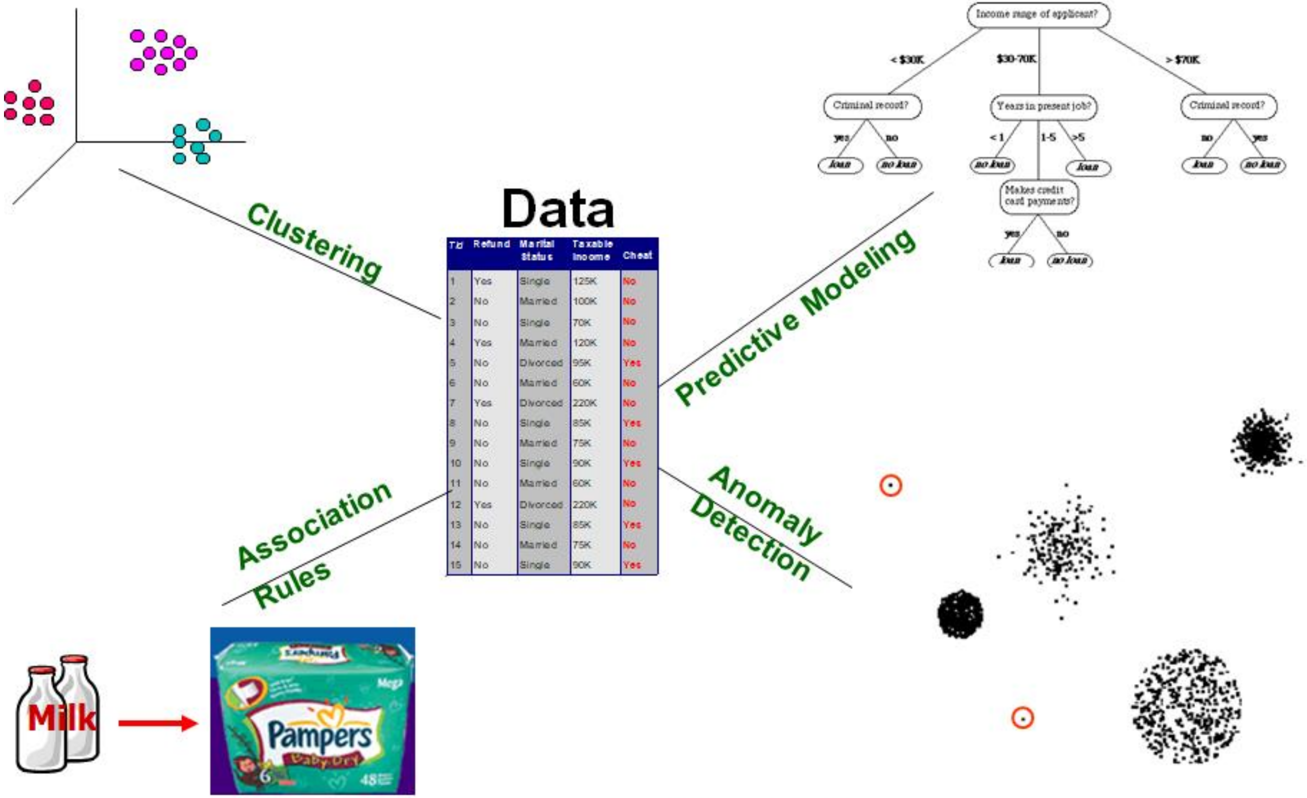
Foundations of Data Science & Analytics: Decision Trees

Ezgi Siir Kibris

Introduction to Data Mining, 2nd Edition
by
Tan, Steinbach, Karpatne, Kumar

Tasks



Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
11	No	Married	60K	No
12	Yes	Divorced	220K	No
13	No	Single	85K	Yes
14	No	Married	75K	No
15	No	Single	90K	Yes

Predictive Modeling

	Output
Classification:	Classes / Categories
Regression:	Continuous Values

Classification Techniques

- **Base Classifiers**
  - **Decision Tree based Methods**
  - Rule-based Methods
  - Instance-based Methods (Nearest-neighbor)
  - Naïve Bayes
  - Support Vector Machines
  - Neural Networks and Deep Learning

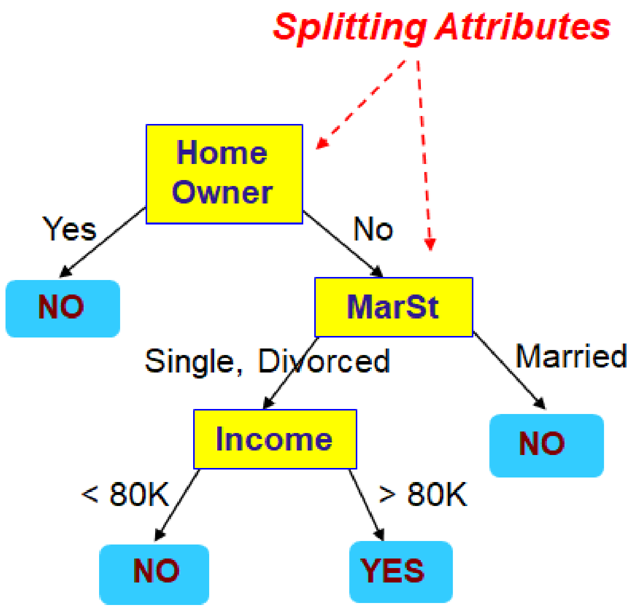
- **Ensemble Classifiers**
  - Boosting, Bagging, Random Forests

Decision Tree

categorical categorical continuous class

ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

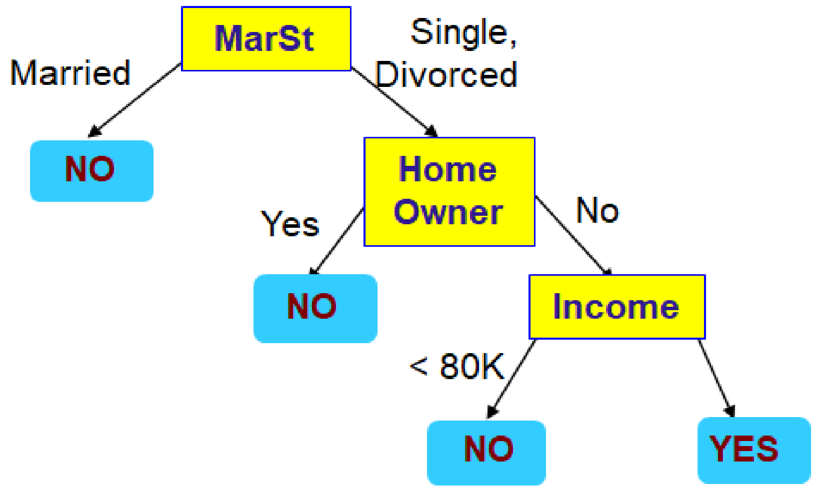
Training Data



Model: Decision Tree

Decision Tree

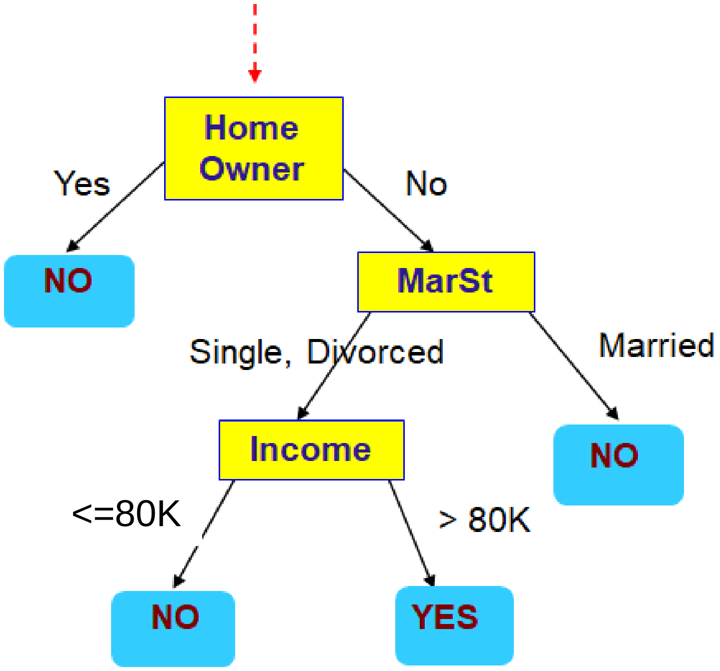
ID	Home Owner (categorical)	Marital Status (categorical)	Annual Income (continuous)	Defaulted Borrower (class)
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



There could be more than one tree that fits the same data!

Predict

Start from the root of tree.



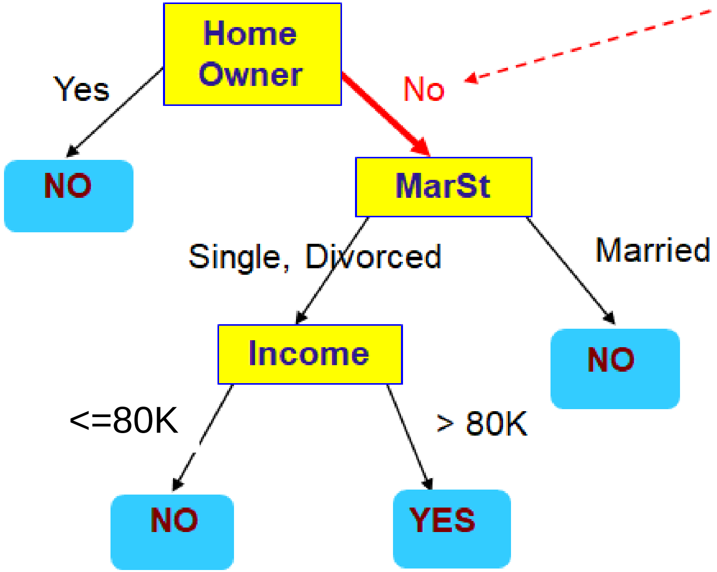
Test Data

Home Owner	Marital Status	Annual Income	Defaulted Borrower
No	Married	80K	?

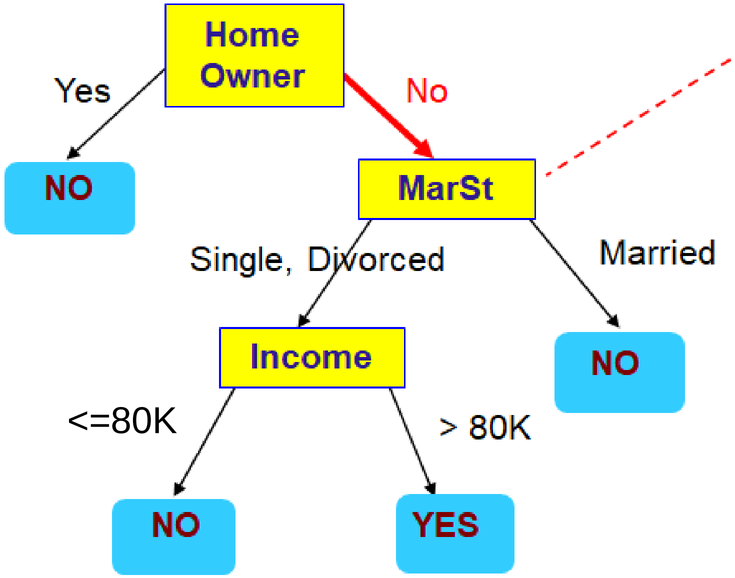
Predict

Test Data

Home Owner	Marital Status	Annual Income	Defaulted Borrower
No	Married	80K	?



Predict



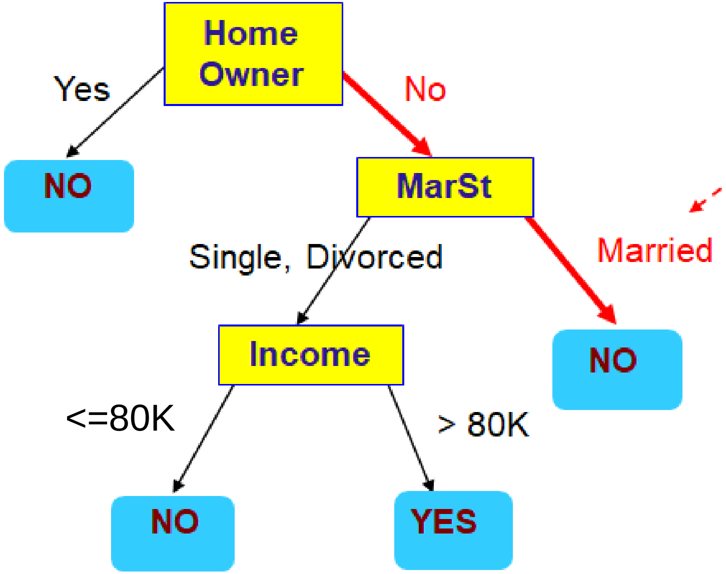
Test Data

Home Owner	Marital Status	Annual Income	Defaulted Borrower
No	Married	80K	?

Predict

Test Data

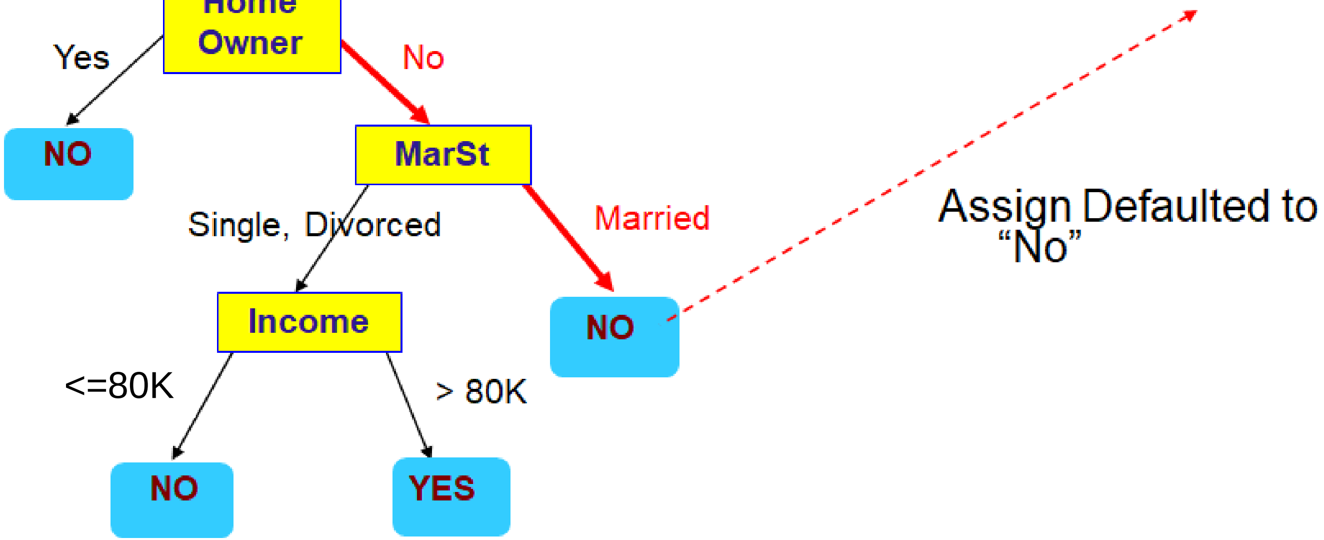
Home Owner	Marital Status	Annual Income	Defaulted Borrower
No	Married	80K	?



Predict

Test Data

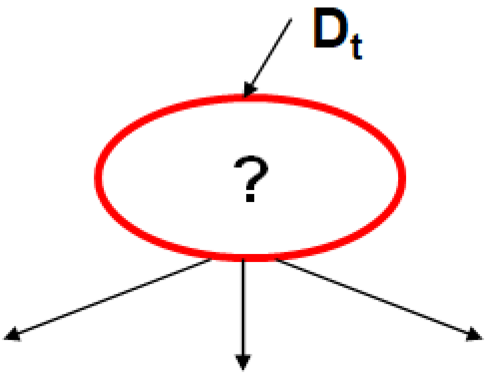
Home Owner	Marital Status	Annual Income	Defaulted Borrower
No	Married	80K	?



How to build a decision tree?

- **Tree induction algorithms:**
  - Hunt's Algorithm (one of the earliest)
  - CART
  - ID3, C4.5
  - SLIQ,SPRINT

Hunt's Algorithm



Let D_t be the set of training records that reach a node t.

General Procedure:

- If D_t only contains records that belong to the same class y_t, then t is a leaf node labeled as y_t
- If D_t contains records that belong to more than one class, use a feature to split the data into smaller subsets. Recursively apply the procedure to each subset.

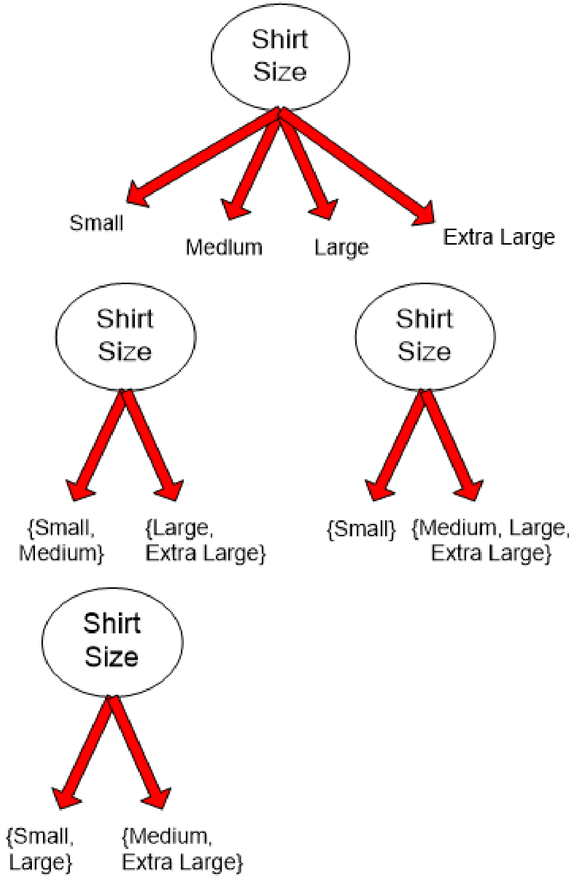
How to split a node?

Multi-way split:

- Use as many partitions as distinct values

Binary split:

- Divides values into two subsets
- Preserve order property among feature values



How to find the Best split?

Greedy approach:

- Nodes with purer class distribution are preferred

Need a measure of node impurity:

C0: 5	C0: 9
C1: 5	C1: 1
High degree of impurity	**Low degree of impurity**

Measure of Node Impurity

- **Gini Index:**

$$Gini\ Index = 1 - \sum_{i=0}^{c-1} p_i(t)^2$$

Where $\boldsymbol{p_i(t)}$ is the frequency of class $\boldsymbol{i}$ at node **t**, and $\boldsymbol{c}$ is the total number of classes

- **Entropy:**

$$Entropy = -\sum_{i=0}^{c-1} p_i(t) log_2 p_i(t)$$

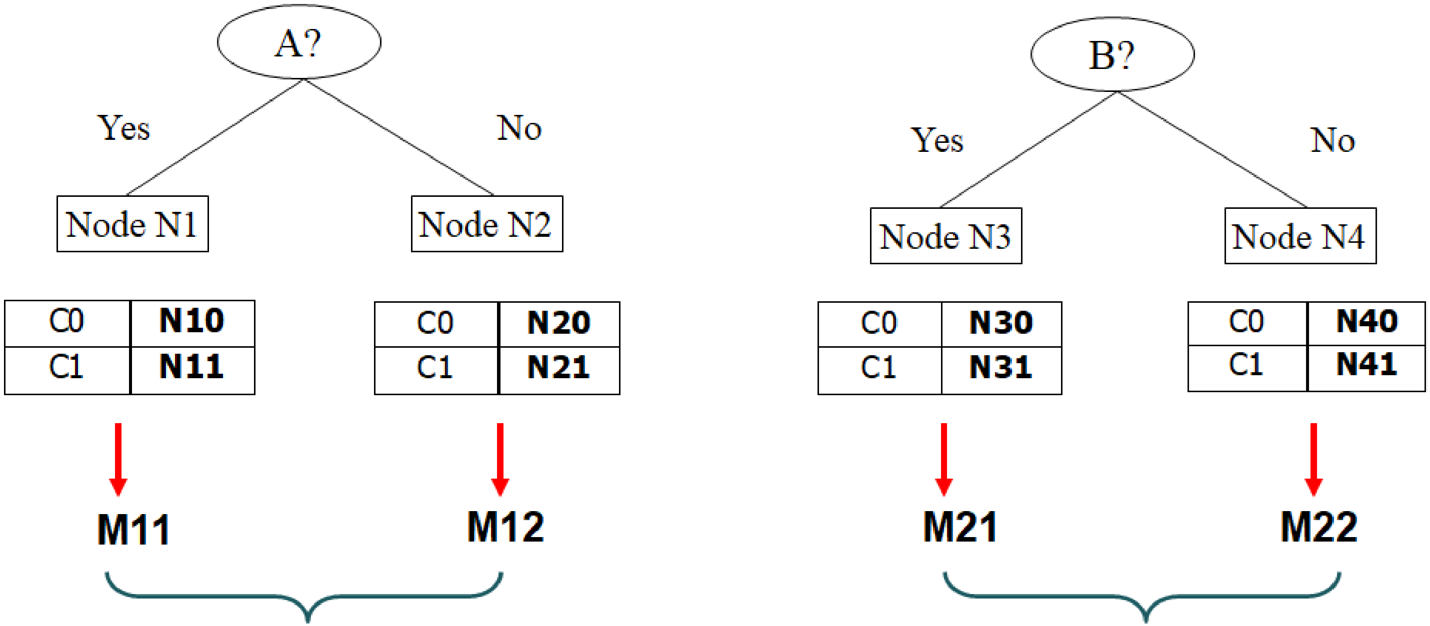
- **Misclassification Error:**

$$Classification\ error = 1 - \max[p_i(t)]$$

Before Splitting:

C0	**N00**
C1	**N01**

→ **P**



M1 = M11*(N1/N0) + M12*(N2/N0) **M2**

Gain = P – M1 vs P – M2

ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

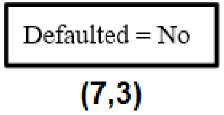
Defaulted = No

(7,3)

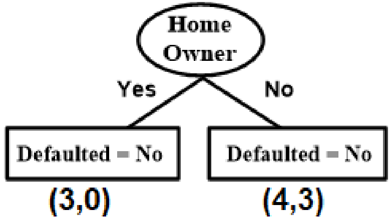
(a)

Before split:

$$P = Gini = 1 - ((7/10)^2 + (3/10)^2) = 0.42$$



(a)



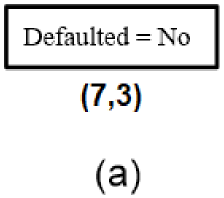
(b)

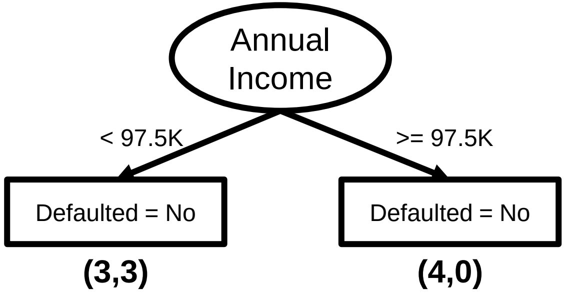
ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

If split on Home Owner:

$$M(HO) = (3/10) \times Gini(HO_{Yes})$$
$$+ (7/10) \times Gini(HO_{No})$$
$$= 0.3 \times (1 - ((3/3)^2 + (0/3)^2)) + 0.7$$
$$\times (1 - ((4/7)^2 + (3/7)^2)) = 0.34$$

$$Gini\ Index = 1 - \sum_{i=0}^{c-1} p_i(t)^2$$

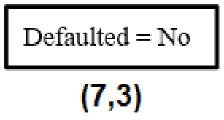




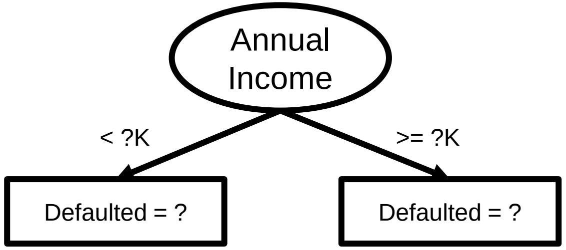
ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

If split on Annual Income:

$$M(AI) = \min Gini(AI(X)) = Gini(AI(97.5K))$$
$$= 0.6 \times (1 - ((3/6)^2 + (3/6)^2)) + 0 = 0.30$$



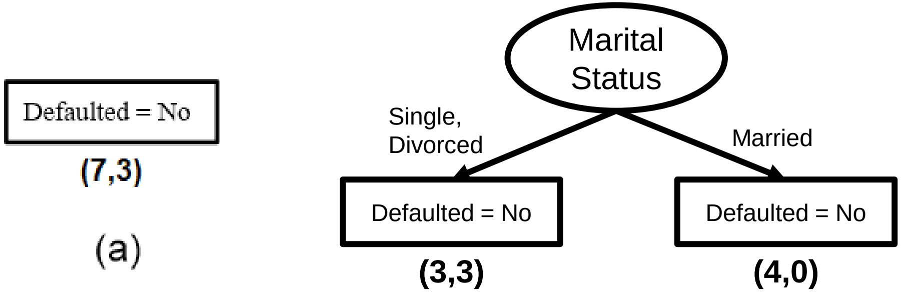
(a)



ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

If split on Annual Income:

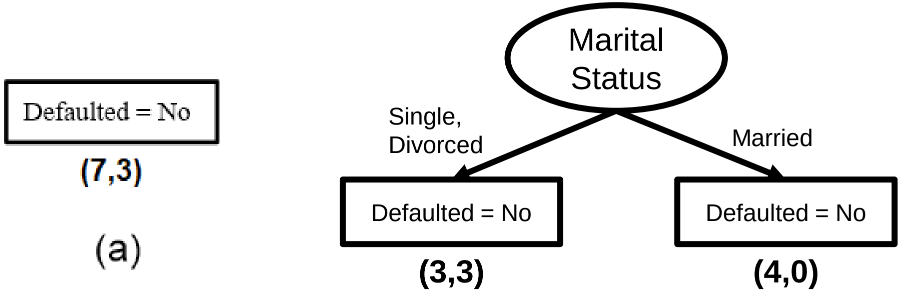
No	No	No	Yes	Yes	Yes	No	No	No	No
<60K	<70K	<75K	<85K	<90K	<95K	<100K	<120K	<125K	<220K
0.42	0.40	0.375	0.34	0.417	0.40	0.30	0.34	0.375	0.40



ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

If split on Marital Status:

$$M(MS) = \min Gini(MS(X)) = Gini(MS(Single \cup Divorced, Married))$$
$$= 0.6 \times (1 - ((3/6)^2 + (3/6)^2)) + 0 = 0.30$$



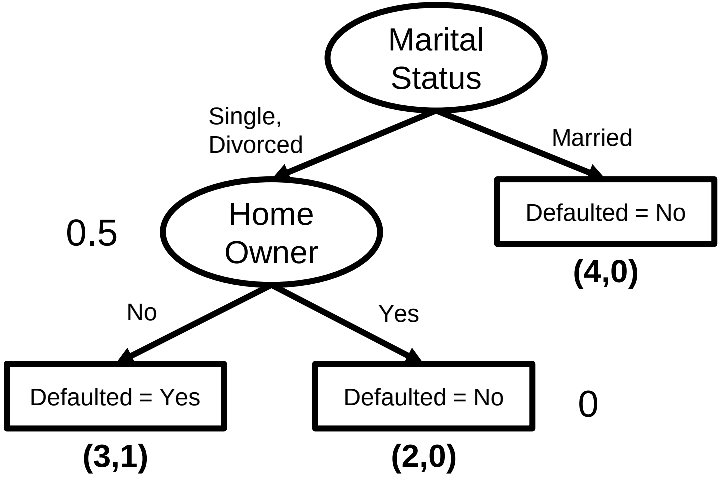
ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Best split:

$$Gain(HO) = P - M(HO) = 0.42 - 0.34 = 0.08$$

$$\boxed{\begin{aligned} Gain(AI) &= P - M(AI) = 0.42 - 0.30 = 0.12 \\ Gain(MS) &= P - M(MS) = 0.42 - 0.30 = 0.12 \end{aligned}}$$

How about split on **Marital Status**?



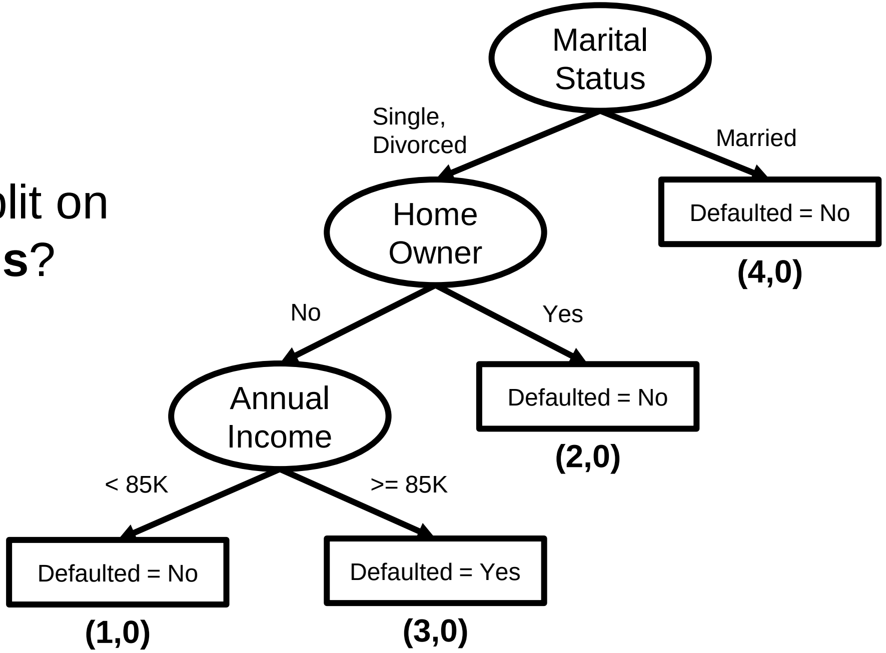
$\frac{3}{8} * \frac{2}{3} = 0.25$

ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
3	No	Single	70K	No
5	No	Divorced	95K	Yes
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
10	No	Single	90K	Yes

Stop until:

- All nodes are pure
- Or early stopping rule is met: Gain < threshold

How about split on **Marital Status**?



ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
3	No	Single	70K	No
5	No	Divorced	95K	Yes
8	No	Single	85K	Yes
10	No	Single	90K	Yes

Decision Tree Classifiers

Advantages:

- Inexpensive to construct
- Extremely fast at classifying unknown records (inference)
- Easy to interpret for small-sized trees
- Robust to noise (especially when methods avoiding overfitting are employed)
- Can easily handle redundant or irrelevant features (unless the features are interacting)

Disadvantages:

- Space of possible decision trees is exponentially large. Greedy approaches are often unable to find the best tree.
- Does not take into account interactions between features
- Each decision boundary involves only a single feature

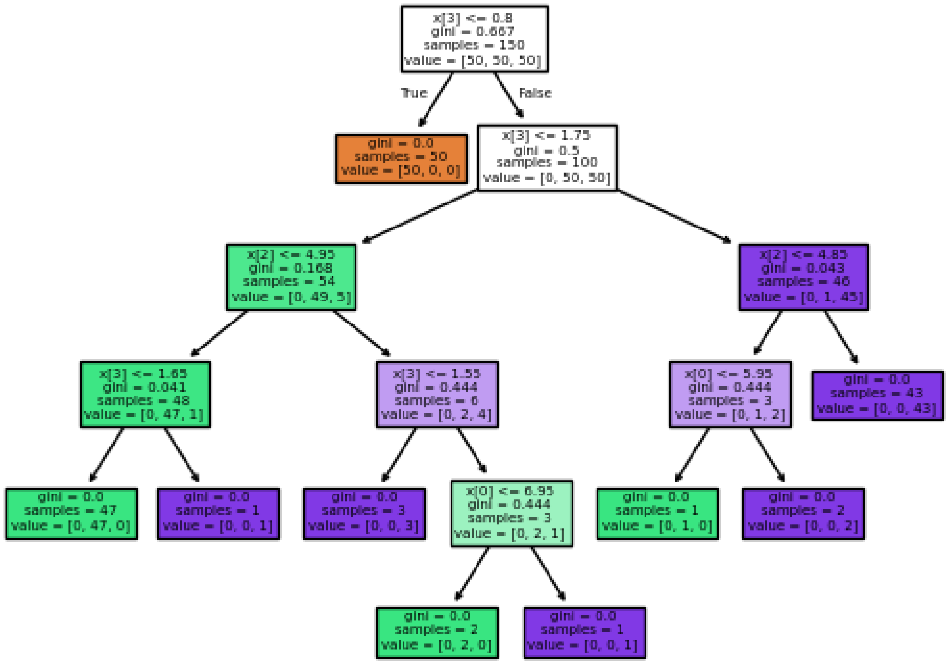
Further Reading

https://scikit-learn.org/stable/modules/tree.html

```
from sklearn.datasets import load_iris
from sklearn import tree
iris = load_iris()
X, y = iris.data, iris.target
clf = tree.DecisionTreeClassifier()
clf = clf.fit(X, y)
```

tree.plot_tree(clf)

Decision tree trained on all the iris features



Assignment 7