CS109 – Data Science

Verena Kaynig-Fittkau

vkaynig@seas.harvard.edu
staff@cs109.org

Announcements

- Register your teams until Thursday!
- Next coming up: Survey for actual project proposal
- Will be due 11/17

What would you like to see in class?

Books

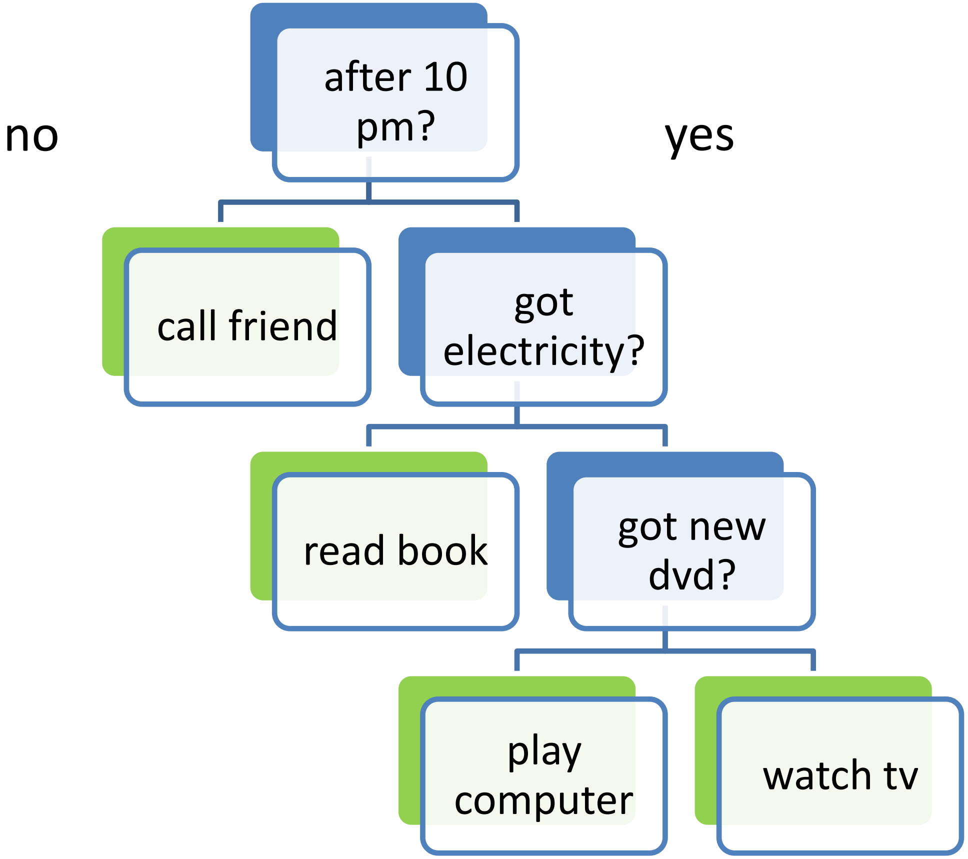
- “Elements of Statistical Learning”
- http://statweb.stanford.edu/~tibs/ElemStatLearn/

- “Pattern Recognition and Machine Learning”
- http://research.microsoft.com/en-us/um/people/cmbishop/PRML/

Next Topics

- Classification and regression trees (CART)
- Bagging
- Random Forest
- Boosting
- Cascade

Decision Tree

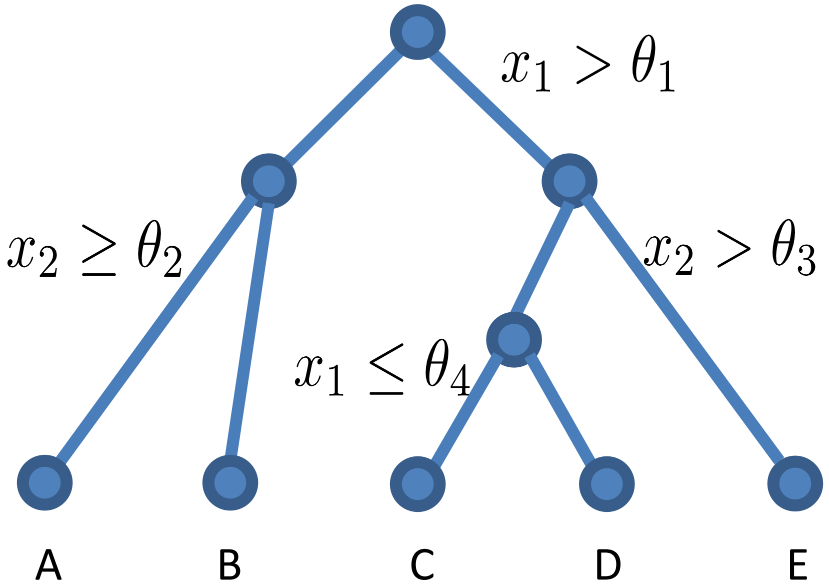


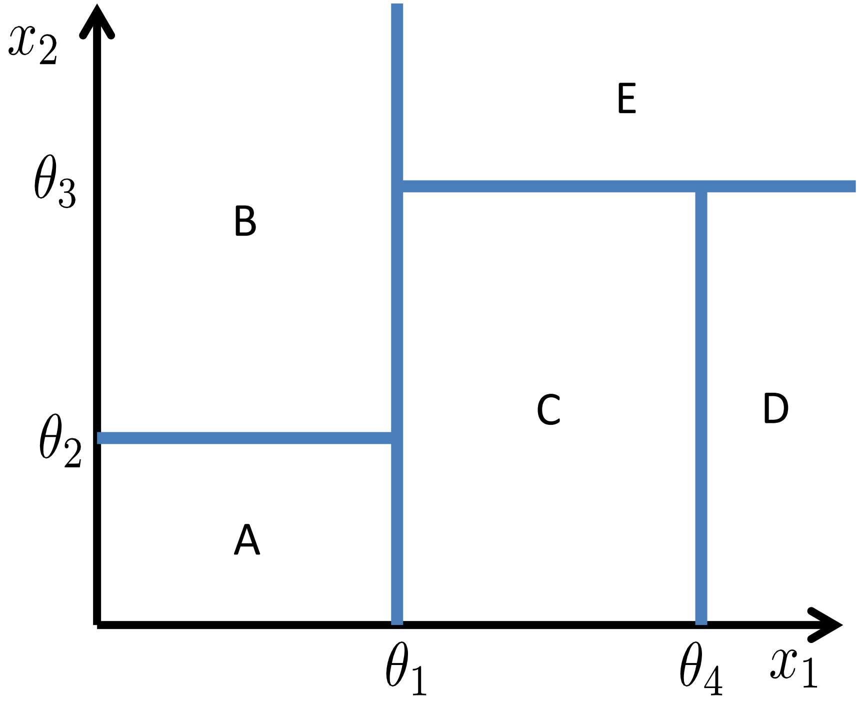
Decision Trees

- Fast training
- Fast prediciton
- Easy to understand
- Easy to interpret

http://en.akinator.com/personnages/jeu

Decision Tree - Idea





Bishop, “Pattern Recognition and Machine Learning”, Springer, 2006

Decision Tree - Idea

- What is a the benefit on using only one feature at a time?

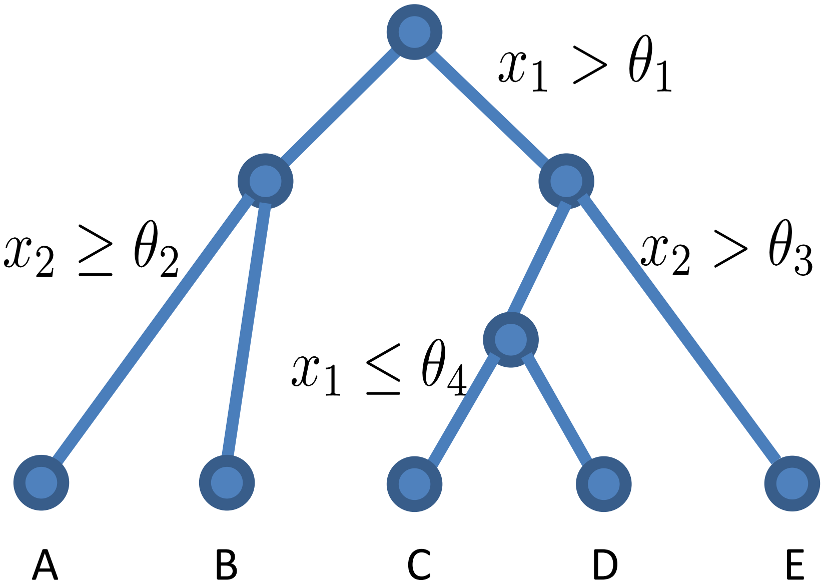
- What is the drawback?

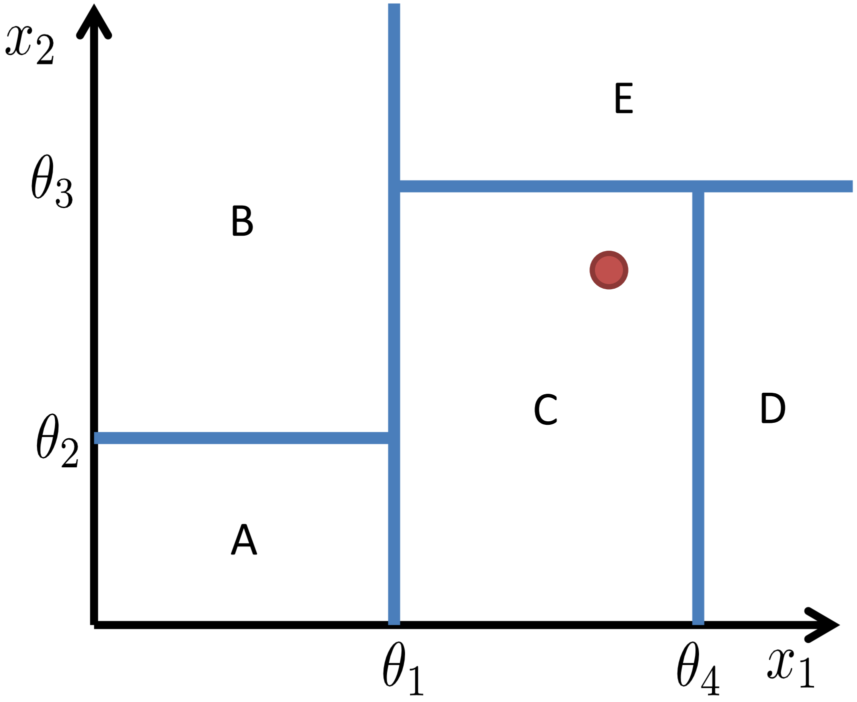
Decision Tree - Idea

- Benefits:
  - Fast in training and prediction
  - Invariant to feature scaling
  - Can handle categorical data

- Drawback:
  - lots of splits for diagonal decision boundary

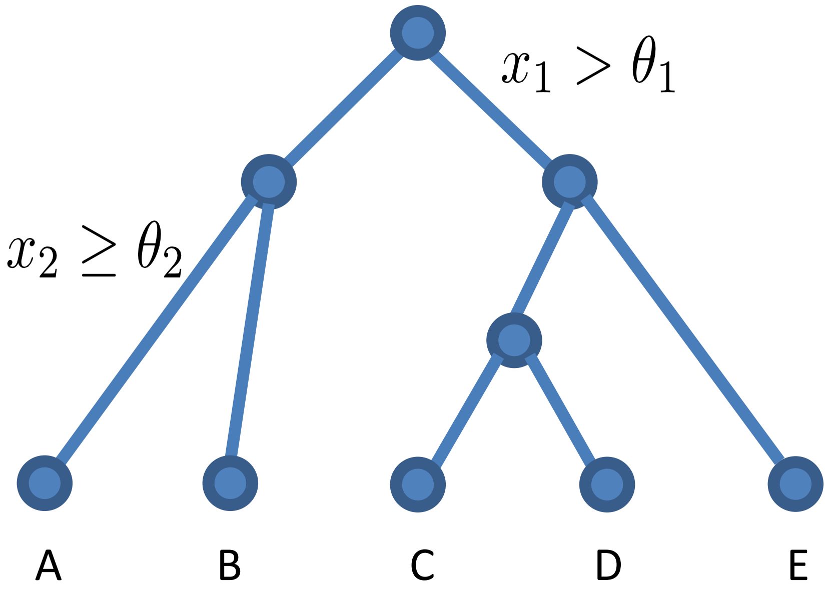
Decision Tree - Prediction



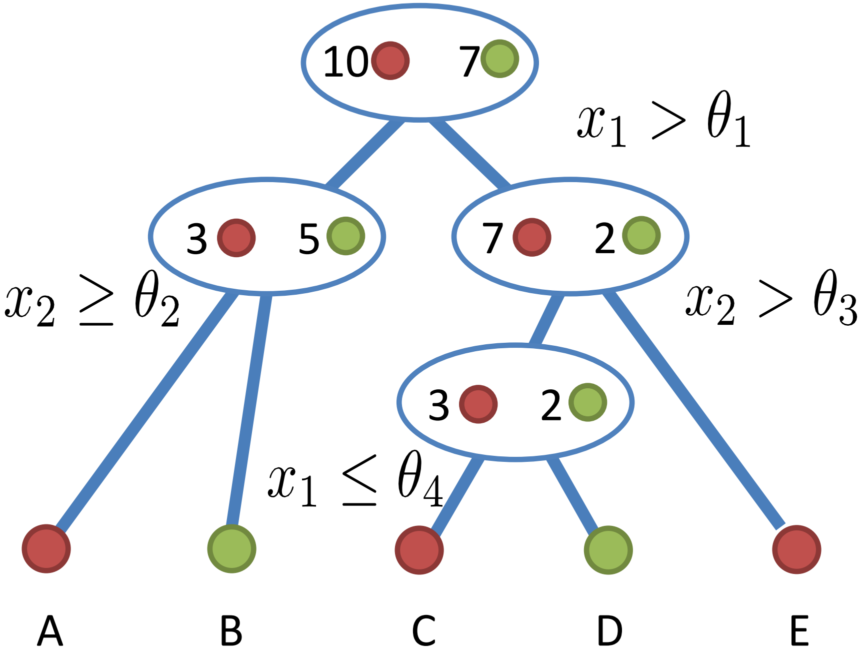


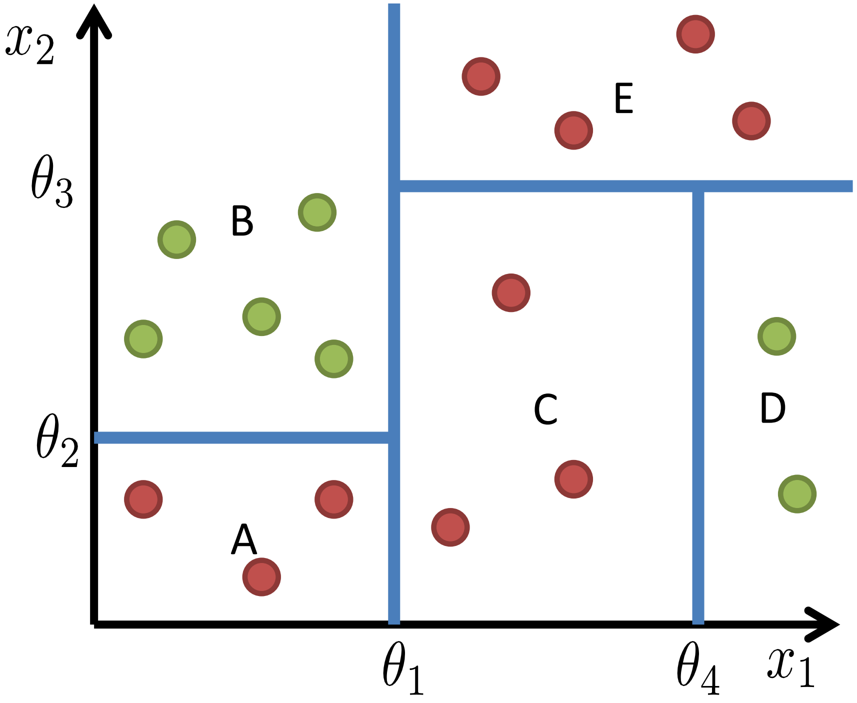
Decision Tree -Training

- Learn the tree structure:
  - – which feature to query
  - – which threshold to choose



Node Purity


10
7
$x_1 > \theta_1$
3
5
7
2
$x_2 \geq \theta_2$
$x_2 > \theta_3$
3
2
$x_1 \leq \theta_4$
A
B
C
D
E


x_2
E
θ_3
B
C
D
θ_2
A
θ_1
θ_4
x_1

Gini Impurity

- Expected error
- if you randomly choose a sample
- and predict the class of the entire node based on it.

Gini Impurity

Example:

4 **red**, 3 **green**, 3 **blue** data points

- Class probabilities:
  - red: 4/10　　green: 3/10　　blue: 3/10
- misclassification:
  - red: $\underbrace{4/10}_{\text{Picking red}} * \underbrace{(3/10 + 3/10)}_{\text{Making an error}}$

Gini Impurity

- misclassification:
  - **red**:

    4/10 * (3/10 + 3/10) = 0.24

  - **green** and **blue**:

    3/10 * (4/10 + 3/10) = 0.21

- gini impurity: **0.24** + **0.21** + **0.21** = 0.66

Gini Impurity

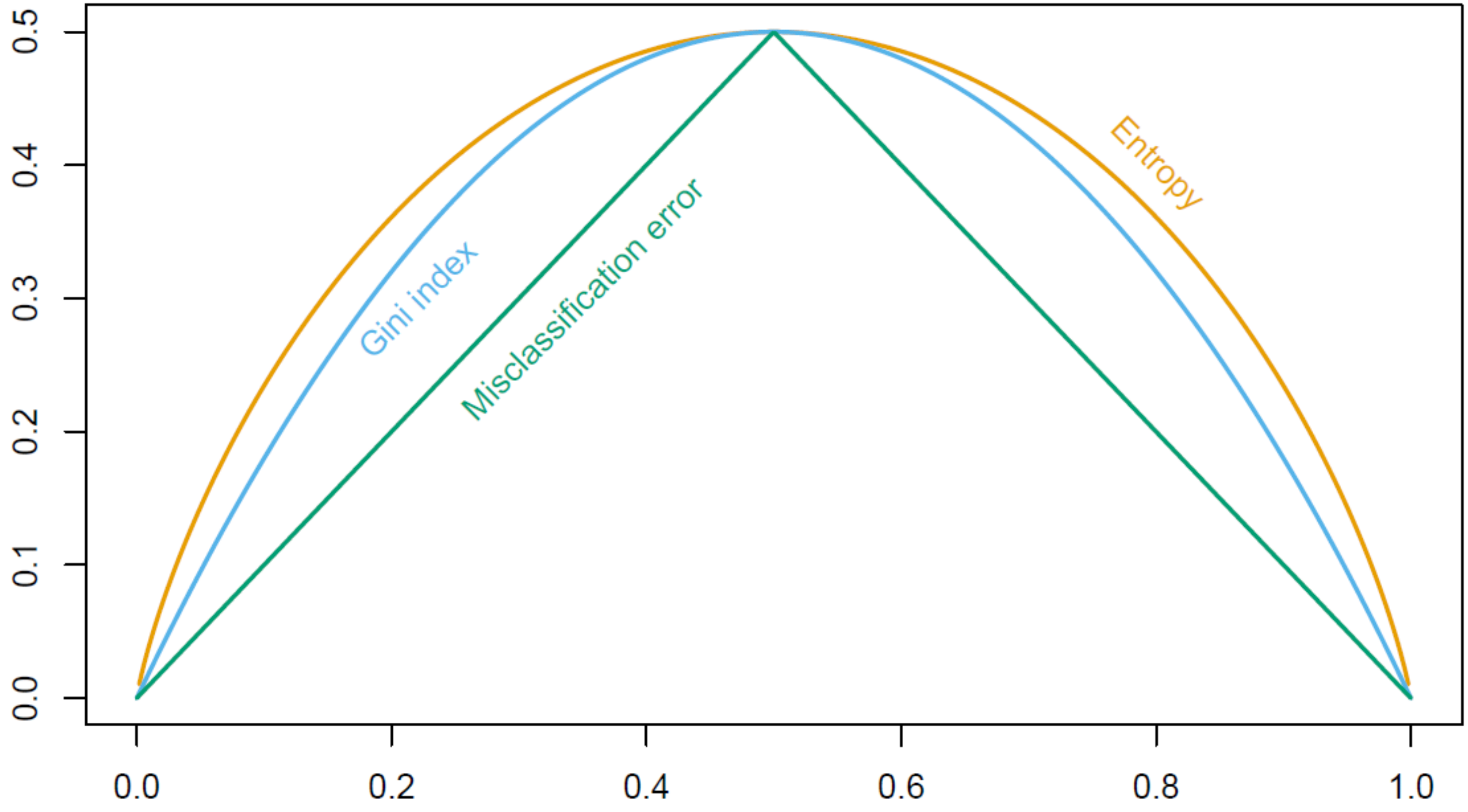
- Number of classes: C
- Number of data points: N
- Number of data points of class i: N_i

$$\boxed{I_G = \sum_{i=1}^{C} \frac{N_i}{N}\left(1 - \frac{N_i}{N}\right)}$$

true class

wrong prediction

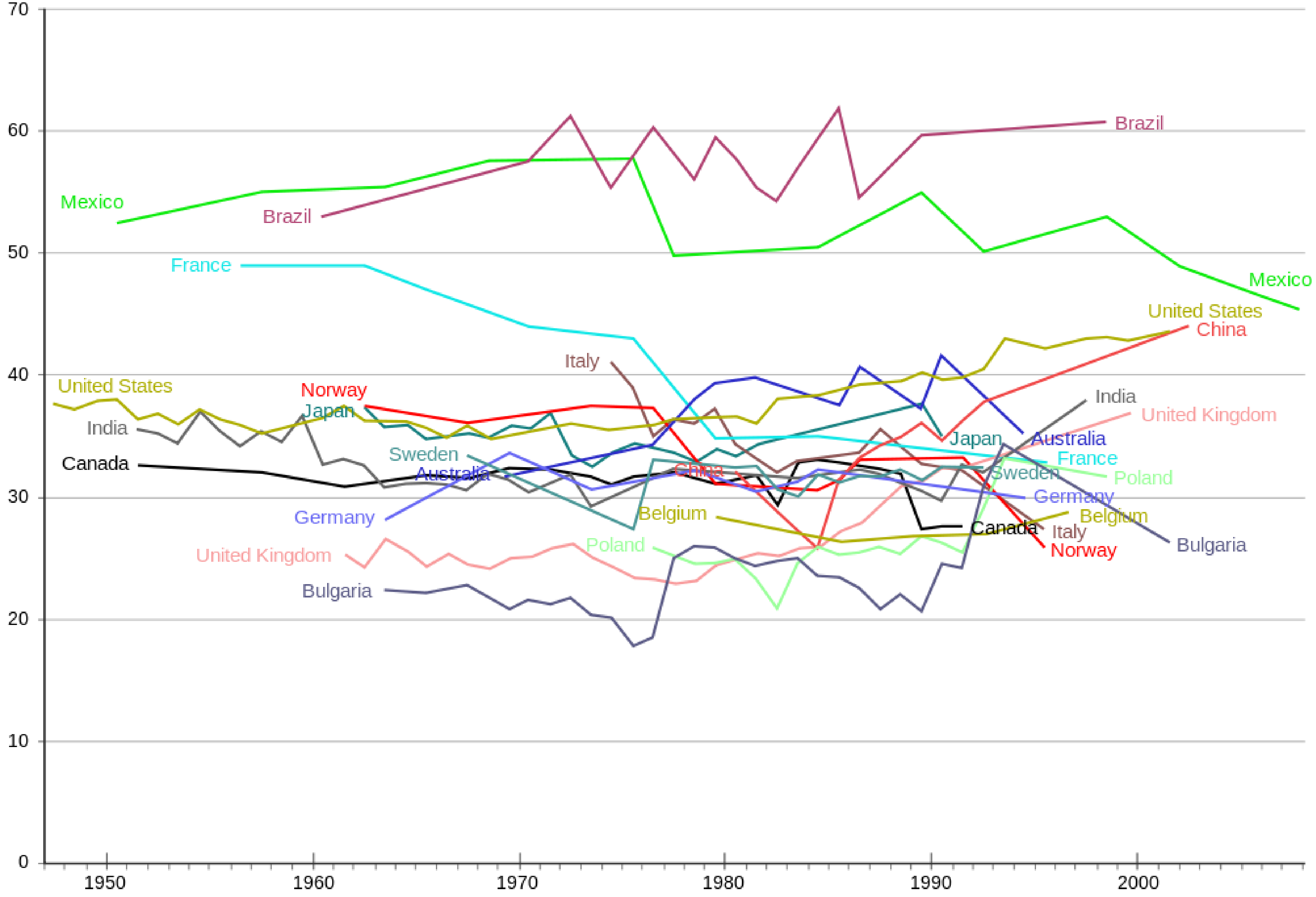
Gini Impurity



Hastie et al.,"The Elements of Statistical Learning: Data Mining, Inference, and Prediction", Springer (2009)

Gini Index - Income Disparity since World War II

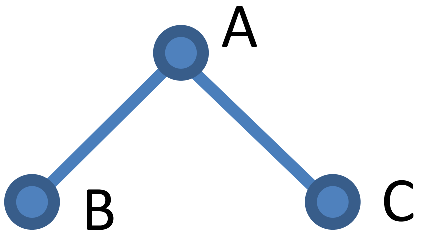
where 0 is perfect equality, and 100 is perfect inequality (i.e., one person has all the income)



http://en.wikipedia.org/wiki/Gini_coefficient

Node Purity Gain

- Compare:
  - Gini impurity of parent node
  - Gini impurity of child nodes



$$\boxed{\Delta I_G = I_G(A) - \frac{N(B)}{N(A)} I_G(B) - \frac{N(C)}{N(A)} I_G(C)}$$

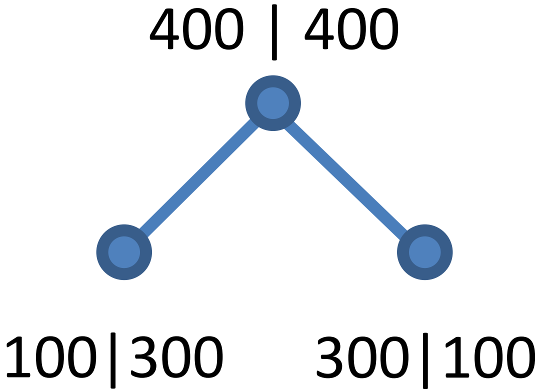
Misclassification

- $\frac{1}{N}\sum_{i}^{N}\mathbf{1}(\hat{\mathrm{y}}_i \neq y_i)$

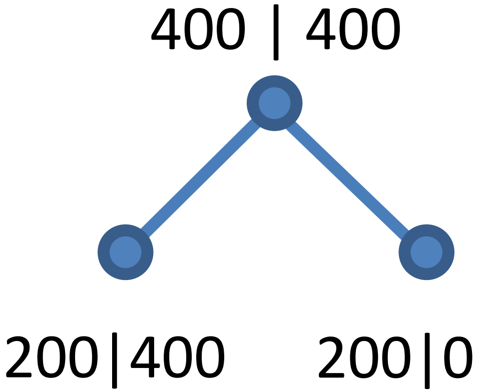
- not differentiable

Comparison Gini vs Misclassification

- Binary problem: 400 samples per class



Misclassification: 0.25
Gini gain: 0.125



Misclassification: 0.25
Gini gain: 0.166

Pseudocode

- Check if already finished
- For each attribute *a*
  - Calculate the gain from splitting on *a*
- Let *a_best* be the attribute with highest gain
- Create a decision *node* that splits on *a_best*
- Repeat on the sub-nodes

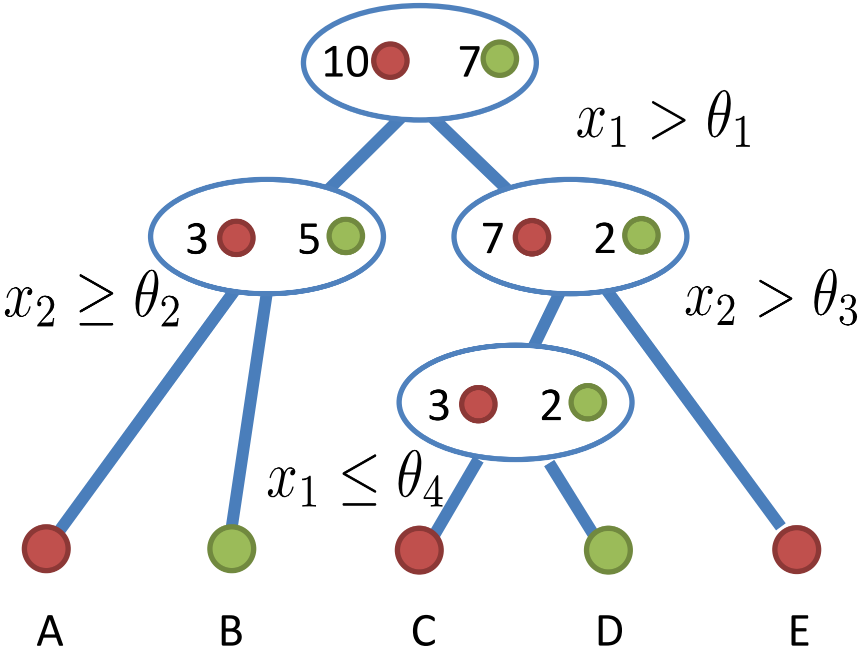
- Does this produce an optimal tree?
- What would an optimal tree be here?

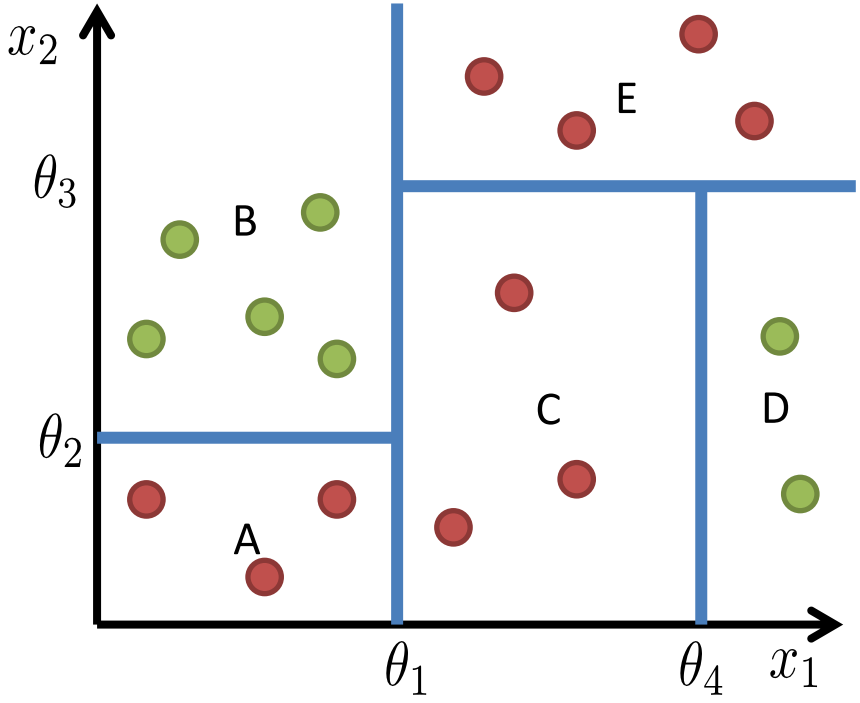
http://en.wikipedia.org/wiki/C4.5_algorithm

When to Stop

- node contains only one class
- node contains less than x data points
- max depth is reached
- node purity is sufficient
- you start to overfit => cross-validation

Tree Pruning





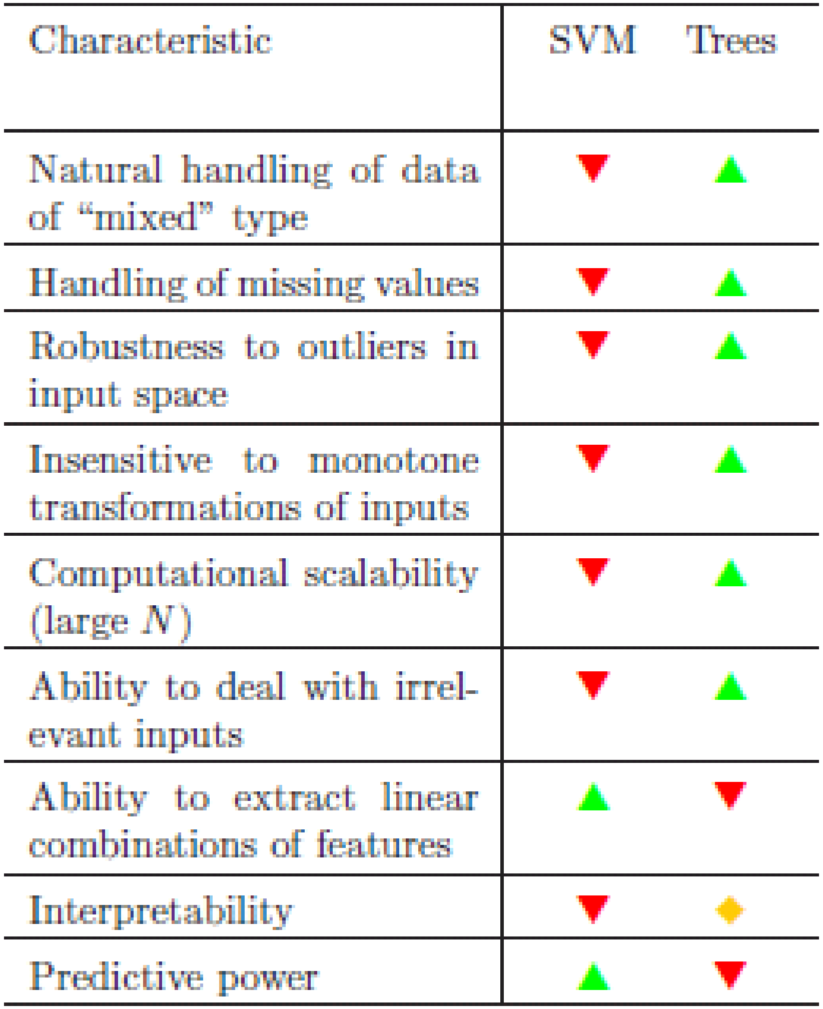
How do you make a prediction for the merged cell?

What is the relation between pruning and k in knn?

Decision Trees - Disadvantages

- Sensitive to small changes in the data
- Overfitting
- Only axis aligned splits

Decision Trees vs SVM



Characteristic	SVM	Trees
Natural handling of data of "mixed" type	▼	▲
Handling of missing values	▼	▲
Robustness to outliers in input space	▼	▲
Insensitive to monotone transformations of inputs	▼	▲
Computational scalability (large N)	▼	▲
Ability to deal with irrelevant inputs	▼	▲
Ability to extract linear combinations of features	▲	▼
Interpretability	▼	◆
Predictive power	▲	▼

Hastie et al.,"The Elements of Statistical Learning: Data Mining, Inference, and Prediction", Springer (2009)

Real Data

DecisionTree in sklearn

- http://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

Wisdom of Crowds

The collective knowledge of a diverse and independent body of people typically exceeds the knowledge of any single individual, and can be harnessed by voting.

James Surowiecki

http://socialmedia4srm.wordpress.com/

https://www.youtube.com/watch?v=ImpV70uLxyw