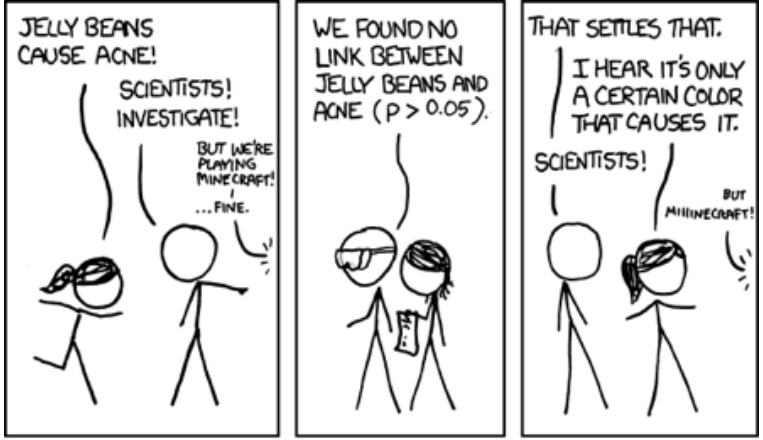
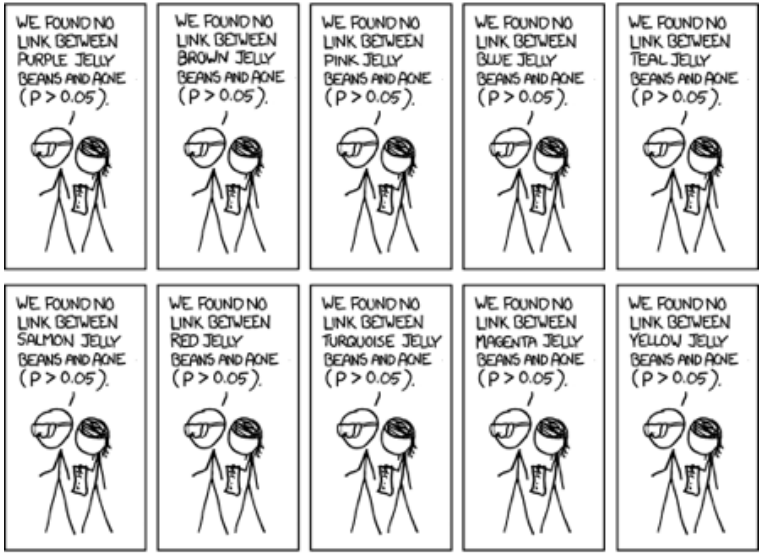
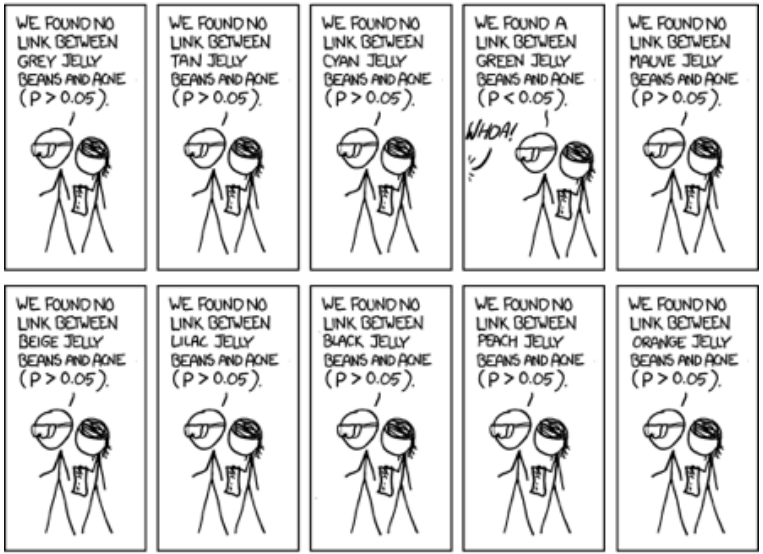
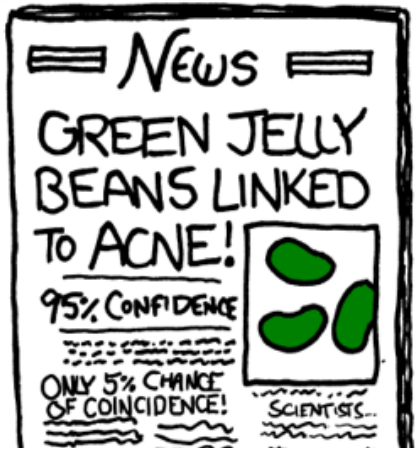

JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MINECRAFT!


WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).


WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).


NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

Supervised Learning Techniques: Cross-validation and Regularization

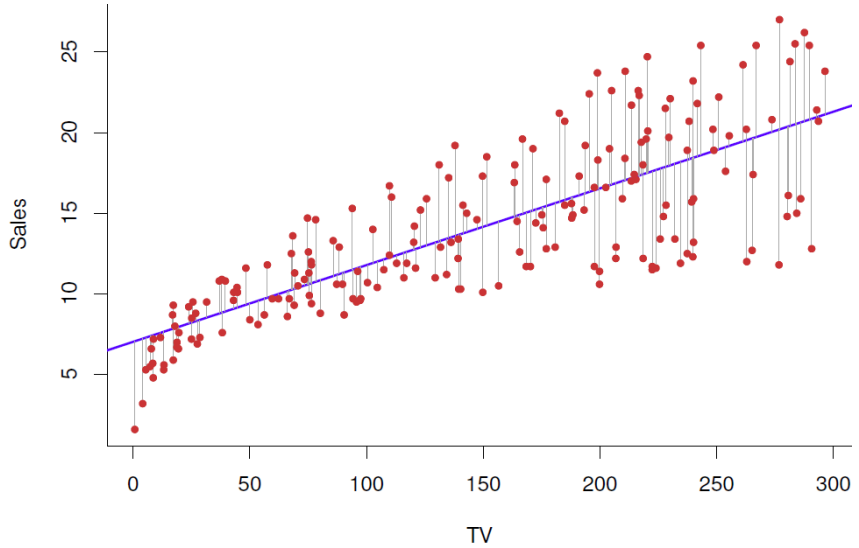
Alexander Ioannidis
ioannidis@stanford.edu

Institute for Computational and Mathematical Engineering,
Stanford University

Loss Functions

Loss Functions

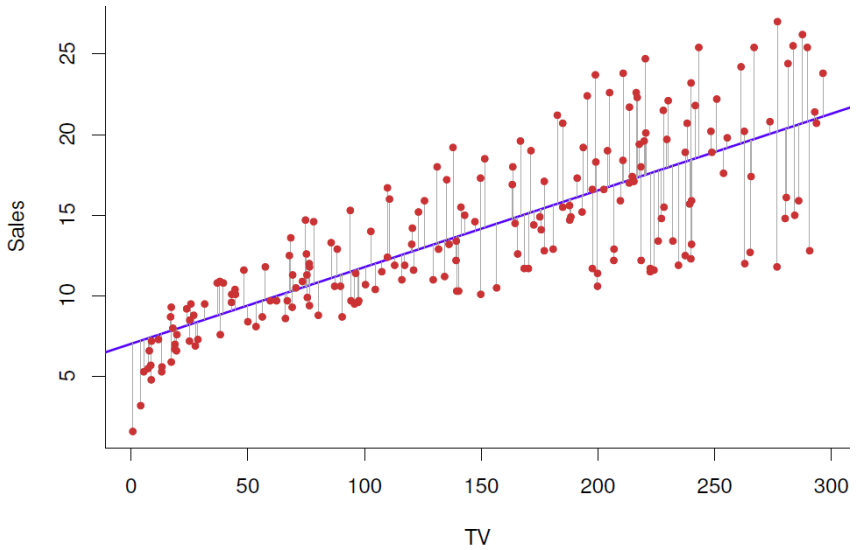
$$L(\theta_i, \hat{\theta}_i)$$



Loss Functions

$$L(\theta_i, \hat{\theta}_i)$$

Squared error $\sum_i (\theta_i - \hat{\theta}_i)^2$

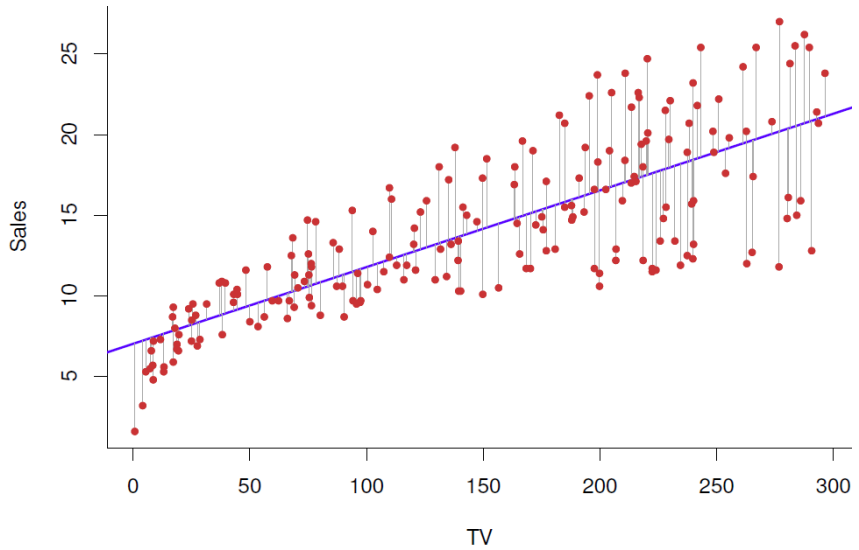


Loss Functions

$$L(\theta_i, \hat{\theta}_i)$$

Squared error $\sum_i (\theta_i - \hat{\theta}_i)^2$

Absolute error $\sum_i \left|\theta_i - \hat{\theta}_i\right|$



Classification Loss Functions

$$L(\theta_i, \hat{\theta}_i)$$

Classification Loss Functions

$$L(\theta_i, \hat{\theta}_i)$$

Indicator error

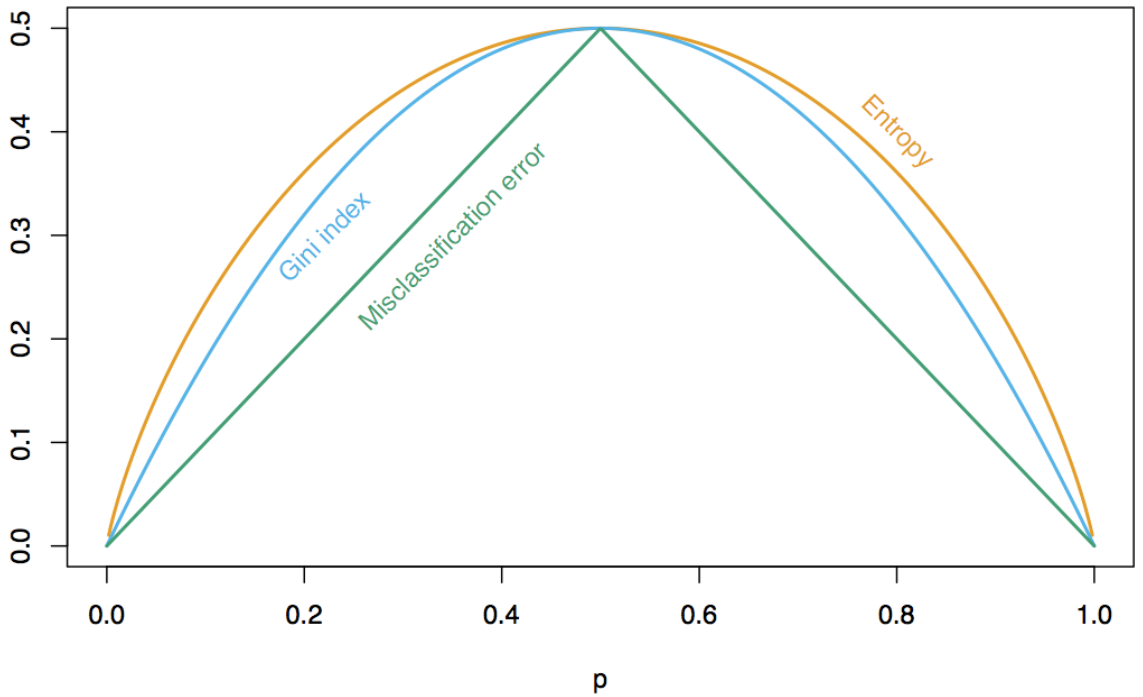
$$\sum_i I(\theta_i \neq \hat{\theta}_i)$$

Misclassification Impurity

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k)$$

Misclassification

For two classes
(also scaled to coincide at .5)



Hastie, Trevor, et al. The elements of statistical learning. Vol. 2. No. 1. New York: Springer, 2009.

Other Classification Loss Functions

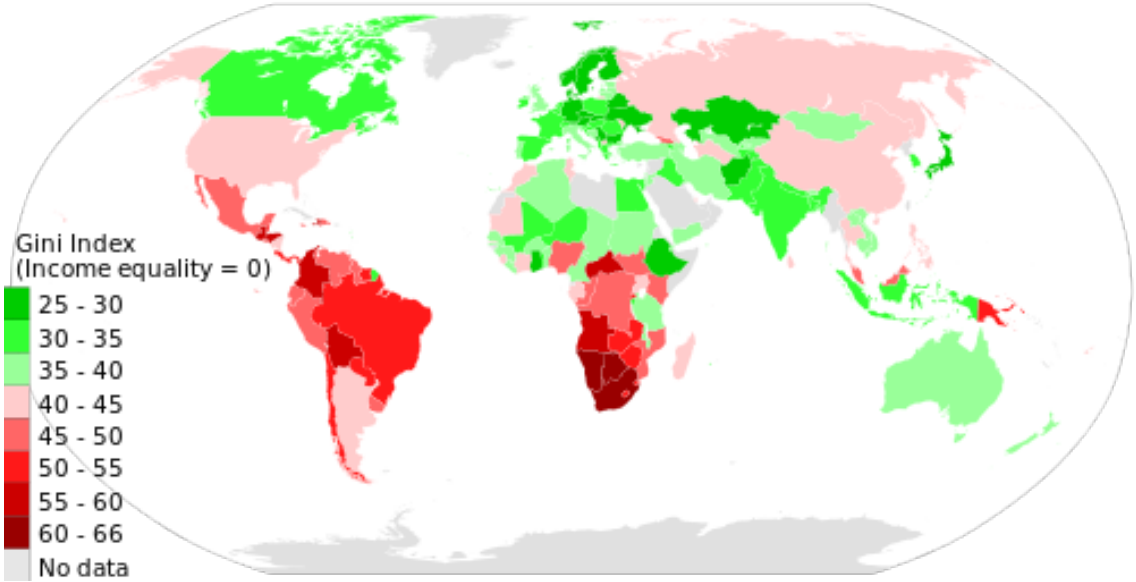
Misclassification error: $\frac{1}{N_m} \sum_{i \in R_m} I(y_i \neq k(m)) = 1 - \hat{p}_{mk(m)}.$

Gini index: $\sum_{k \neq k'} \hat{p}_{mk} \hat{p}_{mk'} = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}).$

Cross-entropy or deviance: $-\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}.$

Hastie, Trevor, et al. The elements of statistical learning. Vol. 2. No. 1. New York: Springer, 2009.

Aside: Gini Index



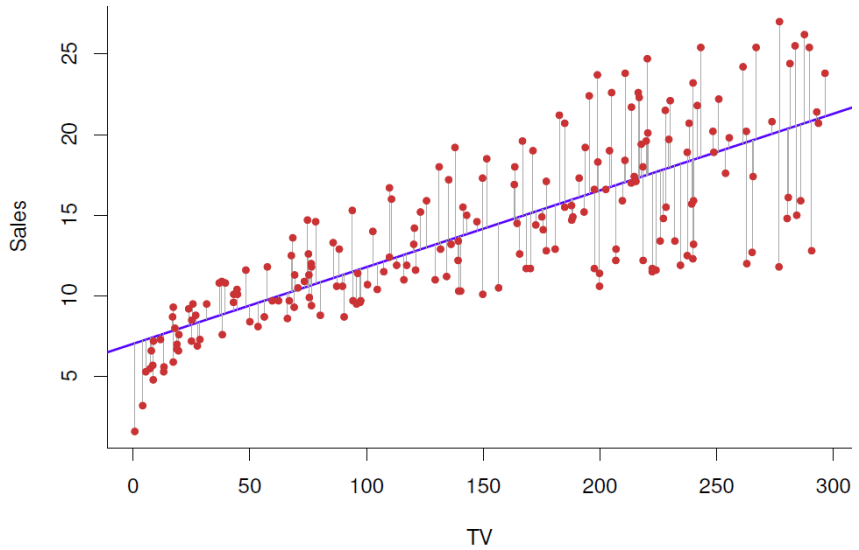
From wikipedia, “Gini Coefficient.”

Loss Functions

$$L(\theta_i, \hat{\theta}_i)$$

Squared error $\sum_i (\theta_i - \hat{\theta}_i)^2$

Absolute error $\sum_i \left|\theta_i - \hat{\theta}_i\right|$



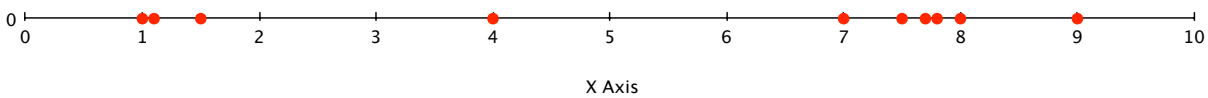
Question

What is the position with minimum squared displacement from a set of points?

$$argmin_{\bar{x}} \sum_i (x_i - \bar{x})^2$$

What is the position with minimum absolute value displacement from a set of points?

$$argmin_{\bar{x}} \sum_i |x_i - \bar{x}|$$



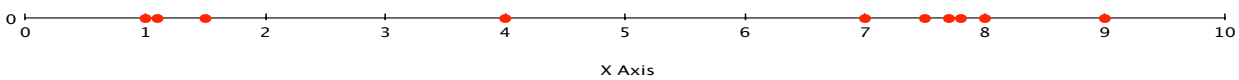
Question

What is the position with minimum distance (L2 norm) from a set of points?

The mean.

What is the position with minimum distance (L1 norm) from a set of points?

The median.



Supervised Learning Theory

$$L\left(Y, \tilde{f}(X)\right)$$

Supervised Learning Theory

$$E[L\left(Y, \tilde{f}(X)\right)]$$

Supervised Learning in one equation

$$\hat{f} = \underset{\tilde{f}}{\operatorname{argmin}} E[L(Y, \tilde{f}(X))]$$

Supervised Learning in one equation

$$\hat{f} = \underset{\tilde{f}}{\operatorname{argmin}} E[L(Y, \tilde{f}(X))]$$

Cross-validation

Cross-validation

"Estimating prediction error using error on a test set"

$$err = E[L(Y, \hat{f}(X))]$$

$$\approx$$

$$\text{average}_{\text{testset}}[L\left(Y, \hat{f}(X)\right)]$$

Cross-validation

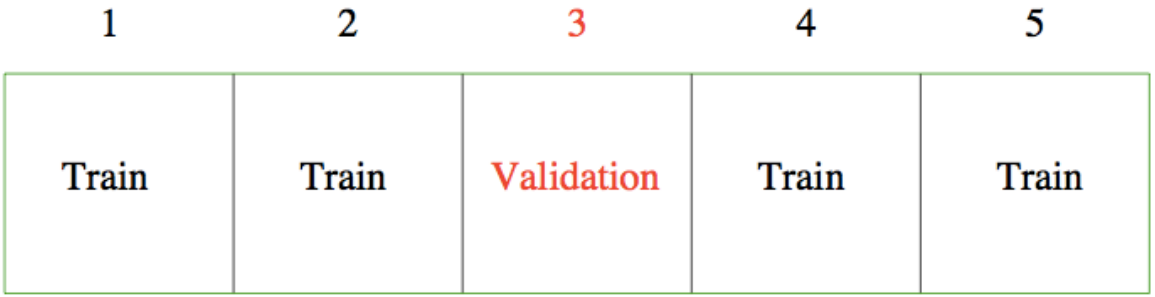
Training Set

Test Set

Validation Set

K-fold cross validation

5-fold example



$$\mathrm{CV}(\hat{f}) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{f}^{-\kappa(i)}(x_i))$$

Hastie, Trevor, et al. The elements of statistical learning. Vol. 2. No. 1. New York: Springer, 2009.

K-fold cross validation

5-fold and 10-fold are most common
(higher bias, lower variance)

N-fold is called “leave one out” cross validation
(lower bias, higher variance)

Hastie, Trevor, et al. The elements of statistical learning. Vol. 2. No. 1. New York: Springer, 2009.

Cross-validation

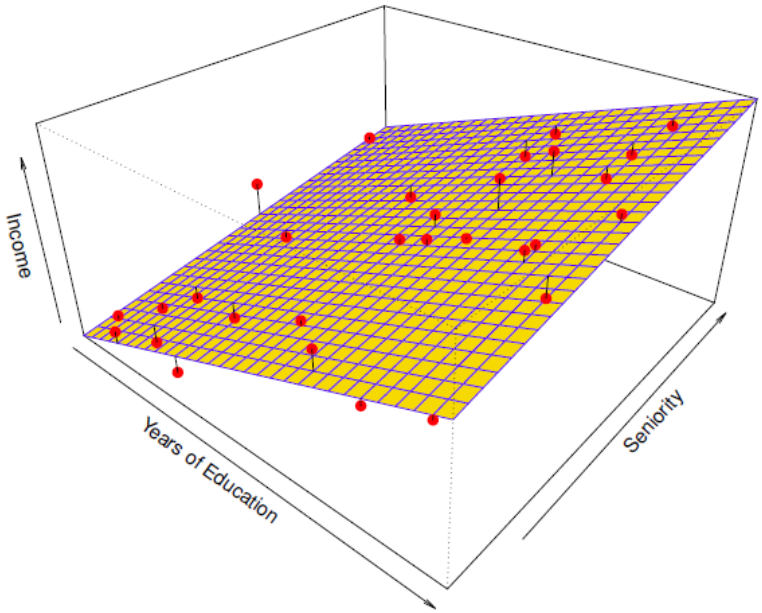
- Estimating prediction error

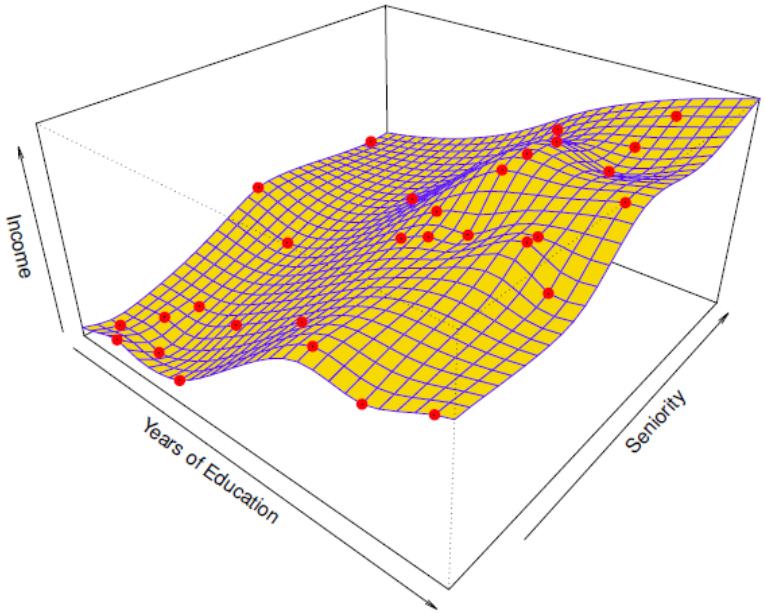
- Choosing appropriate values of model parameters (e.g. k parameter in k-nearest neighbors)

Challenge: Overfitting

- *Overfitting*: learning the random variation in the data rather than the underlying trend
- Characteristic of overfitting
  - good performance on previously-seen data, but poor performance on new data

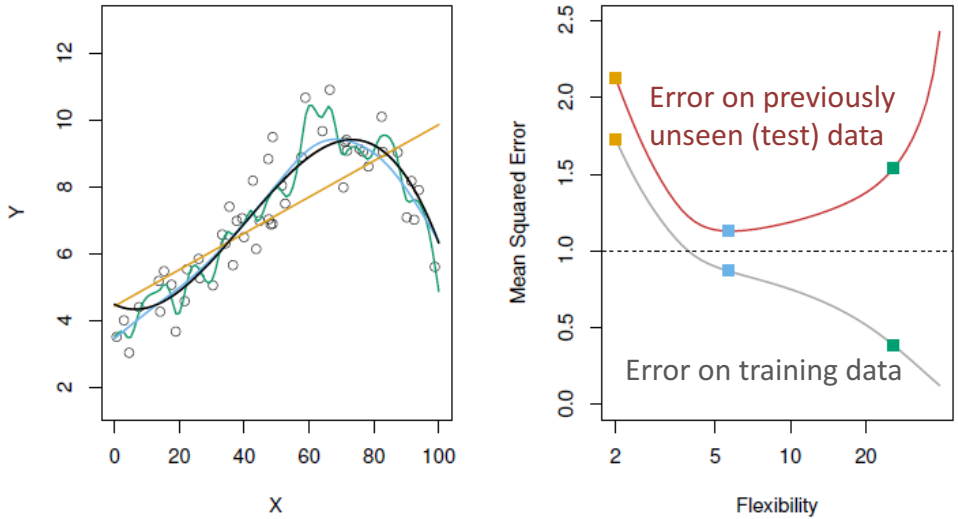
Overfitting





Figures 2.4 and 2.6 , ISL 2013

Overfitting



"With four parameters I can fit an elephant, and with five I can make him wiggle his trunk"

-John von Neumann according to Enrico Fermi

Figure 2.9 , ISL 2013

Choice of K

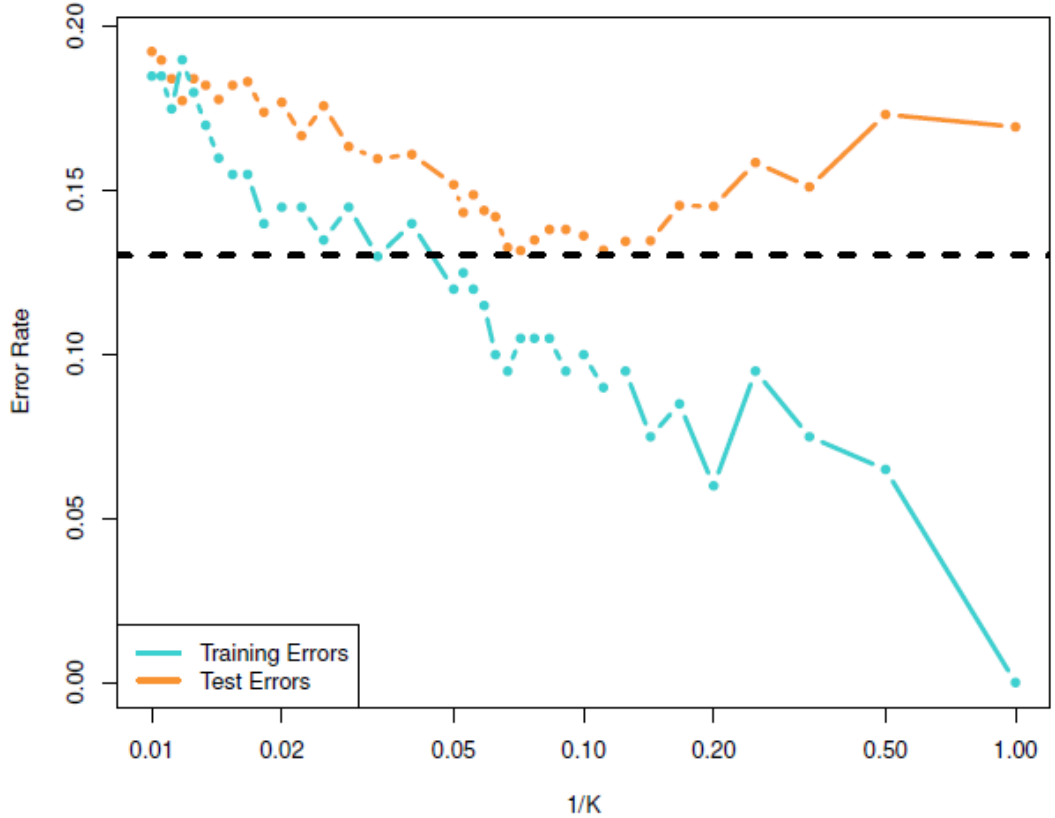
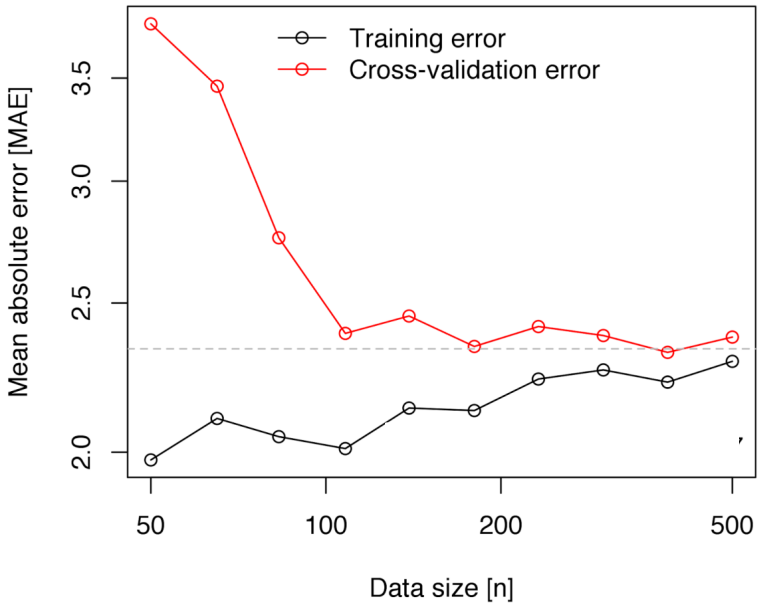


Figure 2.17, ISL 2013

Cross-validation (learning curves)



Do we need more features (new model) or more data?

Mistakes in cross-validation

Common mistakes involve training on the test set.

What's wrong?

Example 1-

You have 40 features. Using your entire dataset you identify the 10 features with the highest correlation to the output you are trying to predict. Now that you've identified the most important variables you fit a linear regression using just those variables on a training set (a subset of your data).*

You then estimate your prediction error by using this linear regression to predict the outputs for the test set that you held out at *.

What's wrong?

Example 2 –

You form a separate training and test set at the outset. You train k-nearest neighbors on the training set with k=5. You estimate the prediction error on a test set. The error is unfortunately bad. You try again with other k values (trying k=2 – 20) on the training set. You also decide to improve these methods' predictions by fitting their errors (so as to correct them) with a locally weighted regression or linear regression.

You estimate the prediction error on the test set for each of these combinations (k=2-20 followed by locally weighted regression or linear regression correction).

The lowest test error results from using k=6 followed by a simple linear regression correction. You will use this combination for future predictions.

Questions?

Regularization

A Reminder: Multiple Linear Regression

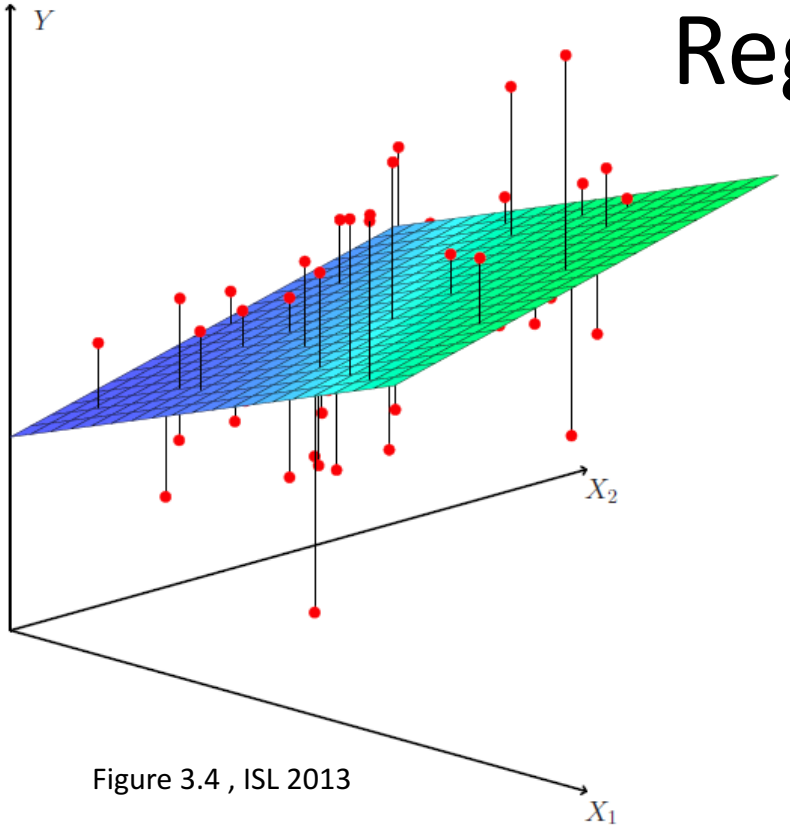


Figure 3.4 , ISL 2013

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2$$

Too Many Cooks

too many variables

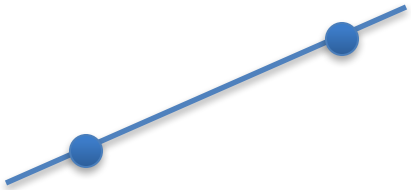
$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3 + \beta_4 \cdot X_4 + \beta_5 \cdot X_5 + \beta_6 \cdot X_6 + \beta_7 \cdot X_7 + \beta_8 \cdot X_8$$

ex. interaction terms and “proprietary” variables

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot (X_1 X_2) + \beta_4 \cdot X_1^2 + \beta_5 \cdot X_2^2 + \beta_6 \cdot \log(X_1 / X_2) + \beta_7 \cdot \sin(X_1 - X_2)$$

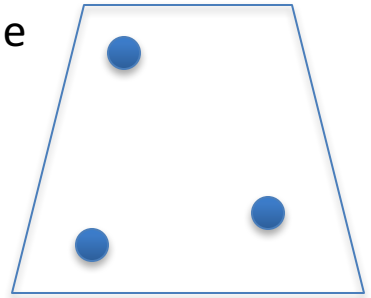
Too Many Cooks

Two samples define a line



$$Y = \beta_0 + \beta_1 \cdot X_1$$

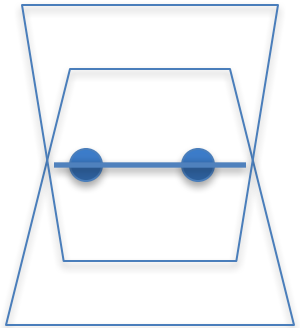
Three samples define a plane



$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2$$

Too Many Cooks

Two samples do not uniquely define a single plane

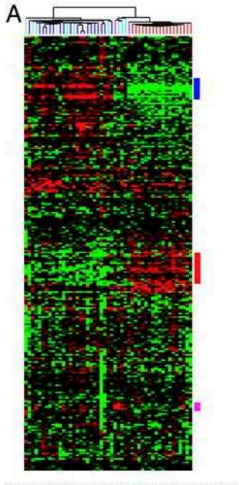


$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2$$

Too Many Cooks

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3 + \beta_4 \cdot X_4 + \beta_5 \cdot X_5 + \beta_6 \cdot X_6 + \beta_7 \cdot X_7 + \beta_8 \cdot X_8$$

Gene expression arrays



What else can go wrong?

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3 + \beta_4 \cdot X_4 + \beta_5 \cdot X_5 + \beta_6 \cdot X_6 + \beta_7 \cdot X_7 + \beta_8 \cdot X_8$$

Question: You have eight features, but hundreds of samples. Two of the features (X_3 and X_4) have little correlation with Y (and so are of little predictive use), but have a large correlation with each other. What might happen to the interpretability of your β coefficients?

Multi-colinearity

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3 + \beta_4 \cdot X_4 + \beta_5 \cdot X_5 + \beta_6 \cdot X_6 + \beta_7 \cdot X_7 + \beta_8 \cdot X_8$$

What do we do?

Penalize large β coefficients.

Multiple Linear Regression

Recall that in linear regression we are trying to minimize the squared error,

$$\sum_{samples} [Y - (\beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2)]^2$$

Ridge Regression

Find β values that minimize the “penalized” error, which equals,

$$\sum_{samples} [Y - (\beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2)]^2 + \lambda \cdot (\beta_1^2 + \beta_2^2)$$

L2

Regularization

Ridge Regression

Find β values that minimize the "penalized" error, which equals

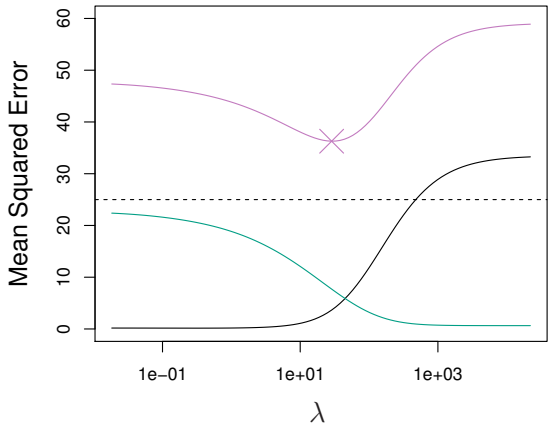
$$\sum_{samples} [Y - (\beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2)]^2 + \lambda \cdot (\beta_1^2 + \beta_2^2)$$

L2

Or written another way,

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left(\|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2 \right)$$

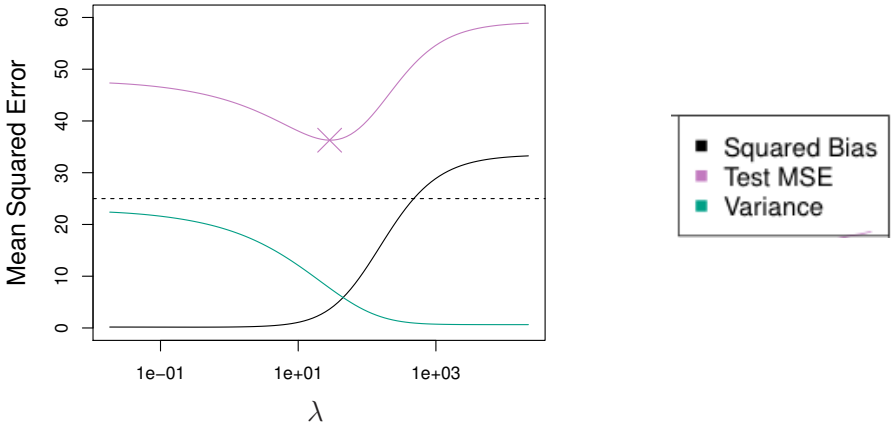
Ridge Regression



Which curve is bias, which is variance, and which is prediction error on the test set?

Hastie, Trevor, et al. Introduction to statistical learning.

Ridge Regression



Hastie, Trevor, et al. Introduction to statistical learning.

Regularization

Now we've addressed:

- *Underdetermined*
- *Overfitting*
- *Multi-colinearity*

What about *sparsity*?

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3 + \beta_4 \cdot X_4 + \beta_5 \cdot X_5 + \beta_6 \cdot X_6 + \beta_7 \cdot X_7 + \beta_8 \cdot X_8$$

$\beta_2 \to 0$, $\beta_4 \to 0$, $\beta_7 \to 0$

Sparsity

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3 + \beta_4 \cdot X_4 + \beta_5 \cdot X_5 + \beta_6 \cdot X_6 + \beta_7 \cdot X_7 + \beta_8 \cdot X_8$$

$\beta_2 \to 0$ $\beta_4 \to 0$ $\beta_7 \to 0$

- For interpretability, “feature selection”
- For computational efficiency

Sparsity

Lasso

"**L**east **a**bsolute **s**hrinkage and **s**election **o**perator"

$$\sum_{samples} [Y - (\beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2)]^2 + \lambda \cdot (|\beta_1| + |\beta_2|)$$

L1

Tibshirani, Robert. "Regression shrinkage and selection via the lasso." Journal of the Royal Statistical Society. Series B (Methodological) (1996): 267-288.

Lasso

"**L**east **a**bsolute **s**hrinkage and **s**election **o**perator"

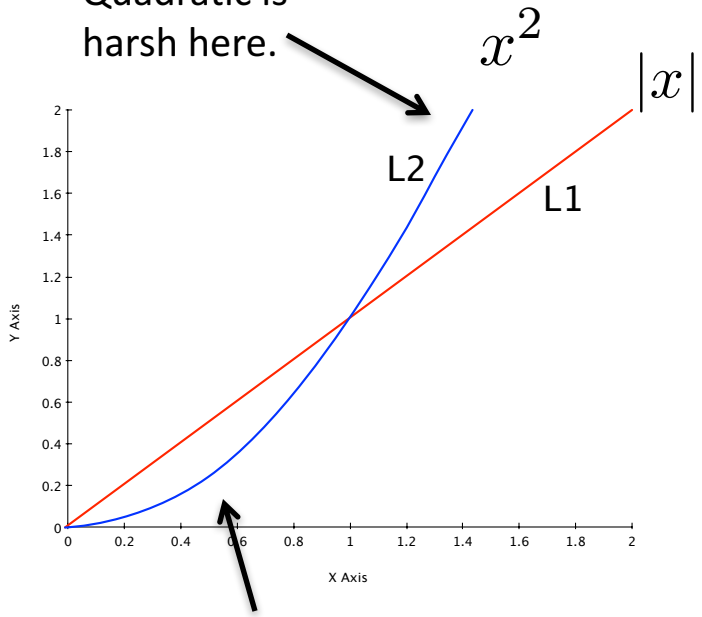
$$\sum_{samples} [Y - (\beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2)]^2 + \lambda \cdot (|\beta_1| + |\beta_2|)$$

Or written another way,

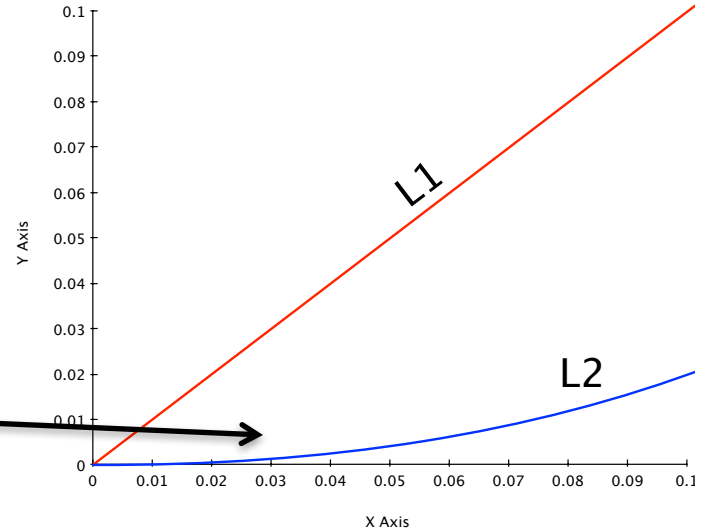
$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left(\|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right)$$

Penalties

Quadratic is harsh here.



Quadratic is kind here.



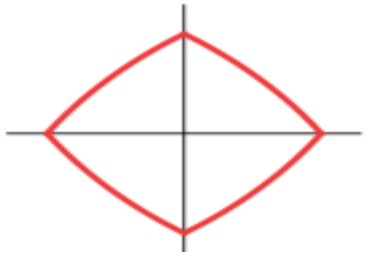
Other Techniques

- Elastic Net
- Principal Component Regression

Ridge + Lasso =

Elastic Net

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}(\|y - X\beta\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1).$$



$\sum_j (\alpha\beta_j^2 + (1-\alpha)|\beta_j|)$ *for* $\alpha = 0.2$

glmnet command in R

Hastie, Trevor, et al. The elements of statistical learning. Vol. 2. No. 1. New York: Springer, 2009.

Questions?

Trees

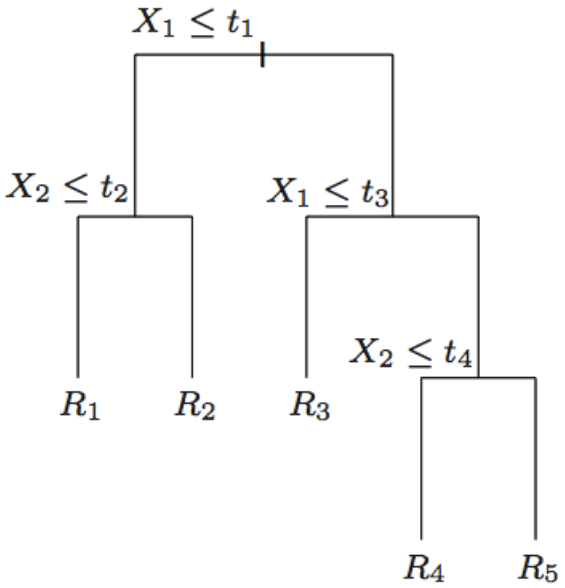
Alexander Ioannidis
ioannidis@stanford.edu

Institute for Computational and Mathematical Engineering,
Stanford University

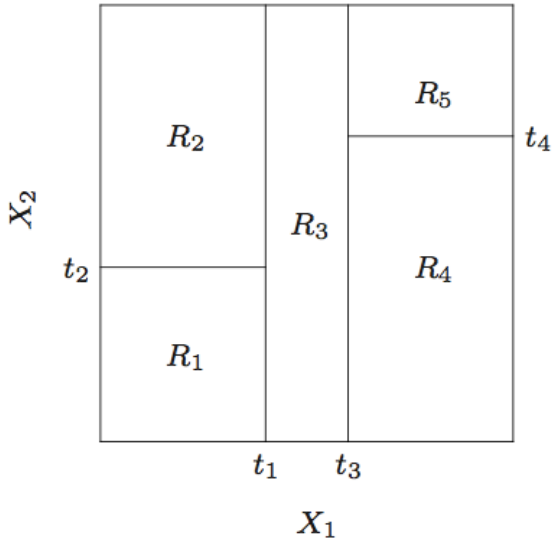
Classification and Regression Trees

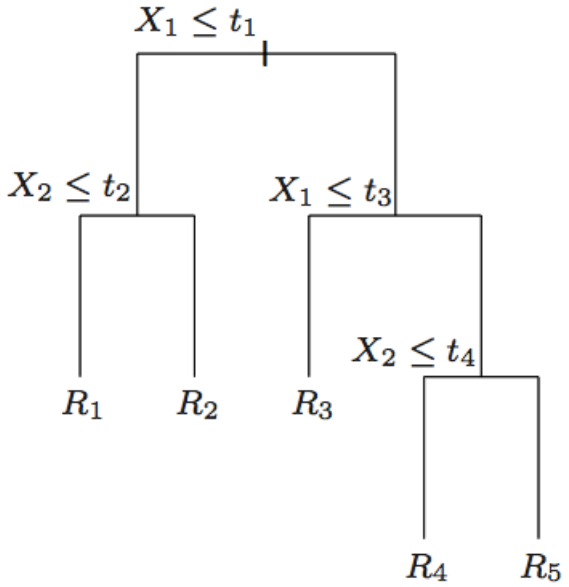
Recursive binary partitions of the feature space

Classification and Regression Trees



Classification and Regression Trees

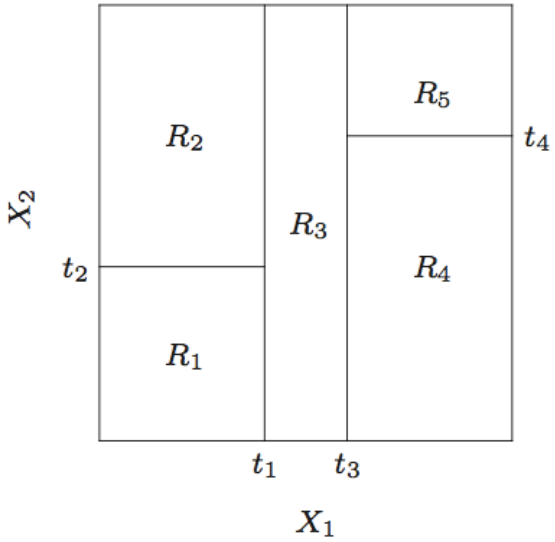




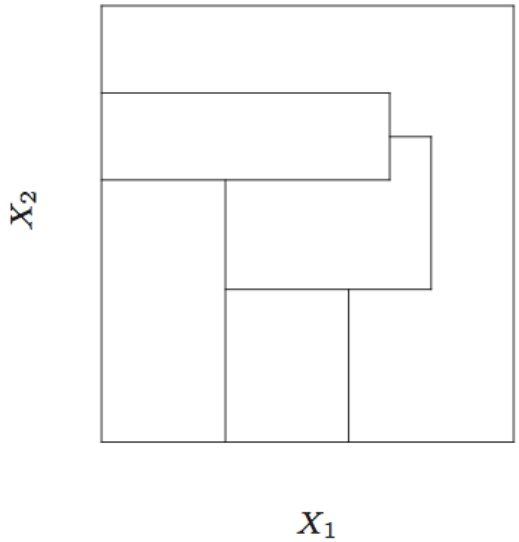
Hastie, Trevor, et al. The elements of statistical learning. Vol. 2. No. 1. New York: Springer, 2009.

Classification and Regression Trees

YES



NO



Hastie, Trevor, et al. The elements of statistical learning. Vol. 2. No. 1. New York: Springer, 2009.

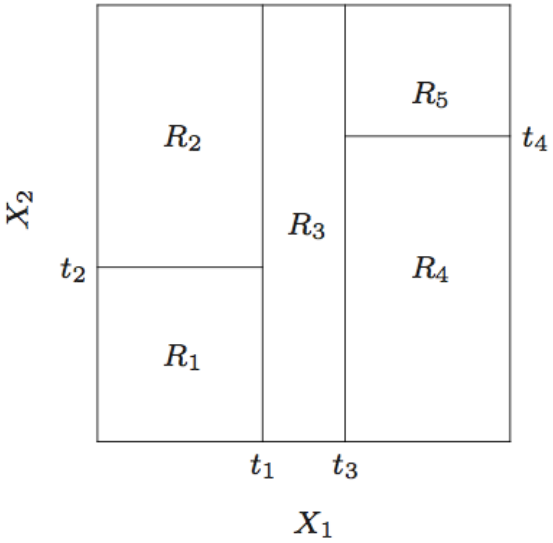
Classification and Regression Trees

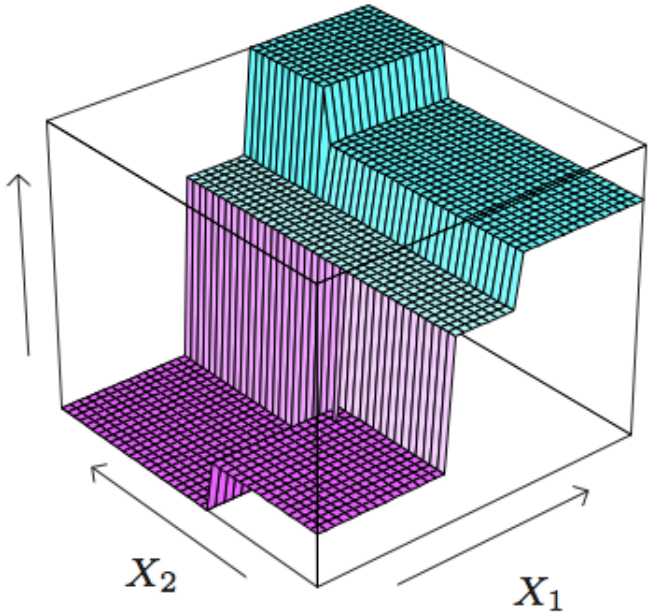
- Constant prediction within each region

$$f(x) = \sum_{m=1}^{M} c_m I(x \in R_m)$$

Hastie, Trevor, et al. The elements of statistical learning. Vol. 2. No. 1. New York: Springer, 2009.

Classification and Regression Trees





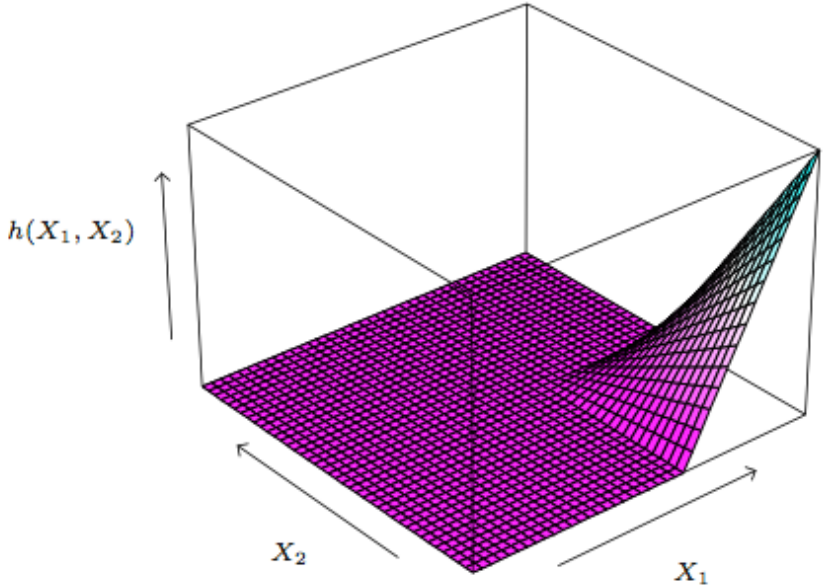
Hastie, Trevor, et al. The elements of statistical learning. Vol. 2. No. 1. New York: Springer, 2009.

Reminder: Machine Learning in one equation

$$\hat{f} = \underset{\tilde{f}}{\operatorname{argmin}} E[L(Y, \tilde{f}(X))]$$

Aside: Piecewise linear model (MARS)

Multivariate Adaptive Regression Splines



Hastie, Trevor, et al. The elements of statistical learning. Vol. 2. No. 1. New York: Springer, 2009.

Classification and Regression Trees

1 TYPE OF HOME
- 1. House
- 2. Condominium
- 3. Apartment
- 4. Mobile Home
- 5. Other

2 SEX
- 1. Male
- 2. Female

3 MARITAL STATUS
- 1. Married
- 2. Living together, not married
- 3. Divorced or separated
- 4. Widowed
- 5. Single, never married

4 AGE
- 1. 14 thru 17
- 2. 18 thru 24
- 3. 25 thru 34
- 4. 35 thru 44

5 EDUCATION
- 1. Grade 8 or less
- 2. Grades 9 to 11
- 3. Graduated high school
- 4. 1 to 3 years of college
- 5. College graduate
- 6. Grad Study

6 OCCUPATION
- 1. Professional/Managerial
- 2. Sales Worker
- 3. Factory Worker/Laborer/Driver
- 4. Clerical/Service Worker
- 5. Homemaker
- 6. Student, HS or College
- 7. Military
- 8. Retired
- 9. Unemployed

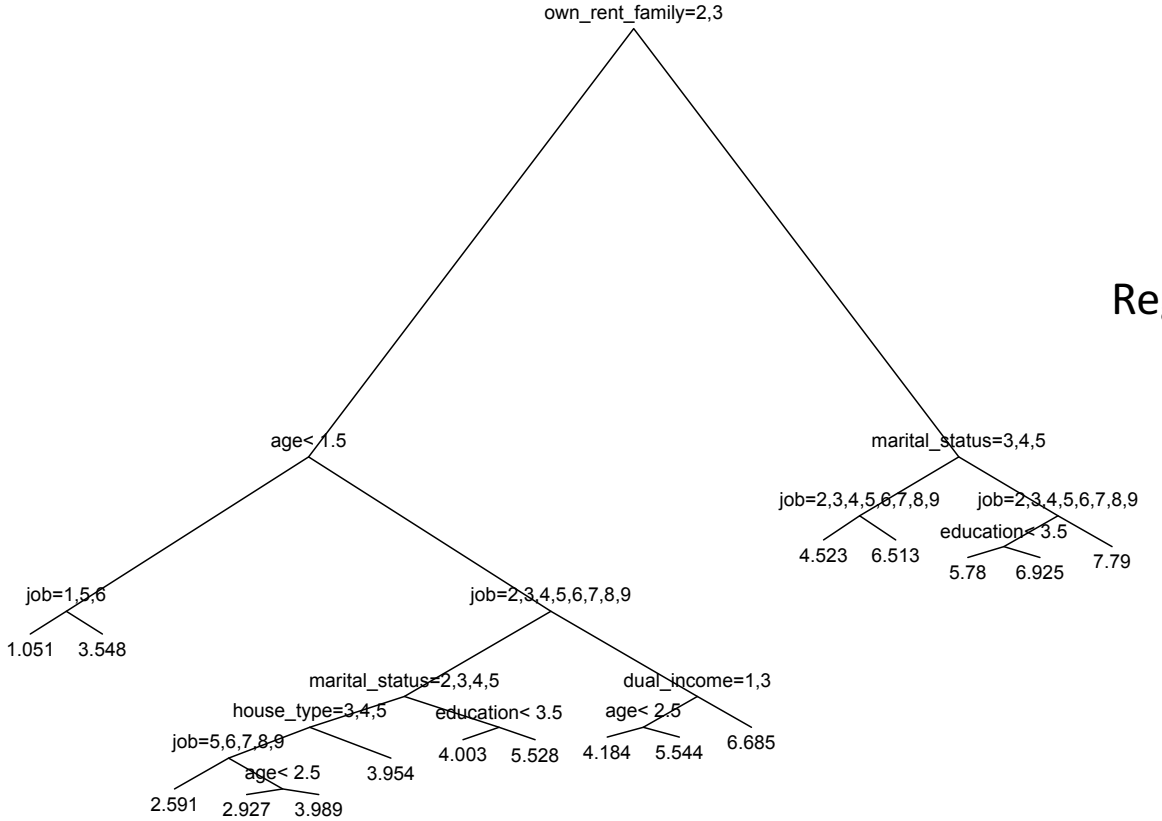
7 ANNUAL INCOME OF HOUSEHOLD (PERSONAL INCOME IF SINGLE
- 1. Less than $10,000
- 2. $10,000 to $14,999
- 3. $15,000 to $19,999
- 4. $20,000 to $24,999
- 5. $25,000 to $29,999
- 6. $30,000 to $39,999

8 HOW LONG HAVE YOU LIVED IN THE SAN FRAN./OAKLAND/SAN JOSE AREA?
- 1. Less than one year
- 2. One to three years
- 3. Four to six years
- 4. Seven to ten years
- 5. More than ten years

9 DUAL INCOMES (IF MARRIED)
- 1. Not Married
- 2. Yes
- 3. No

10 PERSONS IN YOUR HOUSEHOLD
- 1. One
- 2. Two
- 3. Three
- 4. Four
- 5. Five
- 6. Six
- 7. Seven
- 8. Eight
- 9. Nine or more

Classification and Regression Trees



Regression

Regression trees

$$f(x) = \sum_{m=1}^{M} c_m I(x \in R_m)$$

$$\hat{c}_m = \text{ave}(y_i | x_i \in R_m)$$

Classification trees

$$\text{class } k(m) = \arg\max_k \hat{p}_{mk}$$

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k)$$

Greedy algorithm

- Find splitting variable j and split point s that minimize prediction error

Greedy algorithm: classification

- Multiple metrics for prediction error (node impurity)

Misclassification error: $\frac{1}{N_m} \sum_{i \in R_m} I(y_i \neq k(m)) = 1 - \hat{p}_{mk(m)}$

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k)$$

Hastie, Trevor, et al. The elements of statistical learning. Vol. 2. No. 1. New York: Springer, 2009.

Categorical predictors, too many combinations

- $2^{q-1} - 1$ possible splits of q unordered categories
- Improve computation time by ordering the categories based on their mean outcome values

Missing predictor values

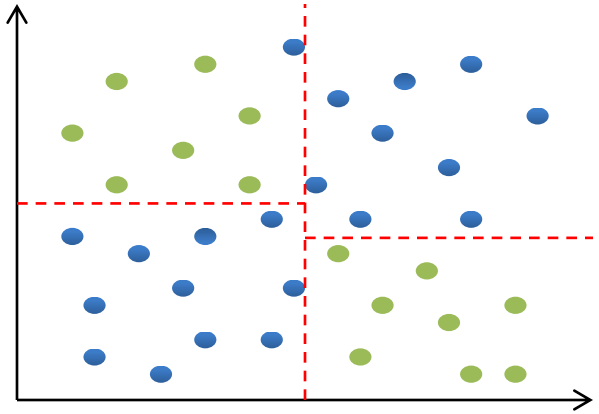
- Store “surrogate” predictors and split points, since predictors are often correlated

- Don’t throw data away when building tree!

Avoiding overfitting

- Stopping criterion? e.g. minimum decrease in prediction error

Problem:



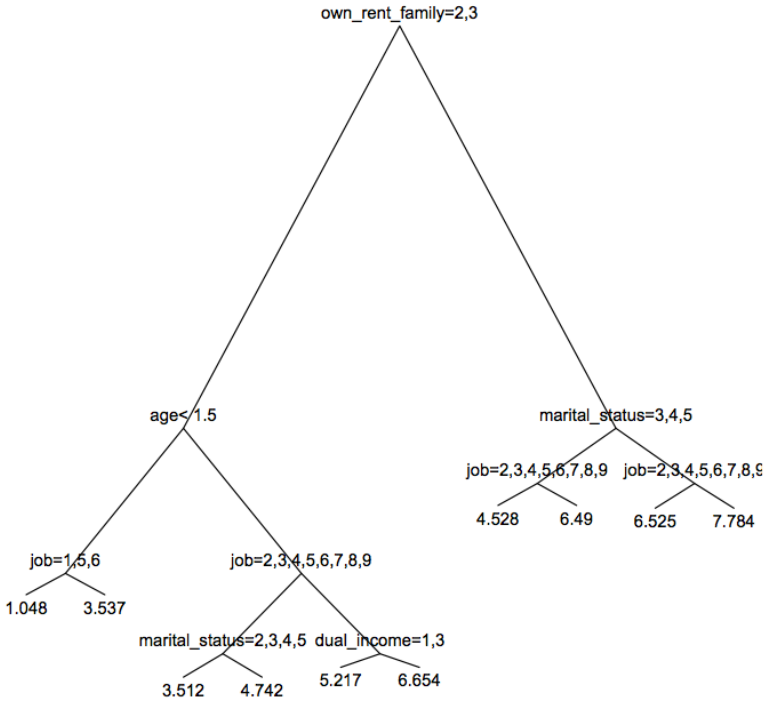
Avoiding overfitting

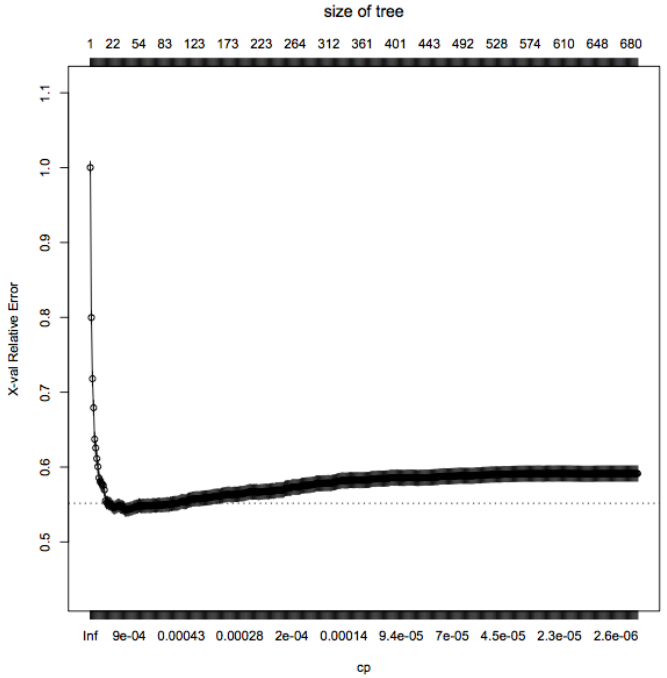
- Pruning:

  Grow large tree T_0

  Prune to some subtree $T \subset T_0$

Cost-complexity pruning



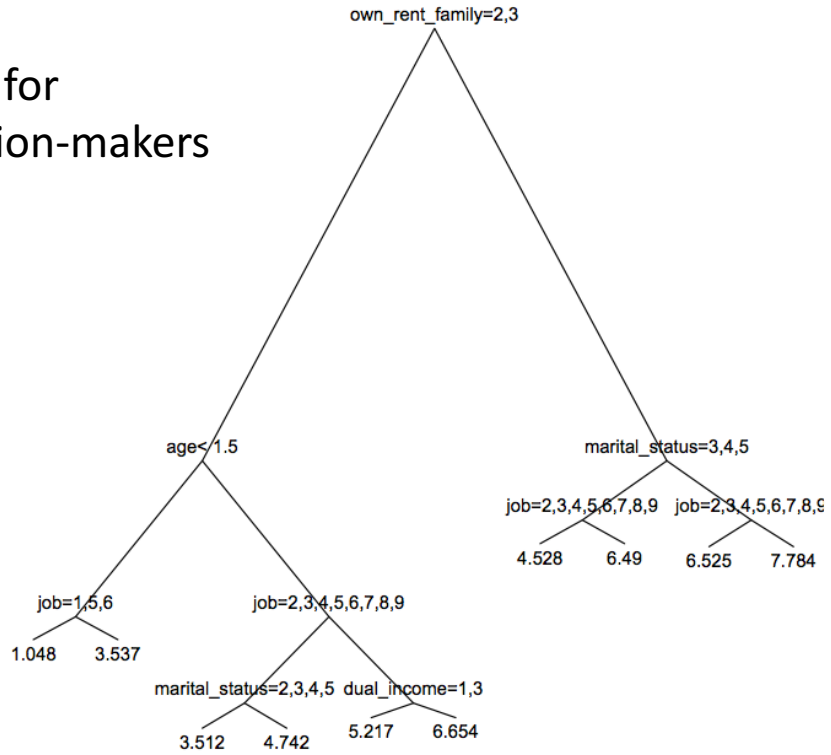


Advantages of CART

1. Handles missing data easily (surrogate splits)
2. Robust to non-informative data
3. Automatic variable selection
4. Easily interpretable, ideal for explaining “why” to decision-makers
5. Captures high order interactions

Advantages of CART

Easily interpretable, ideal for explaining “why” to decision-makers



Advantages of CART

Captures high order interactions

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots$$

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \gamma_1 x_1 x_2 + \gamma_2 x_1 x_3 + \gamma_3 x_2 x_3 + \zeta_1 x_1 x_2 x_3 \ldots$$

Y = 3.5 if ((1<marital_status<6) AND (1<job<9)) AND (age<1.5) OR …

Disadvantages of CART

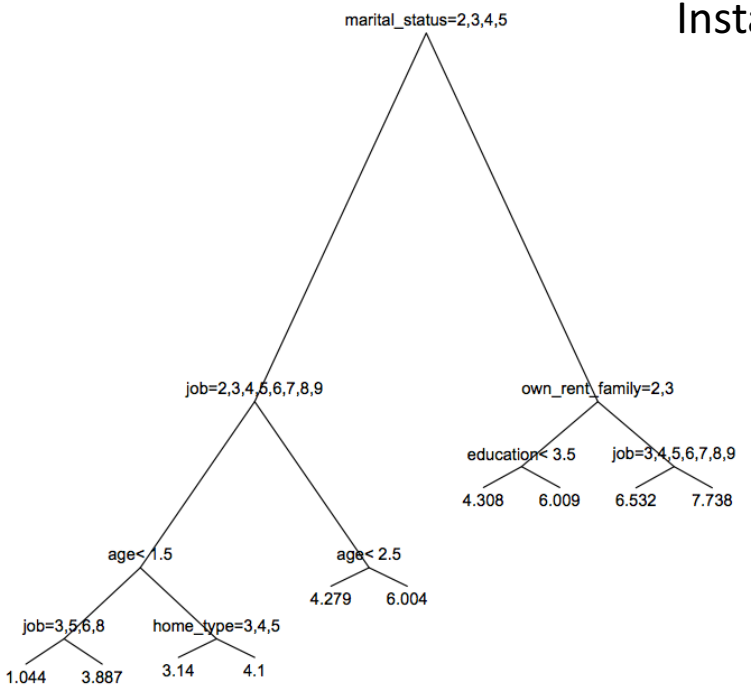
1. Instability of trees

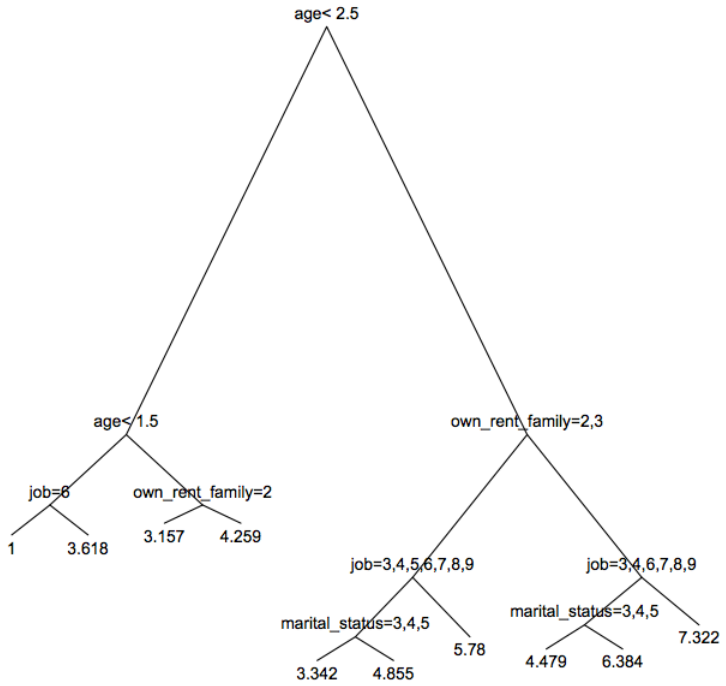
2. Lack of smoothness

3. Hard to capture additivity

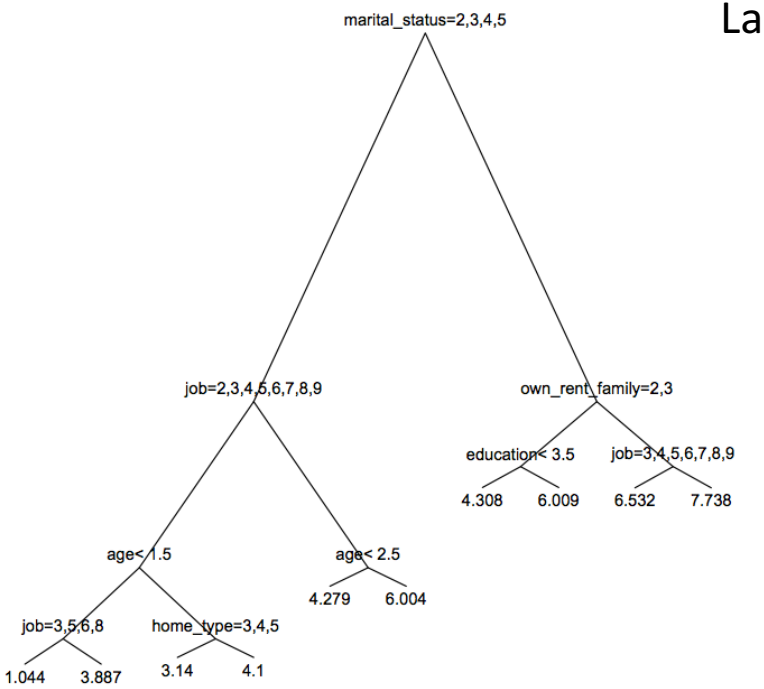
Disadvantages of CART

Instability of trees

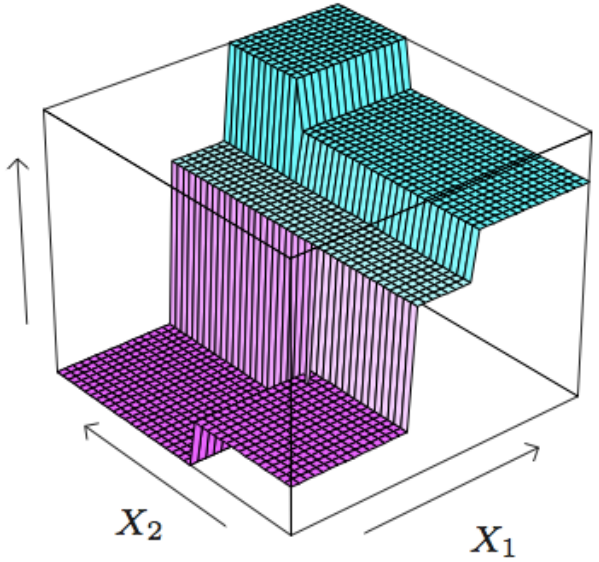




Disadvantages of CART



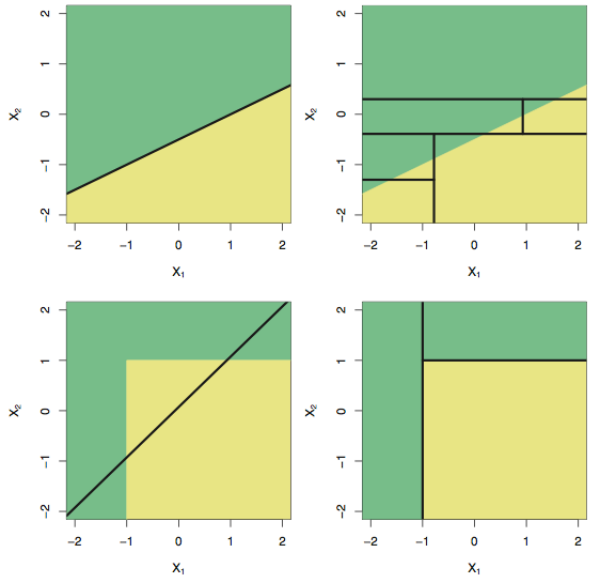
Lack of Smoothness



Disadvantages of CART

Hard to capture additivity

$$Y = c_1 I (X_1 < t_1) + c_2 I (X_2 < t_2) + e$$



Hastie, Trevor, et al. Introduction to statistical learning.

Disadvantages of CART

1. Instability of trees

   Solution - Random Forests

2. Lack of smoothness

   Solution - MARS

3. Hard to capture additivity

   Solution – MART or *MARS*

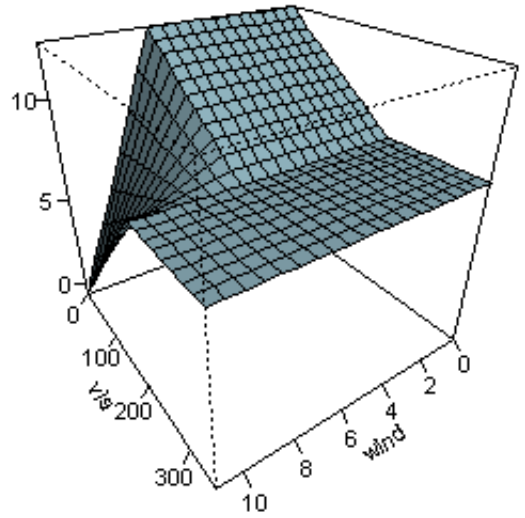
Extensions

- MART – “Multiple Additive Regression Trees”

- MARS – “Multivariate Adaptive Regression Splines”

MARS – “Multivariate Adaptive Regression Splines”

- Invented by Jerome Friedman in 1991



Ensemble Methods: Boosting, Bagging, and Random Forests

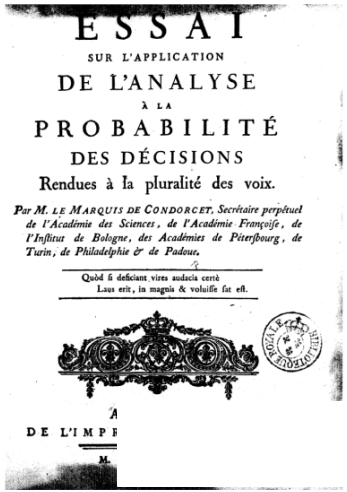
Alexander Ioannidis
ioannidis@stanford.edu

Institute for Computational and Mathematical Engineering,
Stanford University

Ensemble Methods

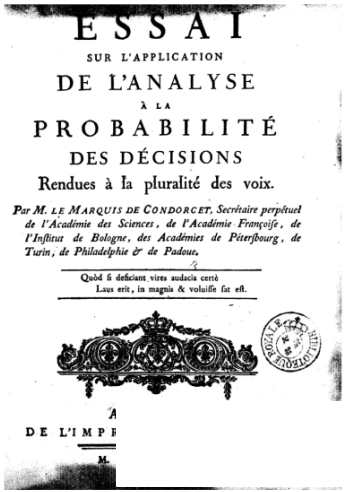
The strength of weak classifiers

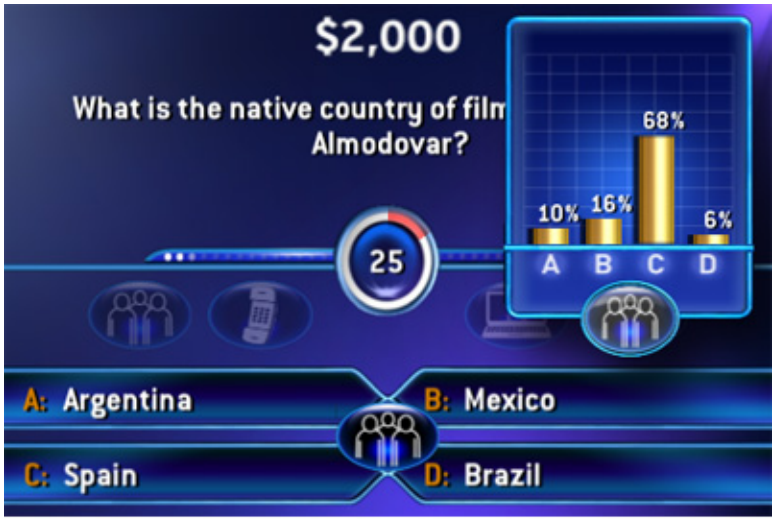
Condorcet's Jury Theorem - If p is greater than 1/2 (each voter is more likely to vote correctly), then adding more voters increases the probability that the majority decision is correct. In the limit, the probability that the majority votes correctly approaches 1 as the number of voters increases.


ESSAI
SUR L'APPLICATION
DE L'ANALYSE
À LA
PROBABILITÉ
DES DÉCISIONS
Rendues à la pluralité des voix.
Par M. LE MARQUIS DE CONDORCET, Secrétaire perpétuel de l'Académie des Sciences, de l'Académie Françoise, de l'Institut de Bologne, des Académies de Pétersbourg, de Turin, de Philadelphie & de Padoue.
Quòd si deficiant vires audacia certè
Laus erit, in magnis & voluisse sat est.

The strength of weak classifiers

Condorcet's Jury Theorem - If p is greater than 1/2 (each voter is more likely to vote correctly), then adding more voters increases the probability that the majority decision is correct. In the limit, the probability that the majority votes correctly approaches 1 as the number of voters increases.


ESSAI
SUR L'APPLICATION
DE L'ANALYSE
À LA
PROBABILITÉ
DES DÉCISIONS
Rendues à la pluralité des voix.
Par M. LE MARQUIS DE CONDORCET, Secrétaire perpétuel de l'Académie des Sciences, de l'Académie Françoise, de l'Institut de Bologne, des Académies de Pétersbourg, de Turin, de Philadelphie & de Padoue.
Quòd fi deficiant vires audacia certè
Laus erit, in magnis & voluiffe fat eft.
DE L'IMPR



The strength of weak classifiers

- Averaging reduces variance without raising bias (bias remains unchanged)

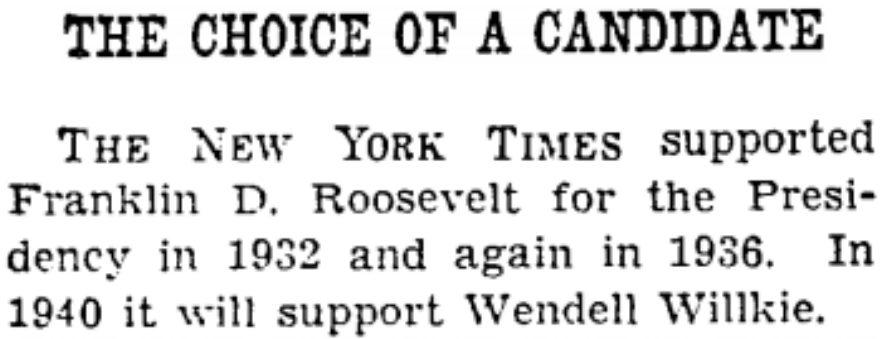
$$\mathrm{Var}[\bar{Y}] = \sigma^2/n$$

The strength of weak classifiers

- Averaging reduces variance without raising bias (bias remains unchanged)

$$\mathrm{Var}[\bar{Y}] = \sigma^2/n$$

- The votes of correlated classifiers don't help as much



THE CHOICE OF A CANDIDATE

THE NEW YORK TIMES supported Franklin D. Roosevelt for the Presidency in 1932 and again in 1936. In 1940 it will support Wendell Willkie.

The strength of weak classifiers

- Averaging reduces variance without raising bias (bias remains unchanged)
  $\mathrm{Var}[\bar{Y}] = \sigma^2/n$

  - The votes of correlated classifiers don't help as much
    $\mathrm{Var}[\bar{Y}] = \sigma^2/n + (\rho\sigma^2)(n-1)/n$

The strength of weak classifiers

- Averaging reduces variance without raising bias (bias remains unchanged)

$$\mathrm{Var}[\bar{Y}] = \sigma^2/n$$

- The votes of correlated classifiers don't help as much → Random Forest

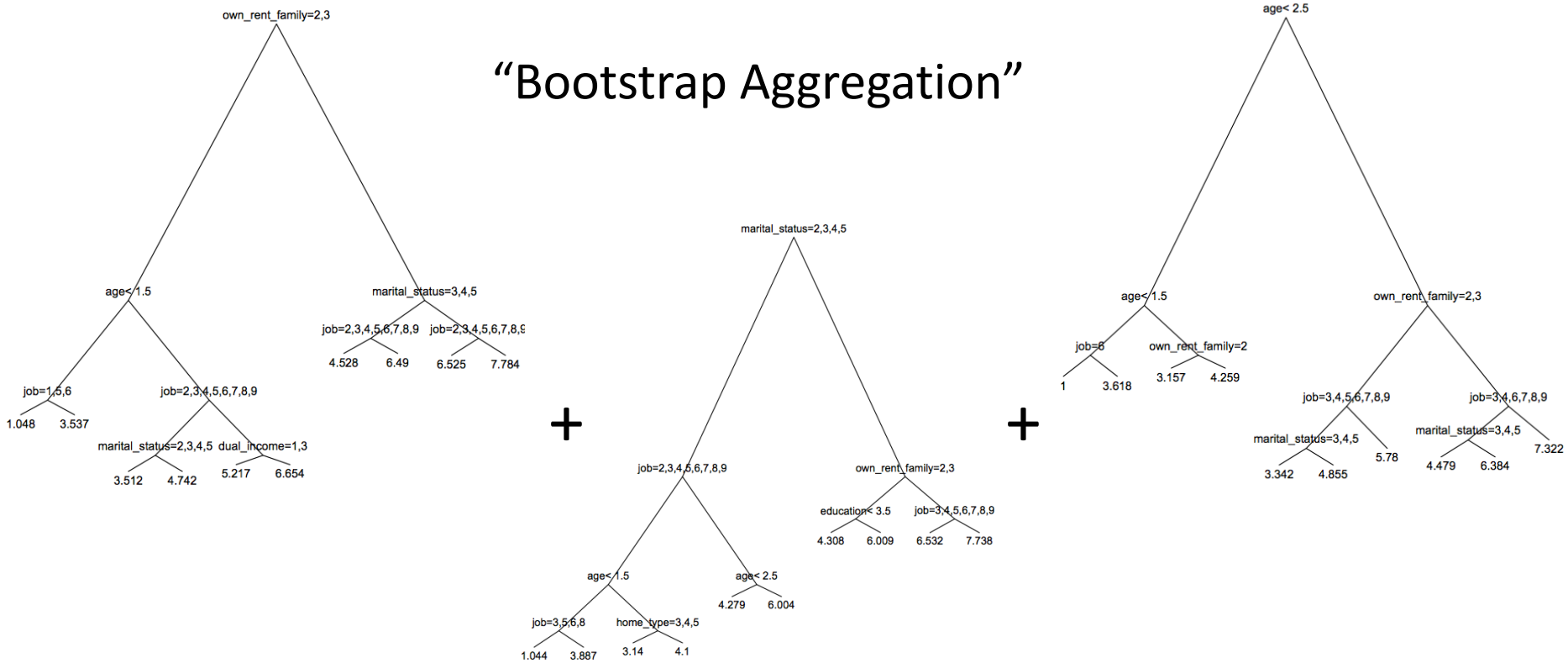
Be Creative

$$\alpha \cdot \{CART\} + (1 - \alpha) \cdot \{LinearModel\}$$

Ensemble Methods: Bagging

What is bagging?

"Bootstrap Aggregation"



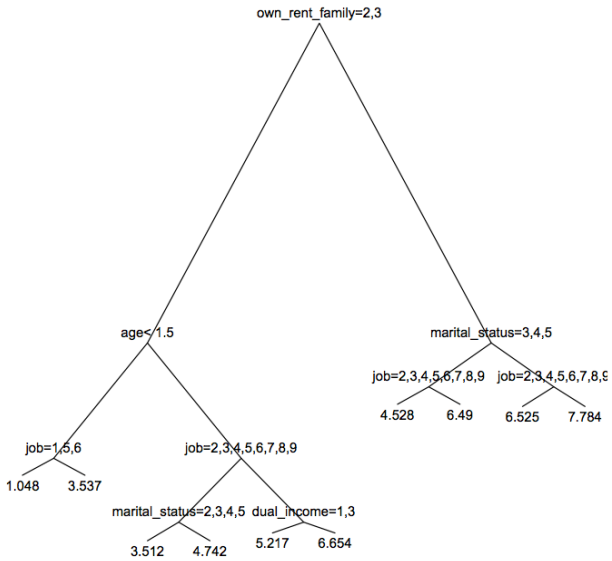
What is bagging?

"Bootstrap Aggregation"

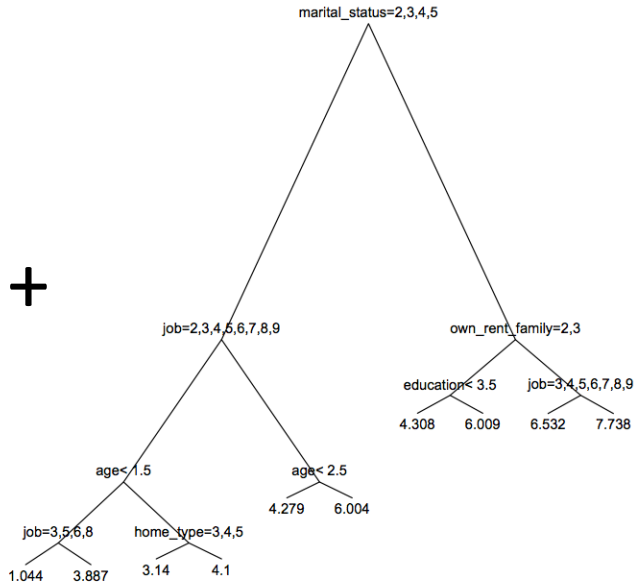
$$\hat{f}_{\text{bag}}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{*b}(x)$$

Bagging

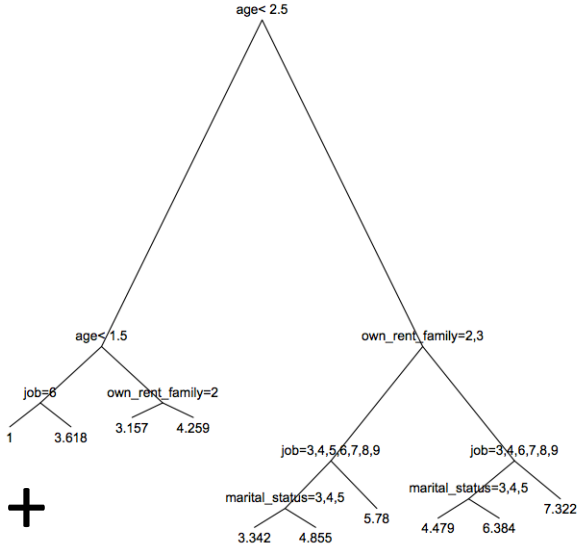
Address the instability of CART



+

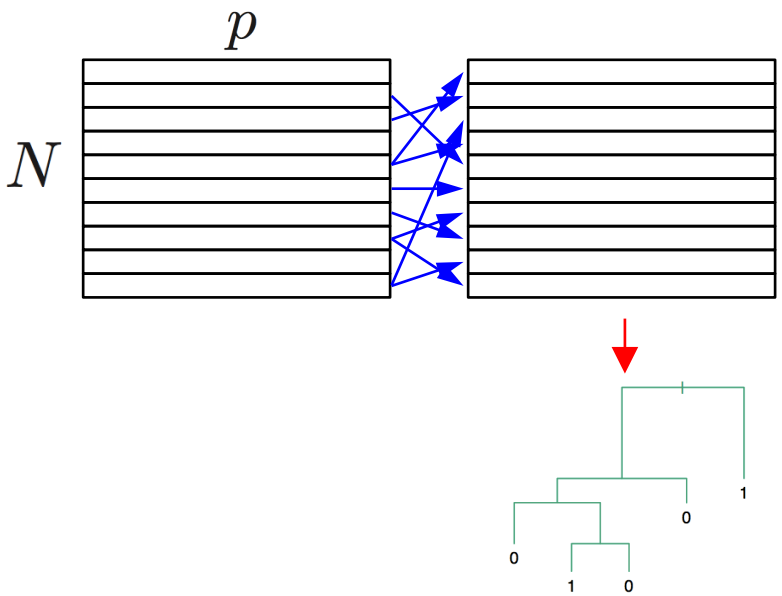


+



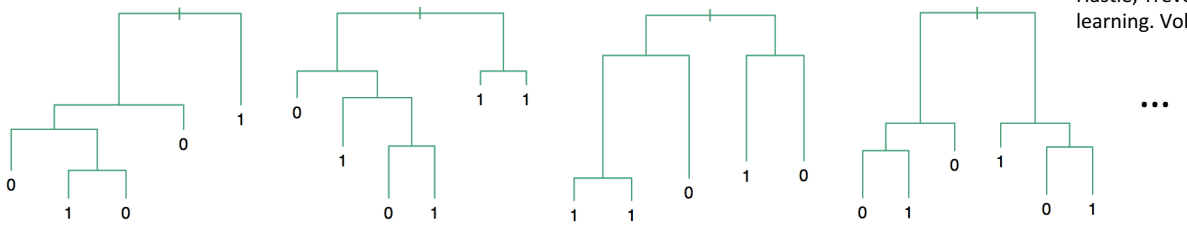
Bagging

- Bootstrap the training data samples to build an ensemble of predictors.



Bagging

- Bootstrap the training data samples to build an ensemble of predictors.



Hastie, Trevor, et al. The elements of statistical learning. Vol. 2. No. 1. New York: Springer, 2009.

- Average (or majority vote) the individual predictions.
- Bagging reduces variance and maintains bias.

Bagging

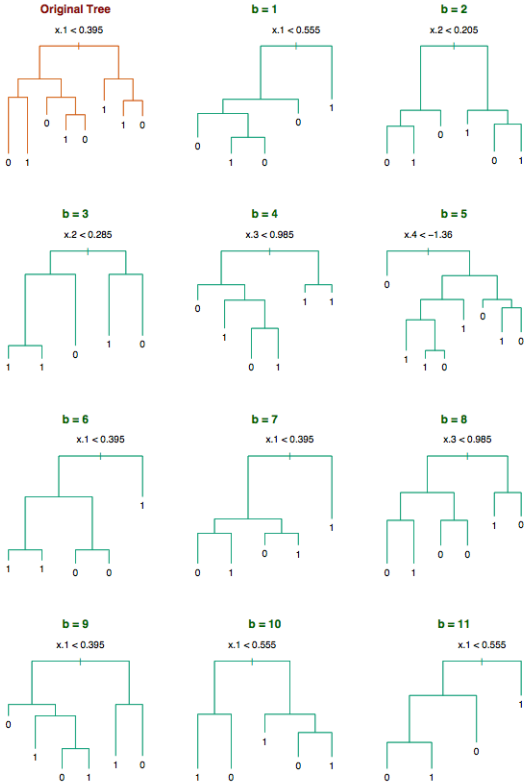
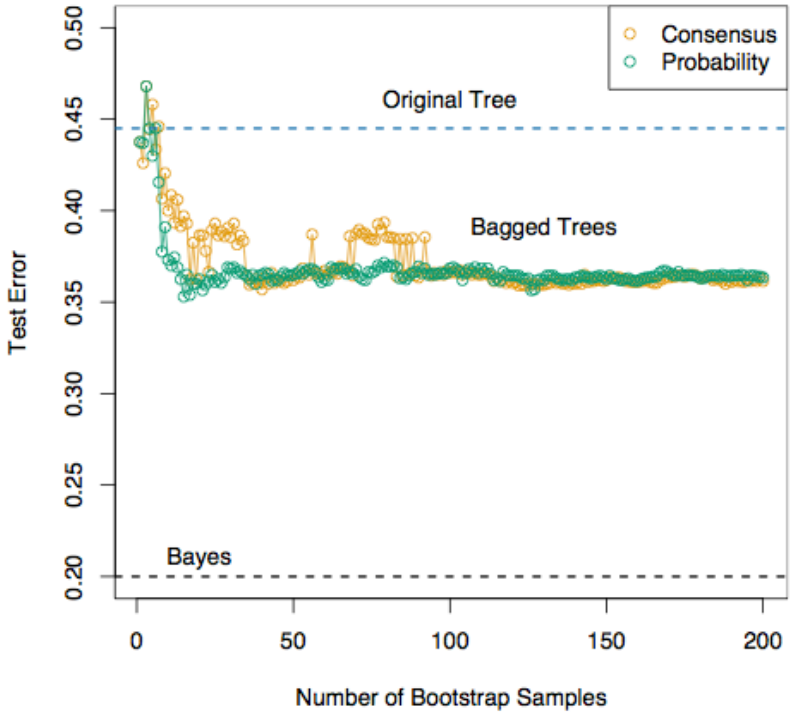


FIGURE 8.9. *Bagging trees on simulated dataset. The top left panel shows the original tree. Eleven trees grown on bootstrap samples are shown. For each tree, the top split is annotated.*

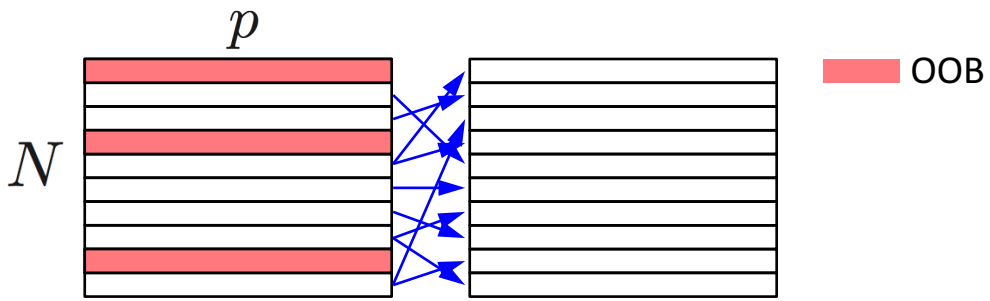
Hastie, Trevor, et al. The elements of statistical learning. Vol. 2. No. 1. New York: Springer, 2009.



Bonus! Out-of-bag cross-validation

Out-of-bag (OOB) samples

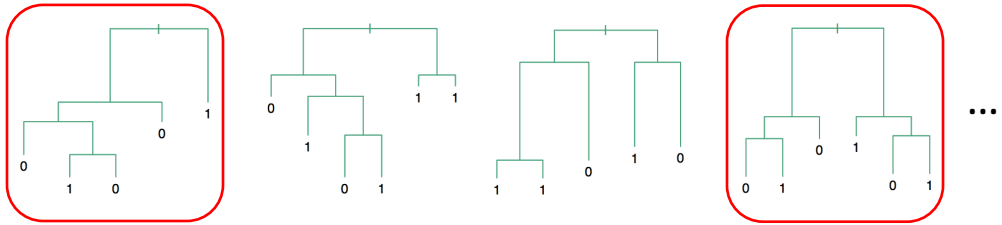
- Bootstrapping process:



- Each tree uses only a subset of the training samples (~2/3 of samples on average).
- Each sample is OOB for ~1/3 of trees.

Predictions for OOB samples

- For each sample, find the trees for which it is OOB.



- Predict its value from each of those trees.
- Estimate prediction error of the bagged trees using all of the OOB predictions.
- Similar to cross-validation.

Random Forests

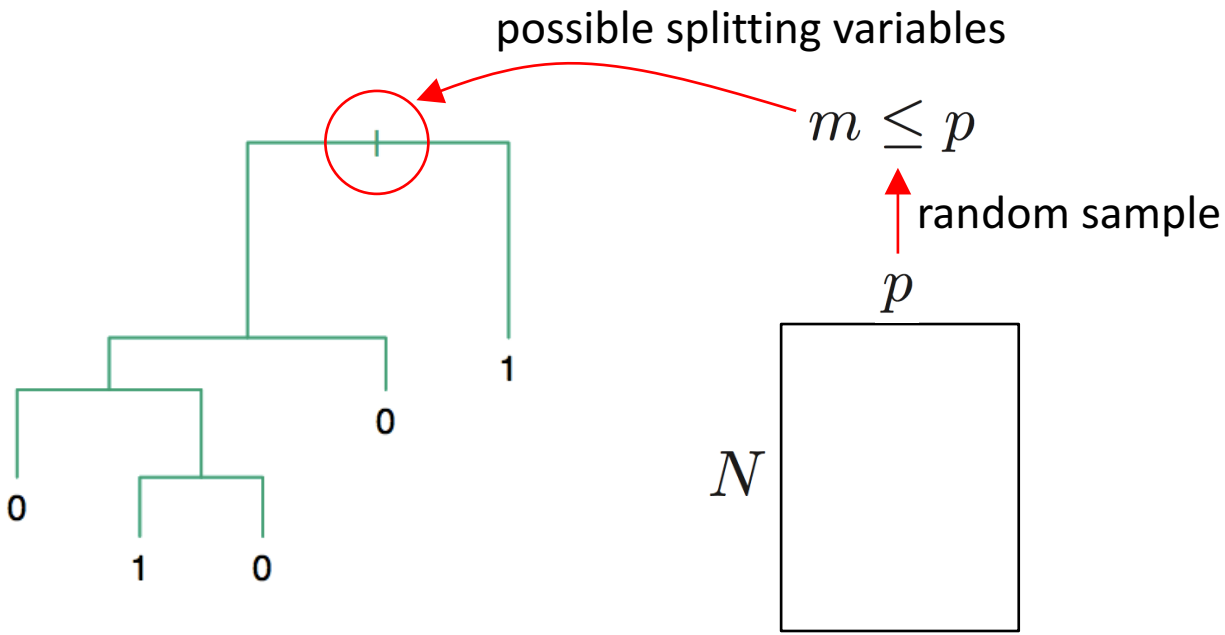
Bagged trees vs. random forests

- Bagging introduces variability between trees by random selection of training data.
- Bagged trees can still be correlated, limiting the reduction in variance.

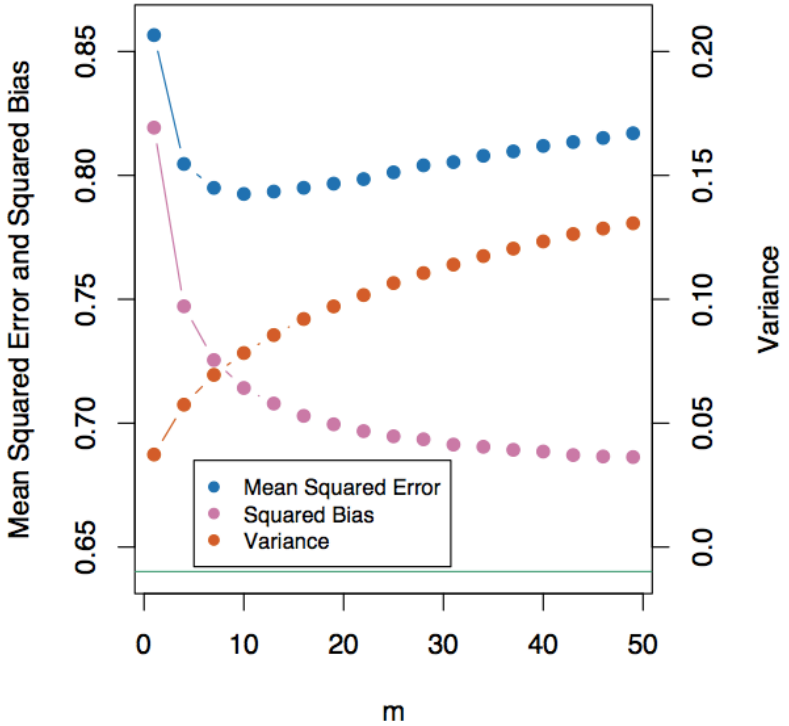
Random forests introduce additional randomness:

- Reduce correlation between trees by randomizing the variables considered for splitting at each node.

Candidate splitting variables



Candidate splitting variables



Hastie, Trevor, et al. The elements of statistical learning. Vol. 2. No. 1. New York: Springer, 2009.

Important parameters

Parameters of random forests:

- # of candidate splitting variables at each node (m)
- Depth of each tree (minimum node size)
- # of trees

Number of splitting variables

Default values

Classification $m = \lfloor \sqrt{p} \rfloor$

Regression $m = \lfloor p/3 \rfloor$

Important parameters

Parameters of random forests:

- # of candidate splitting variables at each node (m)
- Depth of each tree (minimum node size)
- # of trees

Tree depth (minimum node size)

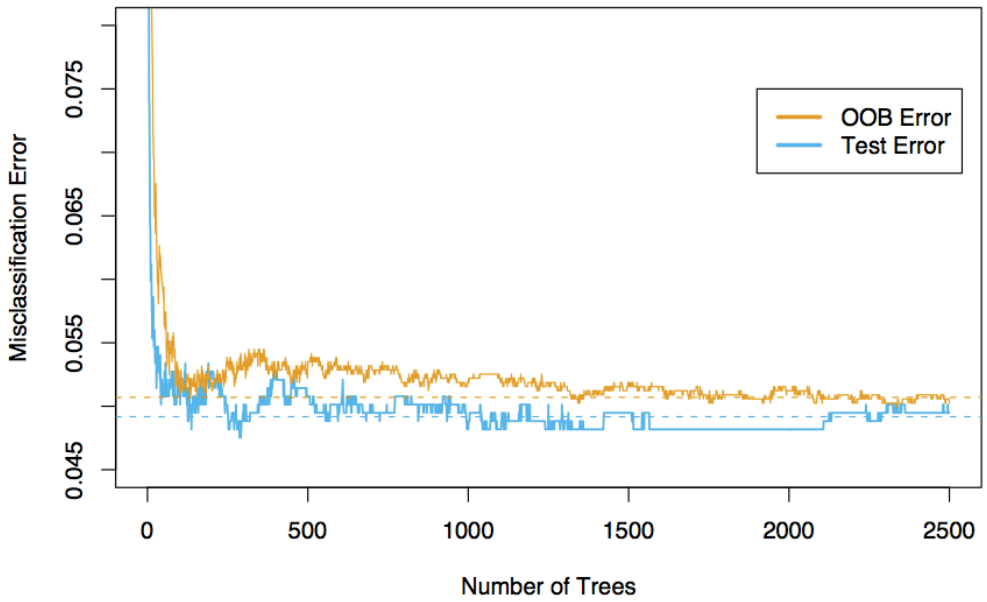
	Default values
Classification	1
Regression	5

Important parameters

Parameters of random forests:

- # of candidate splitting variables at each node (m)
- Depth of each tree (minimum node size)
- # of trees

Number of trees



Hastie, Trevor, et al. The elements of statistical learning. Vol. 2. No. 1. New York: Springer, 2009.

- Adding more trees does not cause overfitting.

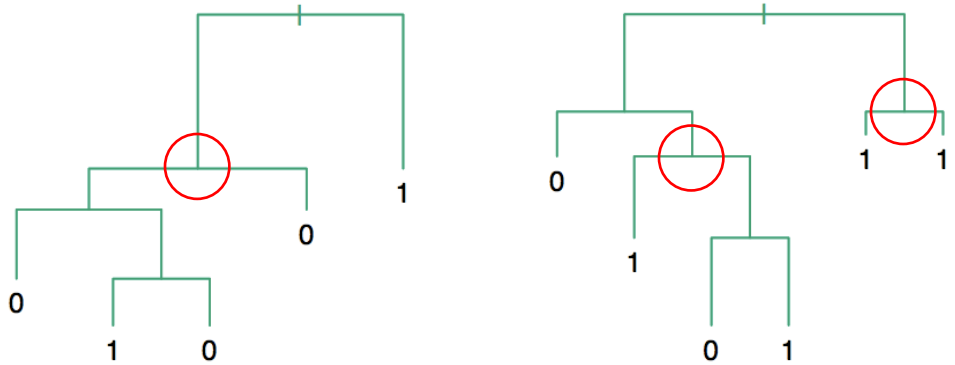
Other features of random forests

- Out-of-bag (OOB) samples
- Variable importance measurements

Variable importance

Metric 1:

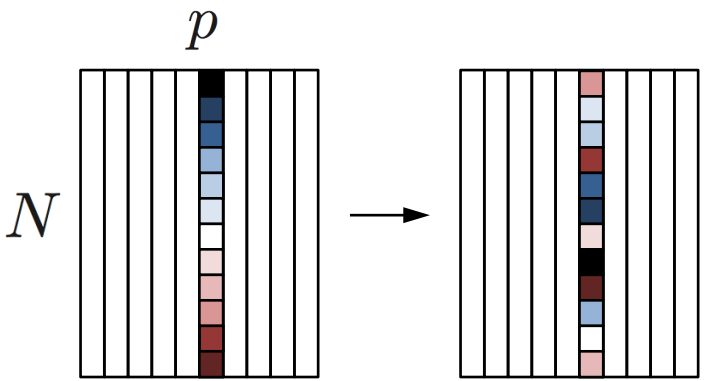
Decrease in prediction error or impurity from all splits involving that variable, averaged over trees.



Variable importance

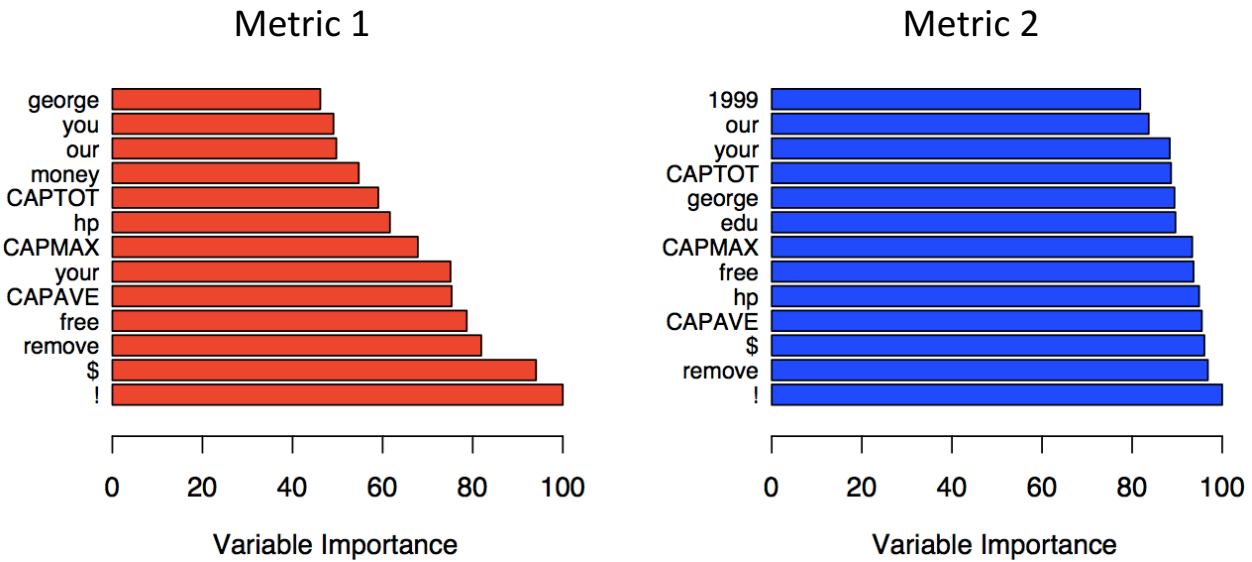
Metric 2:

Increase in overall prediction error when the values of that variable are randomly permuted between samples.



Variable importance example

- The metrics give similar but not identical rankings:



Hastie, Trevor, et al. The elements of statistical learning. Vol. 2. No. 1. New York: Springer, 2009.

Advantages of random forests

Similar to CART:

- Relatively robust to non-informative variables (built-in variable selection)
- Capture high-order interactions between variables
- Low bias
- Naturally handle mixed predictors (quantitative and categorical)

Advantages of random forests

Advantages over CART:

- Lower variance (more robust to choice of training data due to bootstrapping)
- Less prone to overfitting
- No need for pruning
- Built-in cross-validation (using OOB samples)

Disadvantages of random forests

Similar to CART:

- Hard to capture additive effects

Disadvantages relative to CART:

- Hard to interpret/explain the model predictions

Questions?

Computational Consulting

We put the power of experienced graduate students and top researchers within your grasp to answer your trickiest computation questions. We offer a wide range of specialties which include:

- Matrix Problems
- Machine Learning
- PDEs & Simulations
- Optimization
- Discrete Math
- HPC

Send us an email with a description of your mathematical challenge and let us get started helping you.

c2questions@list.stanford.edu