODEL
SPACE
Model bias
Estimation Bias
Shrunken fit
Estimation
Variance
RESTRICTED
MODEL SPACE

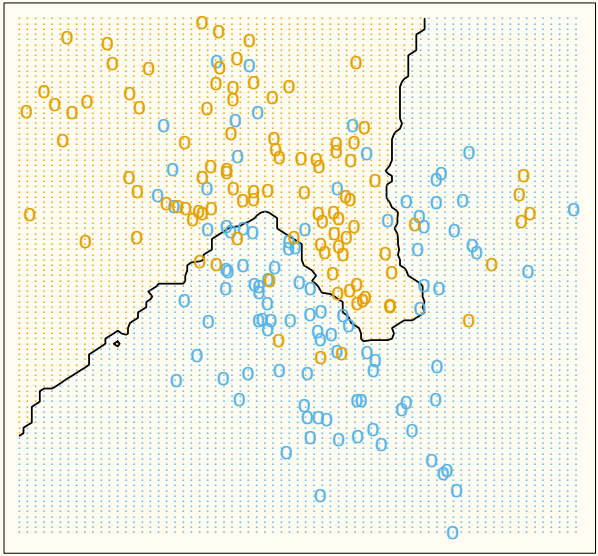
Linear Regression of 0/1 Response



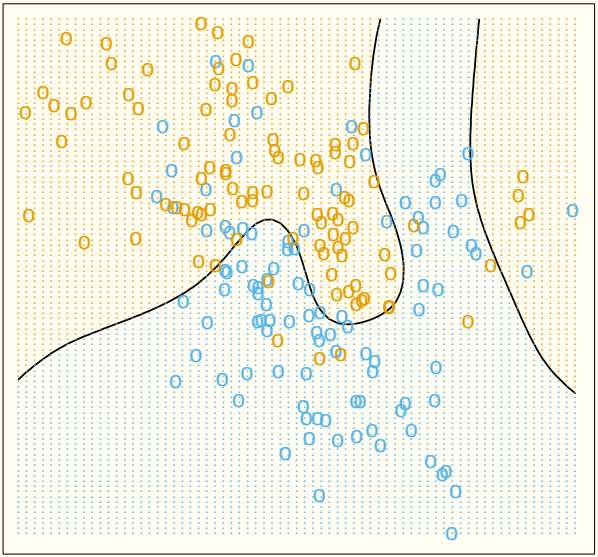
1−Nearest Neighbor Classifier

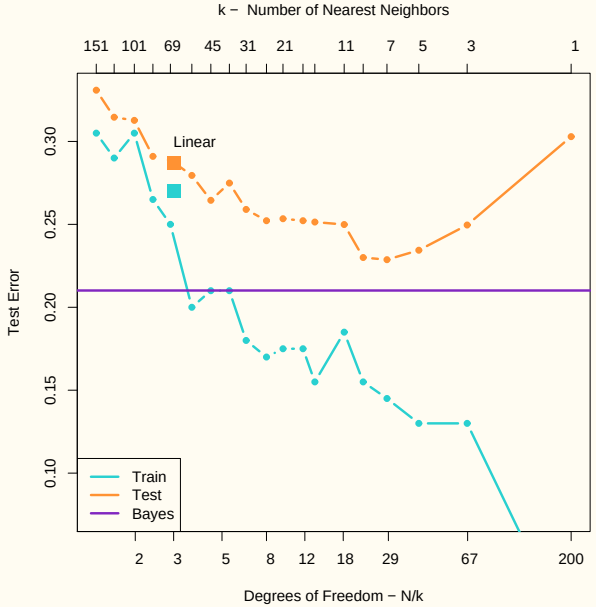


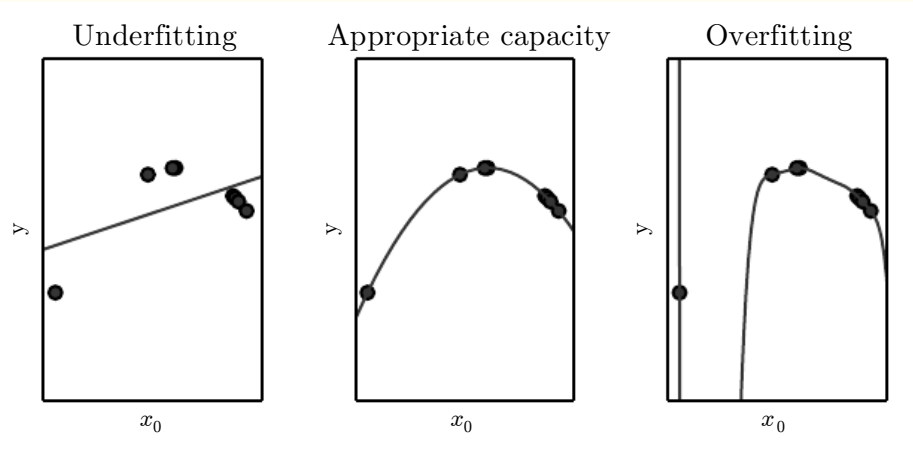
15-Nearest Neighbor Classifier

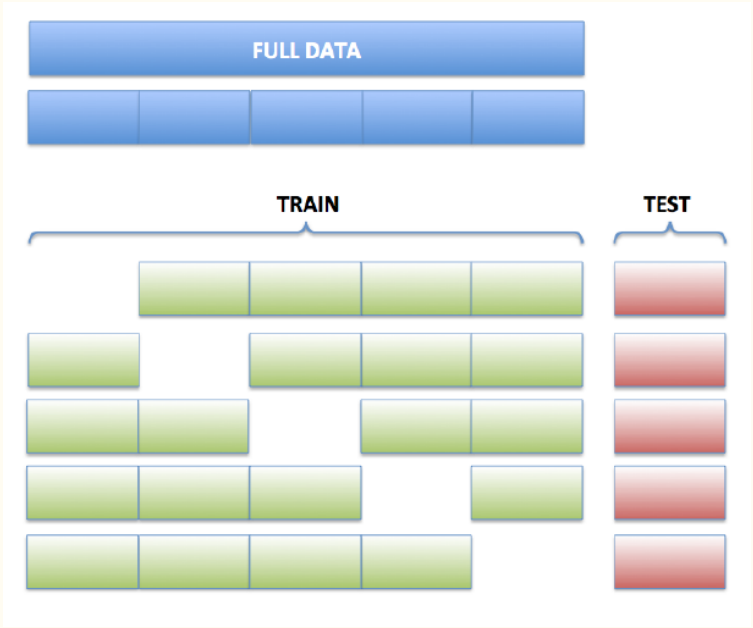


Bayes Optimal Classifier




k − Number of Nearest Neighbors
151
101
69
45
31
21
11
7
5
3
1
Linear
Test Error
0.30
0.25
0.20
0.15
0.10
Train
Test
Bayes
2
3
5
8
12
18
29
67
200
Degrees of Freedom − N/k


Underfitting
Appropriate capacity
Overfitting
y
x_0


FULL DATA
TRAIN
TEST

Example II: Regularized regression

- Standard linear regression:

$$y_i = \mathbf{x}_i'\beta + \varepsilon_t$$

- When number of regressors is 3 or more, OLS is not admissible (Stein's phenomenon).
- In particular, when the number of regressors is large, forecasting properties are poor.
- Zvi Grilliches: "never trust OLS with more than five regressors."
- This problem is becoming more prominent as we get larger datasets with thousands of covariates.
- Also relevant when we have multiple potential IVs.
- Note difference between problem of learning about coefficients and forecasting!

A general formulation

- Solution: regularization (aka as penalization)
  1. Shrink estimates towards zero.
  2. Sparse representation of non-zero parameters.
- Long tradition in econometrics: James–Stein estimator.
- Estimator:

$$\beta_{reg} = \min_{\beta} \sum_{i=1}^{N} \left(y_i - \mathbf{x}_i'\beta\right)^2 + \lambda \left\|\beta\right\|_p^p$$

where

$$\left\|\beta\right\|_p = \left(\sum_{i=1}^{K} |\beta|^p\right)^{\frac{1}{p}}$$

CORE PRINCIPLES IN RESEARCH

OCCAM'S RAZOR

"WHEN FACED WITH TWO POSSIBLE EXPLANATIONS, THE SIMPLER OF THE TWO IS THE ONE MOST LIKELY TO BE TRUE."

OCCAM'S PROFESSOR

"WHEN FACED WITH TWO POSSIBLE WAYS OF DOING SOMETHING, THE MORE COMPLICATED ONE IS THE ONE YOUR PROFESSOR WILL MOST LIKELY ASK YOU TO DO."

JORGE CHAM © 2009

WWW.PHDCOMICS.COM

L_0 norm

- Estimator with $p = 0$ (i.e., the number of non-zero coefficients).
- With an appropriately selected λ, we essentially choose $\boldsymbol{\beta}$ according to AIC/BIC criteria, which have strong statistical foundations.
- Challenge: the optimization problem is difficult, and becomes computationally infeasible when number of coefficients is large.
- Two alternatives:
  1. *Stepwise forward regression*: start with no covariates; add whichever single covariate improves the fit most; continue until no covariate significantly improves fit.
  2. *Stepwise backward regression*: start with all covariates; drop single covariate that leads to the lowest reduction in fit; continue to drop until there is a significant reduction in fit.
- Once we select the relevant covariates, we can perform OLS, but standard errors are not correct since we do not condition on model selection.
- Thus, not very used in practice.

Rigde regression

- Estimator with $p = 2$.
- Close form:

$$\beta = \left(\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}\right)^{-1} \mathbf{X}'\mathbf{Y}$$

- Estimators are shrink towards zero by factor $1 + \lambda$.
- It does not enforce sparsity and, thus, we do not achieve model selection.
- Bayesian interpretation $\lambda = \frac{\sigma^2}{\tau^2}$ where $\beta_{prior} \sim \mathcal{N}\left(\mathbf{0}, \tau^2 \mathbf{I}\right)$ and $\varepsilon_t \sim \mathcal{N}\left(0, \sigma^2\right)$.

LASSO

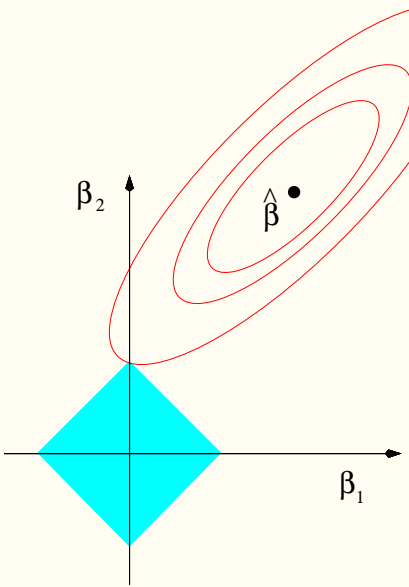
- Least absolute selection and shrinkage operator (Tibshirani, 1996):

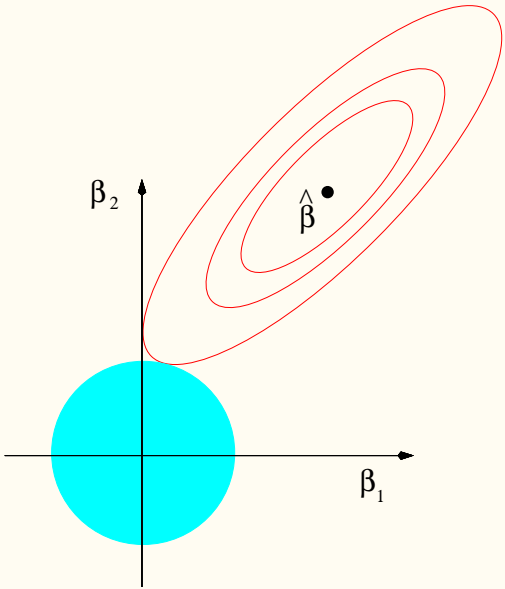
$$\min \sum_{i=1}^{N} \left(y_i - \mathbf{x}_i'\beta\right)^2 + \lambda \left\|\beta\right\|_1$$

- LASSO shrinks some coefficients exactly to zero and others towards zero (variation of relaxed lasso).
- Algorithms for its implementation are easy to code $\Rightarrow$ package in `R package glmnet`.
- Bayesian interpretation: mode of posterior given a Laplace prior

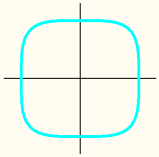
$$p\left(\beta\right) \propto \exp\left(-\lambda \left\|\beta\right\|_1\right)$$

- Bayesian alternative: spike-and-slab prior by Mitchell and Beauchamp (1988).

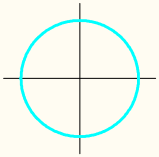

β2
β̂
β1


β2
β̂
β1

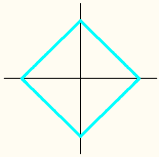
$q = 4$



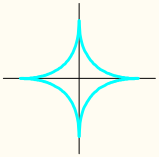
$q = 2$



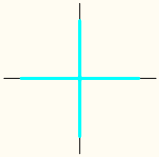
$q = 1$

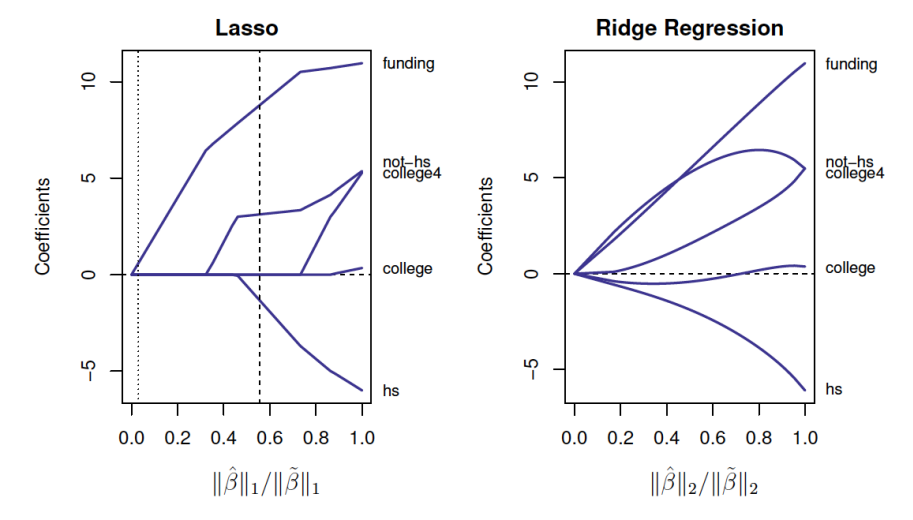


$q = 0.5$



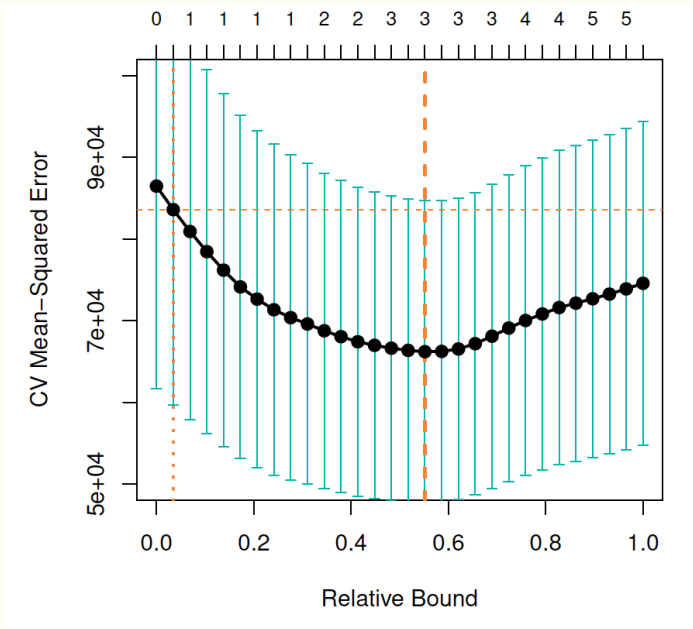
$q = 0.1$




Lasso
Coefficients
10
5
0
−5
0.0
0.2
0.4
0.6
0.8
1.0
$\|\hat{\beta}\|_1/\|\tilde{\beta}\|_1$
funding
not−hs
college4
college
hs
Ridge Regression
Coefficients
10
5
0
−5
0.0
0.2
0.4
0.6
0.8
1.0
$\|\hat{\beta}\|_2/\|\tilde{\beta}\|_2$
funding
not−hs
college4
college
hs

Which penalty?

- The results of LASSO are sensitive to λ.
- Usually chosen via cross-validation by eliminating some test data:
  1. Choose λ to minimize average error of held-out data.
  2. Choose largest λ such that error is within one standard deviation of minimum.
- The choice of penalty parameter also affects model selection consistency.
- Meinshausen and Bühlmann (2006): the prediction-optimal choice of λ is guaranteed NOT to recover the true model asymptotically.
- When λ is chosen by cross validation, we tend to overselect variables and include false positives.
- We need stronger penalization to guarantee model selection consistency, but in finite samples the appropriate choice of λ is not clear.
- Bottom line: in applied work, the LASSO is most useful for variable screening rather than model selection.


0 1 1 1 1 2 2 3 3 3 3 4 4 5 5
CV Mean−Squared Error
5e+04
7e+04
9e+04
0.0
0.2
0.4
0.6
0.8
1.0
Relative Bound

Statistical inference

- We might also be interested in performing hypothesis testing on coefficient estimates from LASSO.
- One option is to put selected variables into an OLS regression. Highly problematic because no conditioning on model selection.
- For example, the first variable selected by LASSO will be the one with the highest partial correlation with the response.
- Suppose number of regressors is large and covariates are randomly generated. One of the covariates is likely to be highly correlated with the response; will be selected by LASSO; and have a significant p-value in an OLS regression.
- This *post-selection inference* problem is not fully resolved in the statistics literature.
- Possibility: sample splitting. Estimate LASSO on some data points, perform OLS regression on selected variables in the remaining data points.

Adaptive LASSO

- Proposed by Zou (2006).
- The LASSO penalty screens out irrelevant variables, but it also biases the estimates of relevant variables.
- Ideally, we want to weaken the penalty on relevant covariates and vice versa.
- Estimator:

$$\min \sum_{i=1}^{N} \left(y_i - \mathbf{x}_i'\beta\right)^2 + \sum_j \omega_j |\beta_j|$$

where

$$\omega_j = \frac{1}{|\widehat{\beta}_j^{OLS}|^{\gamma}}$$

- You can still use standard software to estimate it.

Elastic net

- Advantages and disadvantages of LASSO vs. ridge: saturation in the "large number of regressons, small sample" case.
- We can combine Lasso and ridge:

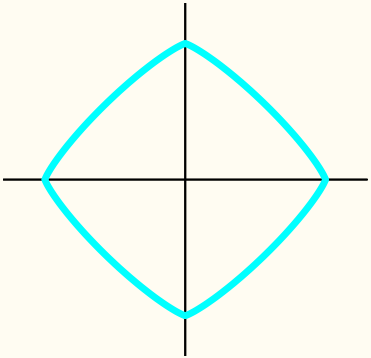
$$\min \sum_{i=1}^{N} \left(y_i - \mathbf{x}_i'\beta\right)^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2$$

or

$$\min \sum_{i=1}^{N} \left(y_i - \mathbf{x}_i'\beta\right)^2 + \lambda \left(\|\beta\|_1 + \alpha \|\beta\|_2\right)$$

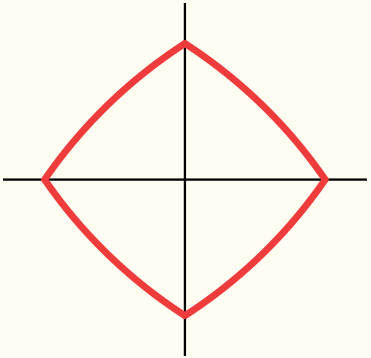
- A elastic net can be expressed as linear support vector machine.
- Particularly useful because we can use available software for support vector machines.

$q = 1.2$



L_q

$\alpha = 0.2$



Elastic Net

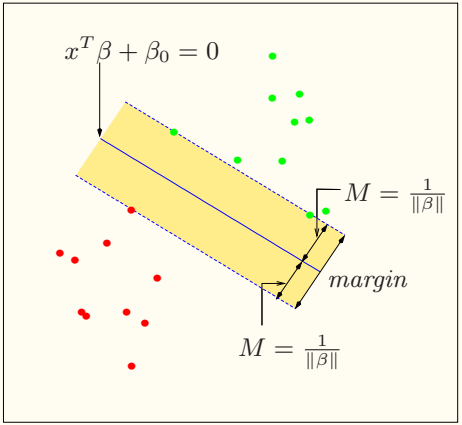
More background

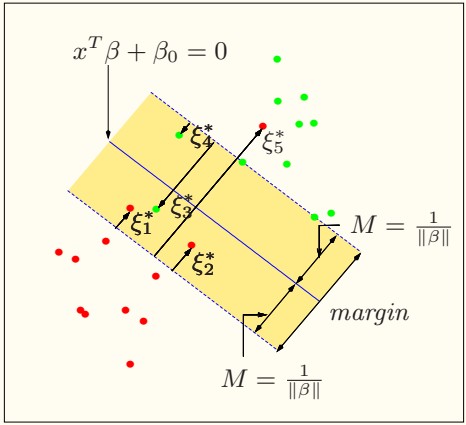
- References:

1. *Statistical Learning with Sparsity: The Lasso and Generalizations*, by Trevor Hastie, Robert Tibshirani, and Martin Wainwright (2015).
2. "High-Dimensional Methods and Inference on Treatment and Structural Effects in Economics," by Belloni, Chernozhukov, and Hansen (2014).
3. "Inference on Treatment Effects after Selection among High-Dimensional Controls," by Belloni, Chernozhukov, and Hansen (2014).

Example II: Support vector machines

- Developed by Vapnik and Alexey Ya. Chervonenkis (1963).
- One of the most popular algorithms in supervised learning.
- Nice example of the "kernel trick."
- Classification problem for $\{y_i, \mathbf{x}_i\}_{i=1}^{N}$ with $y_i \in \{-1, 1\}$.
- Divide observations in two groups by a boundary $f(\mathbf{x}_i)$.
- Maximum-margin hyperplane principle: we search for the boundary that separates as much as possible both regions of data (hard-margin in separable case vs. soft-margin in non-separable case).
- We call observations at the margin support vectors.
- Comparison with logit or probit.


$x^T\beta + \beta_0 = 0$
$M = \frac{1}{\|\beta\|}$
$margin$
$M = \frac{1}{\|\beta\|}$


$x^T\beta + \beta_0 = 0$
ξ_4^*
ξ_5^*
ξ_3^*
ξ_1^*
ξ_2^*
$M = \frac{1}{\|\beta\|}$
$margin$
$M = \frac{1}{\|\beta\|}$

Linear boundaries

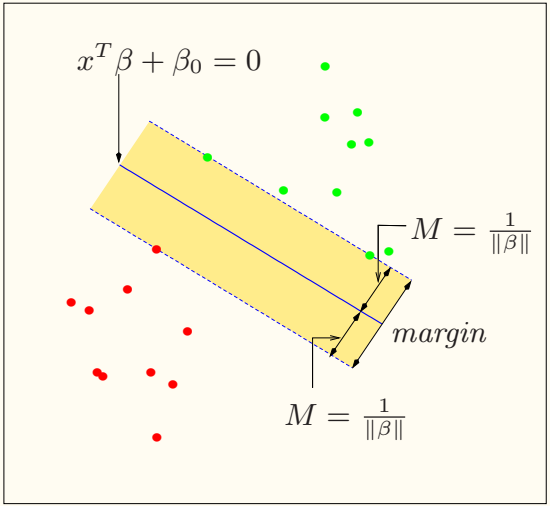
- A standard choice: linear boundary with one regressor

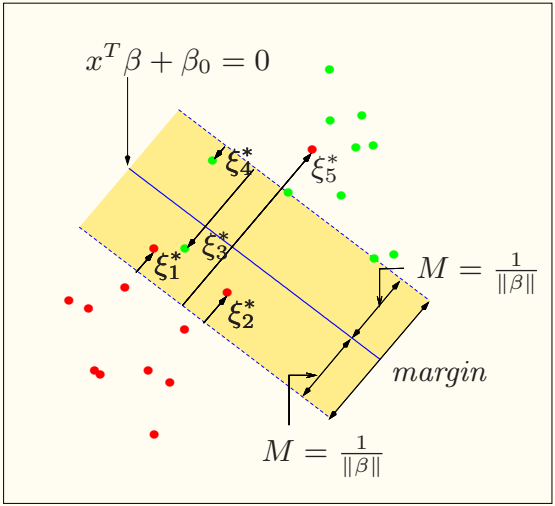
$$\begin{aligned} y_i &= 1 \text{ if } \beta_0 + \beta_1 x \geq 0 \\ y_i &= -1 \text{ otherwise} \end{aligned}$$

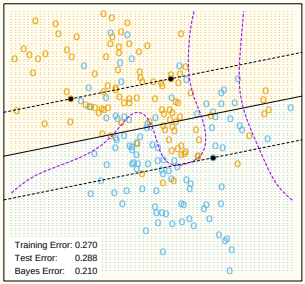
- Find:

$$\min_{\beta_0, \beta_1} \frac{1}{N} \sum_{i=1}^{N} \max\left[0, \left(1 - y_i \left(\beta_0 - \beta_1\right) x\right)\right] + \lambda \left\|\beta\right\|_2^2$$

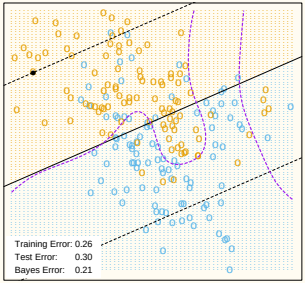
- The regularization term $\lambda \left\|\beta\right\|_2^2$ is known as the hinge loss.
- Note that by setting $\lambda \to 0$ in the hinge loss, we approximate a hard-margin boundary.
- You can solve the primal problem with a sub-gradient descent method or dual problem with a coordinate descent algorithm.


$x^T\beta + \beta_0 = 0$
$M = \frac{1}{\|\beta\|}$
$margin$
$M = \frac{1}{\|\beta\|}$

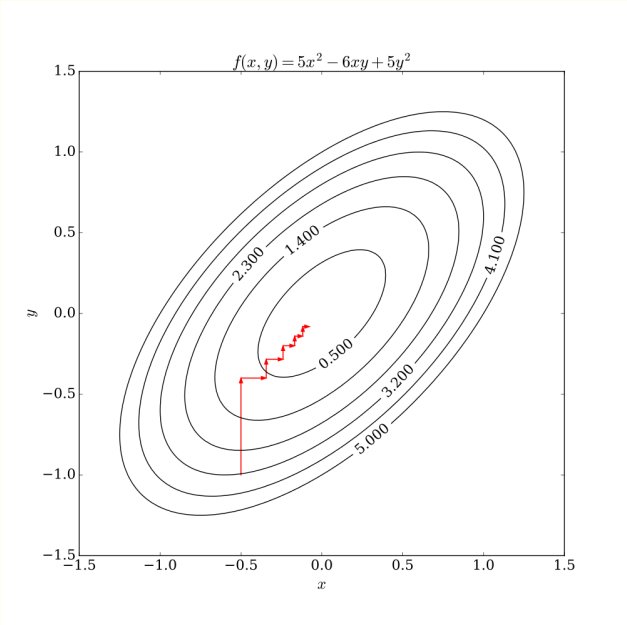

$x^T\beta + \beta_0 = 0$
ξ_4^*
ξ_5^*
ξ_3^*
ξ_1^*
ξ_2^*
$M = \frac{1}{\|\beta\|}$
$margin$
$M = \frac{1}{\|\beta\|}$


Training Error: 0.270
Test Error: 0.288
Bayes Error: 0.210

$C = 10000$


Training Error: 0.26
Test Error: 0.30
Bayes Error: 0.21

$C = 0.01$


$f(x, y) = 5x^2 - 6xy + 5y^2$
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
y
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
x
2.300
1.400
4.100
0.500
3.200
5.000

Example III: Regression trees

- Classification and Regression Trees, Leo Breiman, Jerome Friedman, Charles J. Stone, and R.A. Olshen.
- Main idea: recursively subdivide regressors into subspaces and compute the mean of the regressor in that subspace.
- Result is a flexible step function.
- Works surprisingly well in practice (very popular in data mining), but nearly no theoretical result (asymptotic normality?).
- Cross-validation to determine penalty λ on the number of leaves in the tree.
- "Bet on sparsity" principle.

Recursive partitions I

- Compute

$$f(\mathbf{x}) = \overline{y}$$

with

$$MSE(f(\mathbf{x})) = \sum_{i=1}^{N} (y_i - f(\mathbf{x}))^2$$

- For a regressor x_k and a threshold t, find:

$$\overline{y}|x_k \leq t \text{ and } \overline{y}|x_k > t$$

and define

$$f_{k,t}(\mathbf{x}) = \begin{cases} \overline{y}|x_k \leq t \text{ if } x_k \leq t \\ \overline{y}|x_k > t \text{ if } x_k > t \end{cases}$$

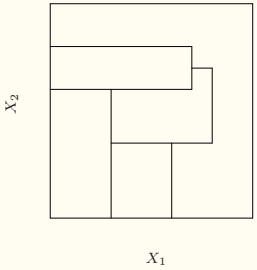
Recursive partitions II

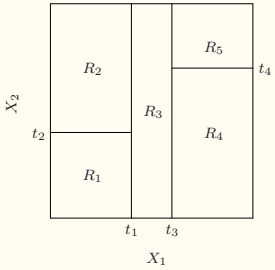
- Select regressor x_k and a threshold t that minimize:

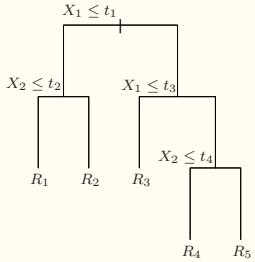
$$(x_k^*, t^*) = \arg\min_{x_k, t} MSE\left(f_{k,t}(\mathbf{x})\right)$$

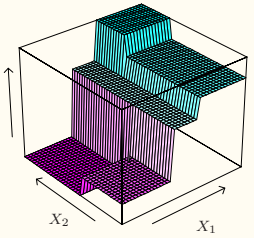
- Keep iterating until

$$\min\left\{MSE\left(f_{k,t}(\mathbf{x})\right) + \lambda \# leaves\right\}$$


X_2
X_1


R_2
R_5
t_4
X_2
R_3
t_2
R_4
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

Improvements

- Pruning an existing tree.
- Boosting: regression tree on new data $(\varepsilon - f(\mathbf{x}), \mathbf{x})$ and iterate.
- Bagging (Bootstrap AGGregatING): regresssion trees on bootstraped data, estimate is average over bootstraped samples.
- Random forest.

Other algorithms

- Multi-task LASSO.
- Least Angle Regression (LARS).
- LARS LASSO.
- Gaussian Process Regression (GPR).
- Dantzig Selector.
- Ensemble methods.
- Super learners/stacking (asymptotic properties).