CSCI567 Machine Learning (Spring 2018)

Michael Shindler

Lecture on February 28, 2018

Outline

Outline

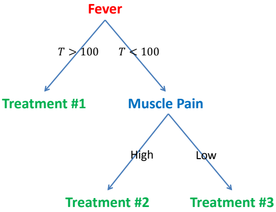
Administrative stuff

- Viewing session?

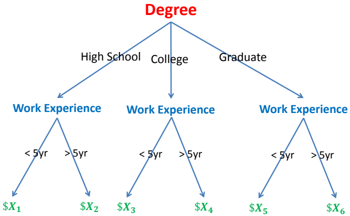
Outline

Many decisions are tree structures

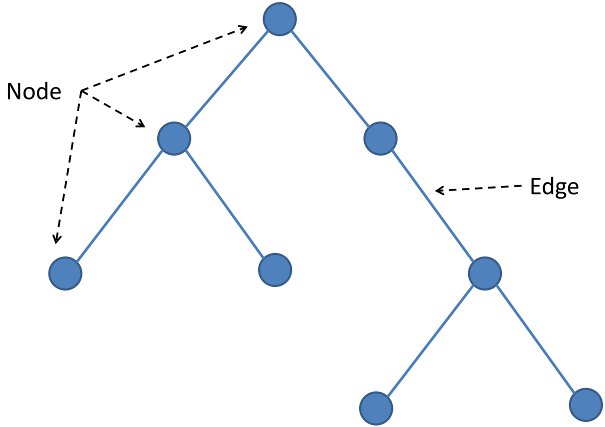
Medical treatment



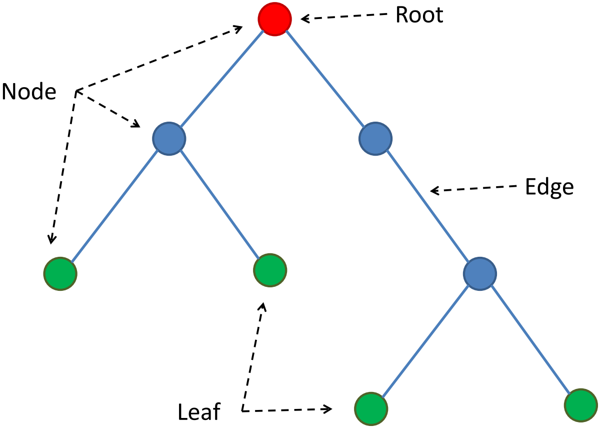
Salary in a company



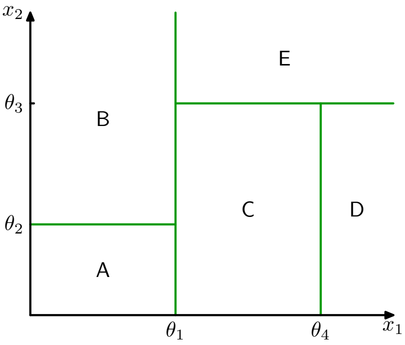
What is a Tree?

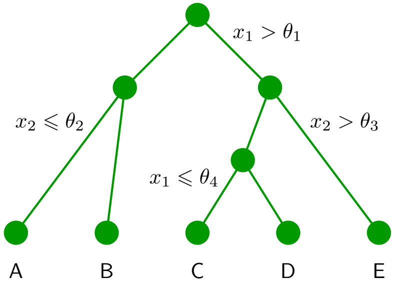


Special Names for Nodes in a Tree



A tree partitions the feature space

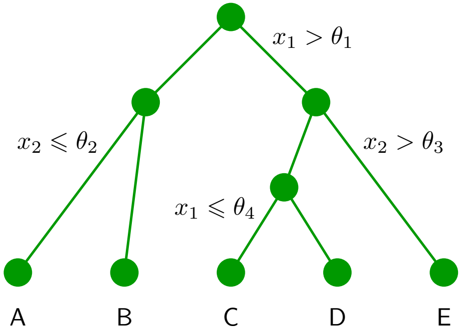




Learning a tree model

Three things to learn:

1. The structure of the tree.
2. The threshold values (θ_i).
3. The values for the leafs ($A, B, \ldots$).



A tree model for deciding where to eat

Choosing a restaurant

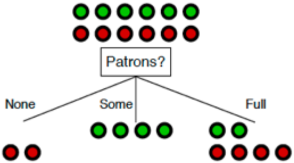
(Example from Russell & Norvig, AIMA)

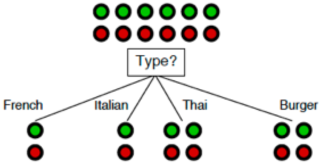
Example	Attributes										Target
	Alt	*Bar*	*Fri*	*Hun*	*Pat*	*Price*	*Rain*	*Res*	*Type*	*Est*	*WillWait*
X_1	T	F	F	T	Some	$$$	F	T	French	0–10	T
X_2	T	F	F	T	Full	$	F	F	Thai	30–60	F
X_3	F	T	F	F	Some	$	F	F	Burger	0–10	T
X_4	T	F	T	T	Full	$	F	F	Thai	10–30	T
X_5	T	F	T	F	Full	$$$	F	T	French	>60	F
X_6	F	T	F	T	Some	$$	T	T	Italian	0–10	T
X_7	F	T	F	F	None	$	T	F	Burger	0–10	F
X_8	F	F	F	T	Some	$$	T	T	Thai	0–10	T
X_9	F	T	T	F	Full	$	T	F	Burger	>60	F
X_{10}	T	T	T	T	Full	$$$	F	T	Italian	10–30	F
X_{11}	F	F	F	F	None	$	F	F	Thai	0–10	F
X_{12}	T	T	T	T	Full	$	F	F	Burger	30–60	T

Classification of examples is positive (T) or negative (F)

First decision: at the root of the tree

Which attribute to split?





Patrons? is a better choice—gives **information** about the classification

Idea: use information gain to choose which attribute to split

How to measure information gain?

Idea:

Gaining information reduces uncertainty

the base can be 2 , though it is not essential (if the base is 2, the unit of the entropy is called "bit")

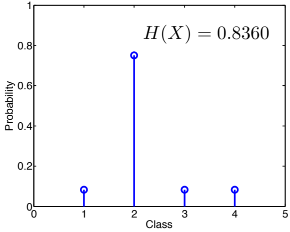
Use to entropy to measure uncertainty

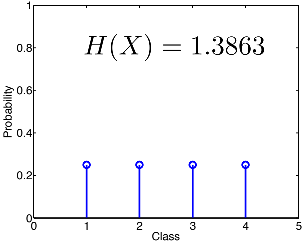
If a random variable X has K different values, a_1, a_2, ...a_K, it is entropy is given by

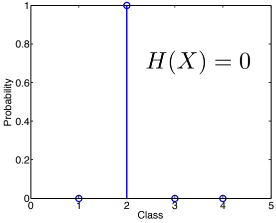
$$H[X] = -\sum_{k=1}^{K} P(X = a_k) \log P(X = a_k)$$

Examples of computing entropy

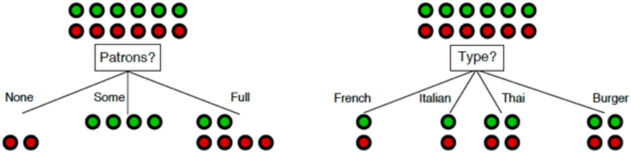
Entropy







Which attribute to split?



Patrons? is a better choice—gives **information** about the classification

Patron vs. Type?

By choosing Patron, we end up with a partition (3 branches) with smaller entropy, ie, smaller uncertainty (0.45 bit)

By choosing Type, we end up with uncertainty of 1 bit.

Thus, we choose Patron over Type.

Uncertainty if we go with "Patron"

For "None" branch

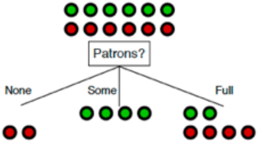
$$-\left(\frac{0}{0+2}\log\frac{0}{0+2}+\frac{2}{0+2}\log\frac{2}{0+2}\right)=0$$

For "Some" branch

$$-\left(\frac{4}{4+0}\log\frac{4}{4+0}+\frac{4}{4+0}\log\frac{4}{4+0}\right)=0$$

For "Full" branch

$$-\left(\frac{2}{2+4}\log\frac{2}{2+4}+\frac{4}{2+4}\log\frac{4}{2+4}\right)\approx 0.9$$



For choosing "Patrons"

weighted average of each branch: this quantity is called conditional entropy

$$\frac{2}{12}*0+\frac{4}{12}*0+\frac{6}{12}*0.9=0.45$$

Conditional entropy

Definition. Given two random variables X and Y

$$H[Y|X] = \sum_k P(X = a_k) H[Y|X = a_k]$$

In our example

X: the attribute to be split

Y: Wait or not

When H[Y] is fixed, we need only to compare conditional entropy

Relation to information gain

$$\mathrm{GAIN} = H[Y] - H[Y|X]$$

Conditional entropy for Type

For "French" branch

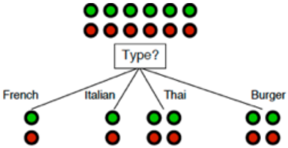
$$-\left(\frac{1}{1+1}\log\frac{1}{1+1}+\frac{1}{1+1}\log\frac{1}{1+1}\right)=1$$

For "Italian" branch

$$-\left(\frac{1}{1+1}\log\frac{1}{1+1}+\frac{1}{1+1}\log\frac{1}{1+1}\right)=1$$

For "Thai" and "Burger" branches

$$-\left(\frac{2}{2+2}\log\frac{2}{2+2}+\frac{2}{2+2}\log\frac{2}{2+2}\right)=1$$

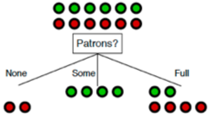


For choosing "Type"

weighted average of each branch:

$$\frac{2}{12}*1+\frac{2}{12}*1+\frac{4}{12}*1+\frac{4}{12}*1=1$$

next split?

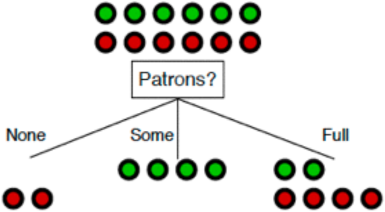


We will look only at the 6 instances with Patrons == Full

Example	Attributes										Target
	Alt	*Bar*	*Fri*	*Hun*	*Pat*	*Price*	*Rain*	*Res*	*Type*	*Est*	*WillWait*
X_1	T	F	F	T	Some	\$\$\$	F	T	French	0–10	T
X_2	T	F	F	T	Full	\$	F	F	Thai	30–60	F
X_3	F	T	F	F	Some	\$	F	F	Burger	0–10	T
X_4	T	F	T	T	Full	\$	F	F	Thai	10–30	T
X_5	T	F	T	F	Full	\$\$\$	F	T	French	>60	F
X_6	F	T	F	T	Some	\$\$	T	T	Italian	0–10	T
X_7	F	T	F	F	None	\$	T	F	Burger	0–10	F
X_8	F	F	F	T	Some	\$\$	T	T	Thai	0–10	T
X_9	F	T	T	F	Full	\$	T	F	Burger	>60	F
X_{10}	T	T	T	T	Full	\$\$\$	F	T	Italian	10–30	F
X_{11}	F	F	F	F	None	\$	F	F	Thai	0–10	F
X_{12}	T	T	T	T	Full	\$	F	F	Burger	30–60	T

Classification of examples is positive (T) or negative (F)

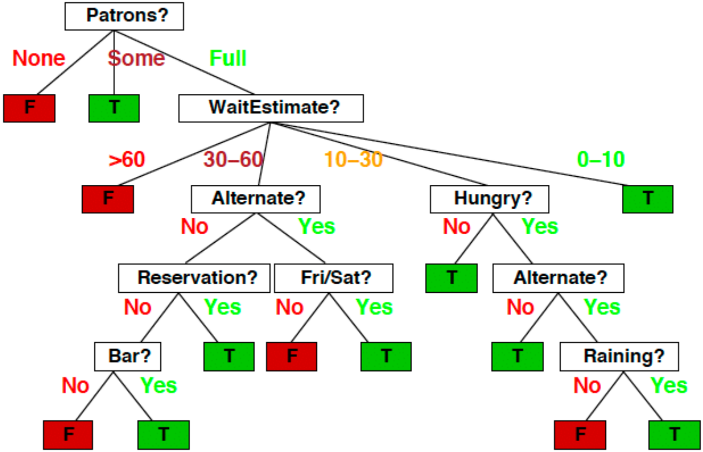
Do we split on "Non" or "Some"?



No, we do not

The decision is deterministic, as seen from the training data

Greedily we build the tree and get this



How deep should we continue to split?

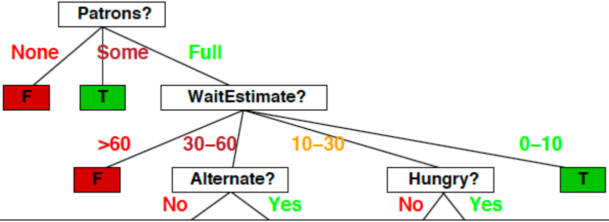
We should be very careful about this

Eventually, we can get all training examples right. But is that what we want?

The maximum depth of the tree is a hyperparameter and should not be tuned by training data — this is to prevent overfitting (we will discuss later)

Control the size of the tree

We would prune to have a smaller one



If we stop here, not all training sample would be classified correctly.

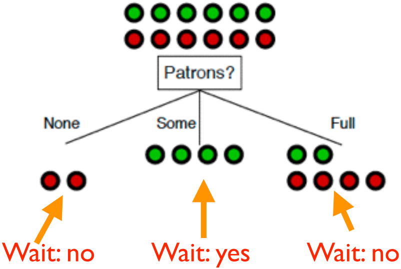
More importantly, how do we classify a new instance?

We label the leaves of this smaller tree with the majority of training samples' labels

Example

Example

We stop after the root (first node)



Splitting and Stopping Criteria

For every leaf m, define the node impurity $Q(m)$ as:

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

The **Misclassification Error** is less sensitive to changes in class probability:

$\Rightarrow$ Use Gini Index or Cross-entropy for growing T_0,

$\Rightarrow$ Use Misclassification Error for pruning T_0 and finding T.

Summary of learning trees

Other ideas in learning trees

- There are other ways of splitting attributes, such as Gini index.
- There are other fast ways of learning tree models.
- There are approaches of learning an ensemble of tree models (more on this later)

Advantages of using trees

- The models are transparent: easily interpretable by human (as long as the tree is not too big)
- It is parametric thus compact: unlike NNC, we do not have to carry our training instances around

Outline

Random Forests

- Idea: build a large collection of de-correlated trees
- Use them to vote on a classification
- This is similar in effect to *boosting* (next lecture)
- but these are simpler to train and tune.

Rule for growing trees

The Rule: before each split, select $m \leq p$ of the input variables at random as candidates for splitting.
Why?

- Trees (on their own) are noisy
- We'd like to reduce variance
- Average of B i.i.d random variables
  each with variance σ^2
  Variance σ^2/B
- If identically distributed, but not independent,
  with positive pairwise correlation ρ, variance of average is:

$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2$$

Random Forests are popular

- Software freely available:
  http://math.usu.edu/ adele/forests/
- Many claims about the success:
  - Most accurate
  - Most interpretable

Details of Random Forests

The Rule: before each split, select $m \leq p$ of the input variables at random as candidates for splitting.

- Build many trees
- When classifying, give each tree a vote.
- Use majority vote for classification
- Use average for regression problems.
- In general, $\lfloor \sqrt{p} \rfloor$ suggested value m

Out of bag samples

For each observation z_i, construct its random forest predictor by averaging only those trees corresponding to bootstrap samples in which z_i did not appear.

- OOB error estimate is almost identical to N-fold cross validation

Out of bag samples

For each observation z_i, construct its random forest predictor by averaging only those trees corresponding to bootstrap samples in which z_i did not appear.

- OOB error estimate is almost identical to N-fold cross validation
- Random forests can be fit in one sequence
- Cross-validation is performed along the way
- Training can stop when OOB error stabilizes.

Variable Importance

- How strong is predictive power of each variable?
- When bth tree is grown:
  - Pass OOB samples down the tree
  - Record prediction accuracy.
  - Then values for jth are randomly permuted
    Accuracy is again computed.
  - Decrease in accuracy as a result is averaged across trees
    This is a measure of the importance of j in the random forest.