Decision Trees

Farzaneh Mirzazadeh

University of California, Santa Cruz

Winter' 17

Slide credit

Most of the slides are by Prof Aarti Singh, CMU.

Another nonparametric method: decision trees

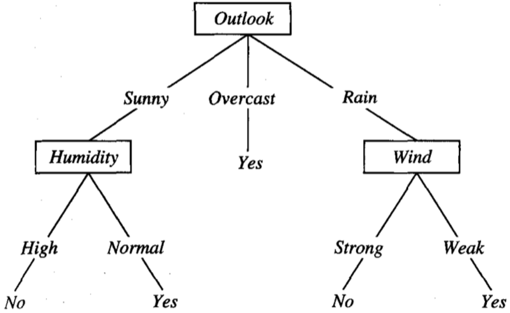
- An old supervised learning method.
- Were popular and commonly used in 80s.
- Simple, interpretable, easy to implement
- An efficient representation for a set of if-then else rules.
- Could be used for both classification and regression.
- Decision trees will overfit. Must use tricks to find Simple Trees.
- Not much used as standalone method now.
- Nowadays more often used as weak learners for the ensemble learning method so that together they form a strong method in boosting or in ensemble method of random forests.

We focus on classification.

Question I

- What does a decision tree represent?
- How should we do prediction with decision trees?

- Each node specifies a test on some feature.
- Each branch descending that node corresponds to one of the possible values for this attribute.



Prediction scheme

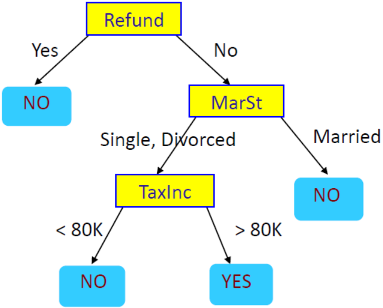
- Given an example, at every internal node, a question about a feature is asked
- Based on the value of that feature for the example, a branch leading to a child of the node is selected and we go down one step to that child.
- Until we get to a leaf.
- The target value at the leaf will be announced as the prediction.

Demo

Decision Tree for Tax Fraud Detection

$\mathcal{F}$ – Decision Trees

$f(X_1, X_2, X_3) \in \mathcal{F}$



Query Data

X_1 Refund	X_2 Marital Status	X_3 Taxable Income	Y Cheat
No	Married	80K	?

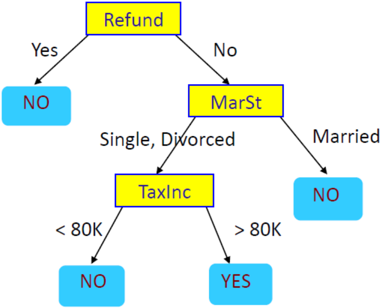
- Each internal node: test one feature X_i
- Each branch from a node: selects one value for X_i
- Each leaf node: predict Y

Demo

Decision Tree for Tax Fraud Detection

$\mathcal{F}$ – Decision Trees

$f(X_1, X_2, X_3) \in \mathcal{F}$



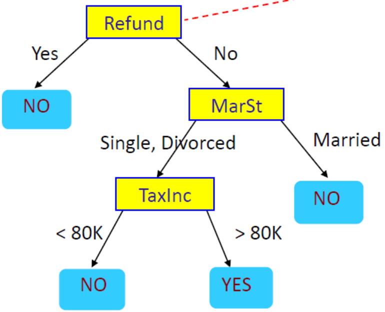
Query Data

X_1 Refund	X_2 Marital Status	X_3 Taxable Income	Y Cheat
No	Married	80K	?

Decision Tree for Tax Fraud Detection

$\mathcal{F}$ – Decision Trees

$f(X_1, X_2, X_3) \in \mathcal{F}$



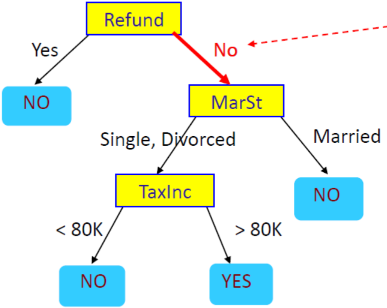
Query Data

X_1 Refund	X_2 Marital Status	X_3 Taxable Income	Y Cheat
No	Married	80K	?

Decision Tree for Tax Fraud Detection

$\mathcal{F}$ – Decision Trees

$f(X_1, X_2, X_3) \in \mathcal{F}$



Query Data

X_1 Refund	X_2 Marital Status	X_3 Taxable Income	Y Cheat
No	Married	80K	?

Prediction algorithm in decision trees

1. Choose one feature to descent at each level Note
    - Condition on earlier (higher) choices.
    - Generally restrict only one input dimension (feature) at a time.
2. Declare the output target value corresponding to the leaf when you get to the bottom (a leaf).

Question II

- How to learn (train) a decision tree from data?

Notes about constructing a decision tree

- The tree is constructed from data by employing a top-down greedy search through the space of possible decision trees
- Beginning with the question: Which attribute should be tested at the root of the tree?
- Each attribute is evaluated using a statistical test to determine how well it alone classifies the training data points
- The best feature is selected and used as the test at the root node of the tree.
- A descendant of the node is then created for each possible value of this feature and training examples are sorted and sent to the appropriate descendant node.
- The entire process is then repeated using the training points associated.

ID3 Tree

- Question: Which feature is the best classifier to correspond to this node of the tree?
- ID3 answer: The feature with largest information gain
- Entropy Given a set of data points S, containing positive and negative classification of examples, the entropy of S is defined by

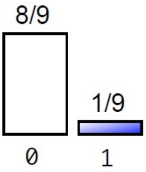
$$H(Y) = -p_{\oplus} \log_2 p_{\oplus} - p_{\ominus} \log_2 p_{\ominus}$$

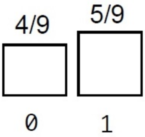
$\oplus$ data points have positive class, $\ominus$ has negative class.

$$H([9+, 5-]) = -9/14 \log_2(9/14)_{-}(5/14) \log_2(5/14) = 0.940$$

To compute Information Gain: Compute the entropy of the target class on this feature ($H(Y)$) - sum of entropy the nodes formed from the attribute values $H(Y|X_i)$

Entropy H:





$$-\frac{8}{9}\log_2\frac{8}{9} - \frac{1}{9}\log_2\frac{1}{9} \approx \frac{1}{2} \qquad -\frac{4}{9}\log_2\frac{4}{9} - \frac{5}{9}\log_2\frac{5}{9} \approx 0.99$$

Information Gain

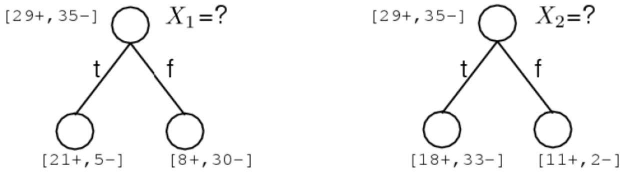
S = A set of examples
F = a possible feature out of a set of features
S_f is examples in which feature F has the value f

Information Gain:

$$InfoGain(S, F) = H(S) - t \sum_{f \in values(F)} \frac{|S_f|}{|S|} H(S_f)$$

Pick the feature that maximizes Information Gain
i.e. reduces entropy most

Which feature is best to split?



Pick the attribute/feature which yields maximum information gain:

$$\arg\max_i I(Y, X_i) = \arg\max_i [H(Y) - H(Y|X_i)]$$

H(Y) – entropy of Y   H(Y|X_i) – conditional entropy of Y

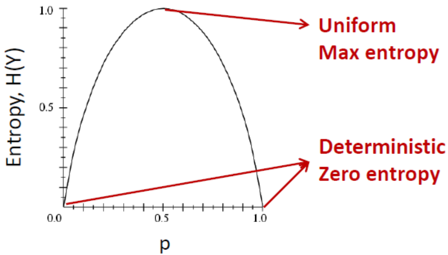
Entropy

- Entropy of a random variable Y

$$H(Y) = -\sum_{y} P(Y = y) \log_2 P(Y = y)$$

More uncertainty, more entropy!

Y ~ Bernoulli(p)



<u>**Information Theory interpretation**</u>: $H(Y)$ is the expected number of bits needed to encode a randomly drawn value of Y (under most efficient code)

Information Gain

- Advantage of attribute = decrease in uncertainty
  - Entropy of Y before split

$$H(Y) = -\sum_y P(Y=y)\log_2 P(Y=y)$$

  - Entropy of Y after splitting based on X_i
    - Weight by probability of following each branch

$$\begin{aligned} H(Y \mid X_i) &= -\sum_x P(X_i = x) H(Y \mid X_i = x) \\ &= -\sum_x P(X_i = x) \sum_y P(Y = y \mid X_i = x) \log_2 P(Y = y \mid X_i = x) \end{aligned}$$

- Information gain is difference

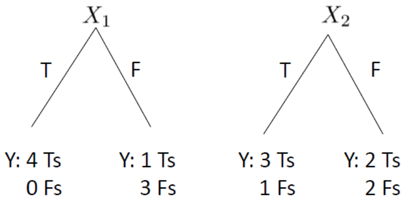
$$I(Y, X_i) = H(Y) - H(Y \mid X_i)$$

Max Information gain = min conditional entropy

Information Gain

$$H(Y \mid X_i) \;=\; -\sum_x P(X_i = x) \sum_y P(Y = y \mid X_i = x) \log_2 P(Y = y \mid X_i = x)$$

X_1	X_2	Y
T	T	T
T	F	T
T	T	T
T	F	T
F	T	T
F	F	F
F	T	F
F	F	F



$$\widehat{H}(Y|X_1) = -\cancel{\frac{1}{2}[1\log_2 1 + 0\log_2 0]} - \frac{1}{2}[\frac{1}{4}\log_2\frac{1}{4} + \frac{3}{4}\log_2\frac{3}{4}]$$

$$\widehat{H}(Y|X_2) = -\frac{1}{2}[\frac{3}{4}\log_2\frac{3}{4} + \frac{1}{4}\log_2\frac{1}{4}] \underbrace{- \frac{1}{2}[\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}]}_{>0}$$

$$\widehat{H}(Y|X_1) < \widehat{H}(Y|X_2)$$

Which feature is best to split?

Pick the attribute/feature which yields maximum information gain:

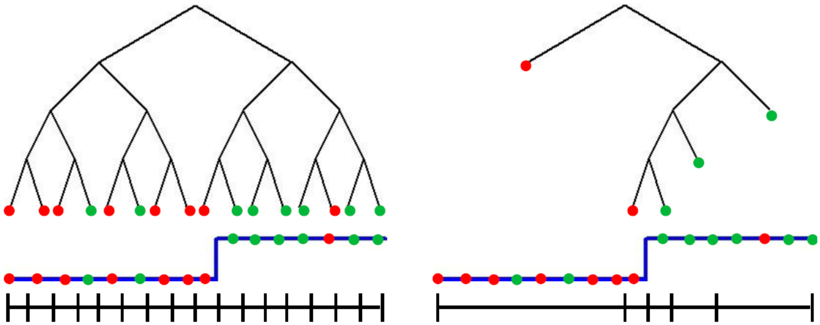
$$\arg\max_i I(Y, X_i) = \arg\max_i [H(Y) - H(Y|X_i)]$$

H(Y) – entropy of Y   H(Y|X_i) – conditional entropy of Y

Feature which yields maximum reduction in entropy provides maximum information about Y

Decision Trees - Overfitting

One training example per leaf – overfits, need compact/pruned decision tree



t

Sample exam question

Consider the training dataset given below. In the dataset, X_1, X_2, and X_3 are the attributes and Y is the class variable.

Example#	X_1	X_2	X_3	Y
E1	0	0	0	+
E2	0	0	1	-
E3	0	1	0	-
E4	0	1	1	+
E5	1	0	0	-

(a) Which attribute has the highest information gain? Justify your answer?

(b) Draw the (full) decision tree for this dataset using the information gain criteria.