upGrad

PLACEMENT REFRESHER PROGRAM

Session 9 - ML 2

Classification

By
Ritesh Kumar Pandey

Agenda

- Classification
- Decision Tree
- Model Performance

Introduction

upGrad

Supervised learning uses labeled data to predict outcomes. For instance, in email spam detection (a classification problem), the algorithm is trained on a dataset where emails are already marked as ‘spam’ or ‘not spam’. It learns patterns from this training set and applies them to new data.

Unsupervised learning, however, doesn’t have labels for its training data. The algorithm must find structure within the data itself. An example of a classification problem here would be customer segmentation. Given purchasing behavior data, an unsupervised algorithm could group customers into different segments without prior knowledge of what these groups might be.

Classification

upGrad

Classification is a data mining task of assigning a data instance to one of the predefined classes/groups based upon the knowledge gained from previously seen(classified data)

Types of attributes

Attribute	Continuous	Discrete
Age	11,12,15,25, etc.	Child, youngster, senior citizen, etc.
Income	30K, 45K, 60K, etc.	Low, High, etc.

Prediction:
predicts continuous values
Eg. How much will a customer spend?

Classification:
predicts discrete values
Eg. Is giving loan to a customer safe or risky?

Types of Classification Algorithm

upGrad

- Naive Bayes
- Decision Tree
- Random Forest
- K-Nearest Neighbour
- Neural Networks
- Logistic Regression

Decision Tree

upGrad

A decision tree is a flowchart-like tree structure, where each internal node (non-leaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label. The topmost node in a tree is the root node.

Decision Tree

upGrad

Tid	House Owner	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

categorical, categorical, continuous, class

Training Data

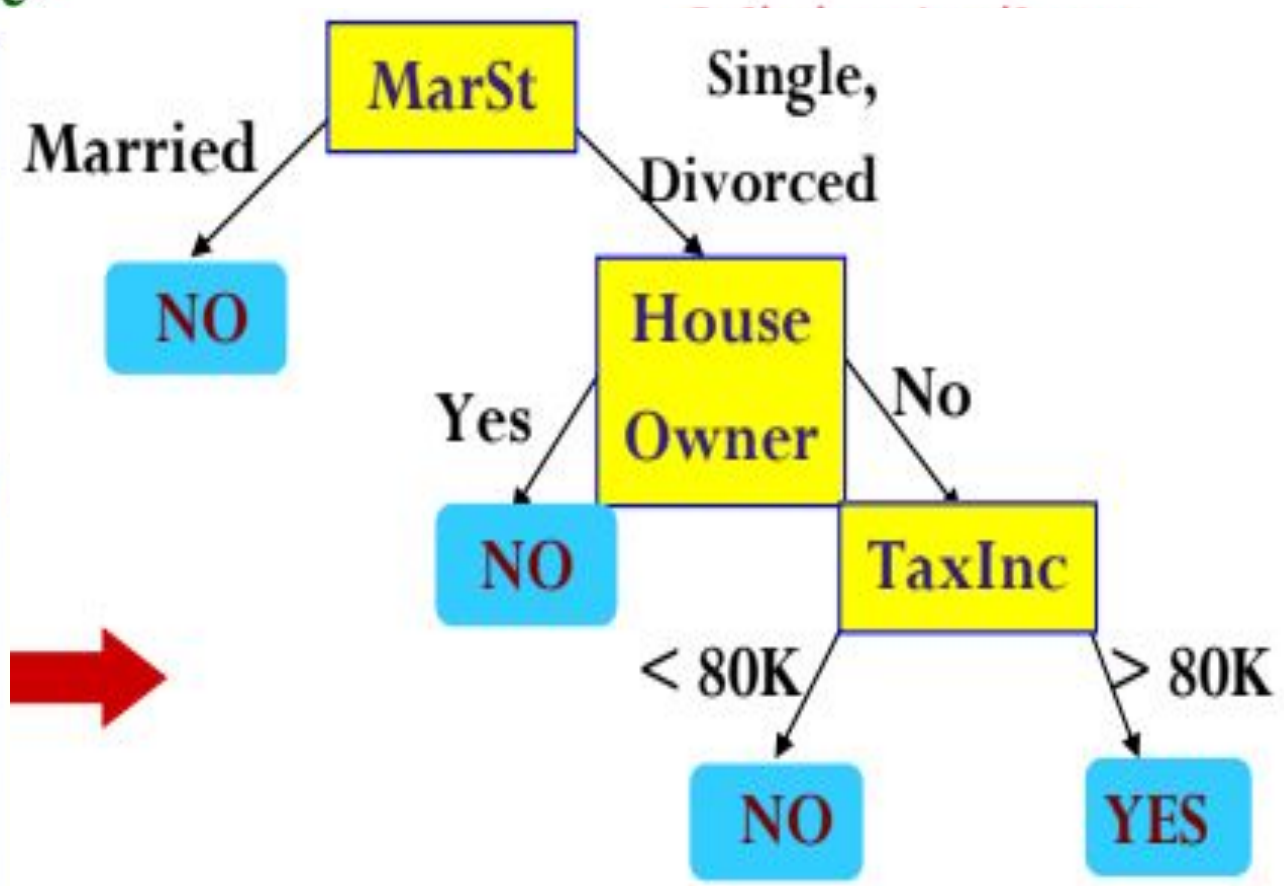
Model: Decision Tree

Decision Tree

upGrad

Tid	House Owner	Marital Status	Taxable Income	Cheat ?
	categorical	categorical	continuous	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

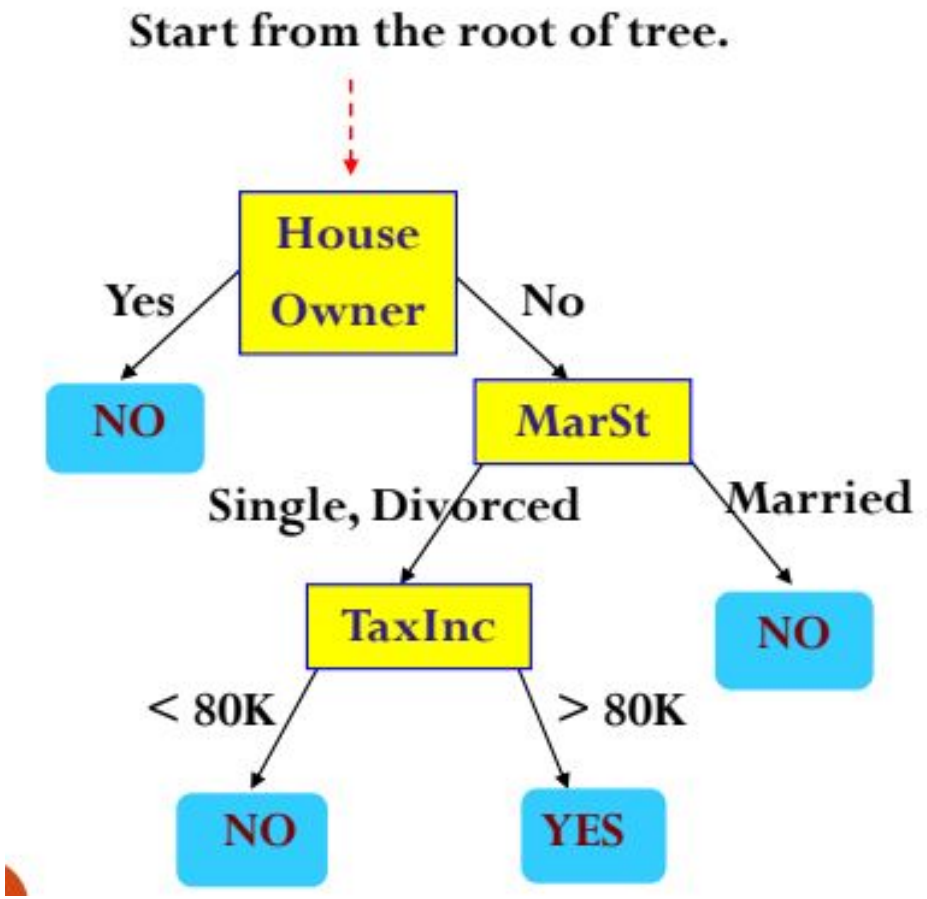
Training Data



Model: Decision Tree

There can be more than one tree that fits the same data

Apply Model to Test Data



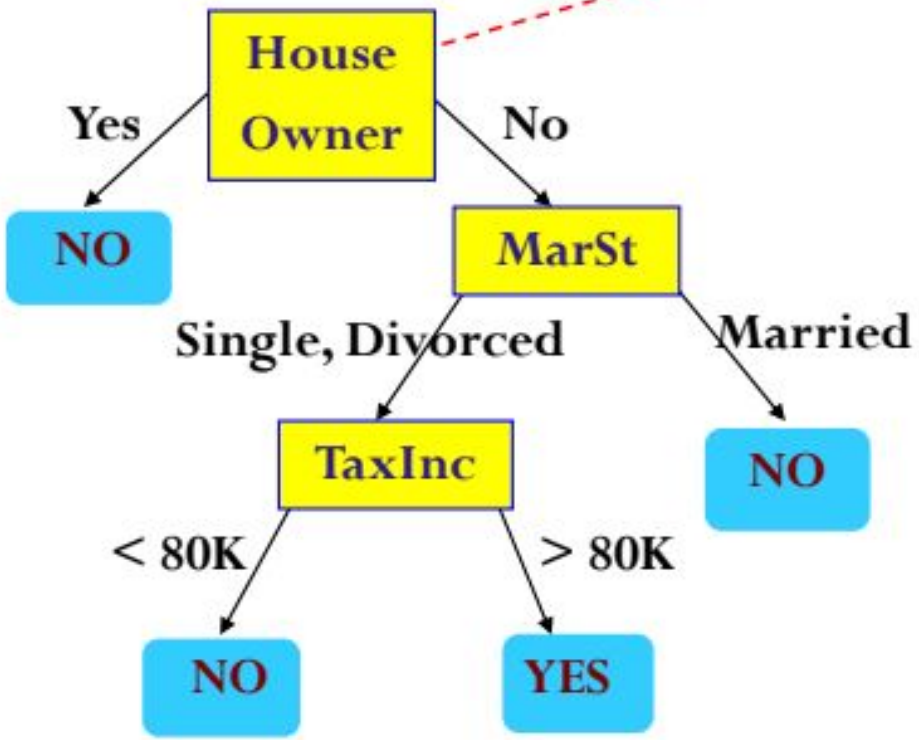
Test Data

House Owner	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Apply Model to Test Data

Test Data

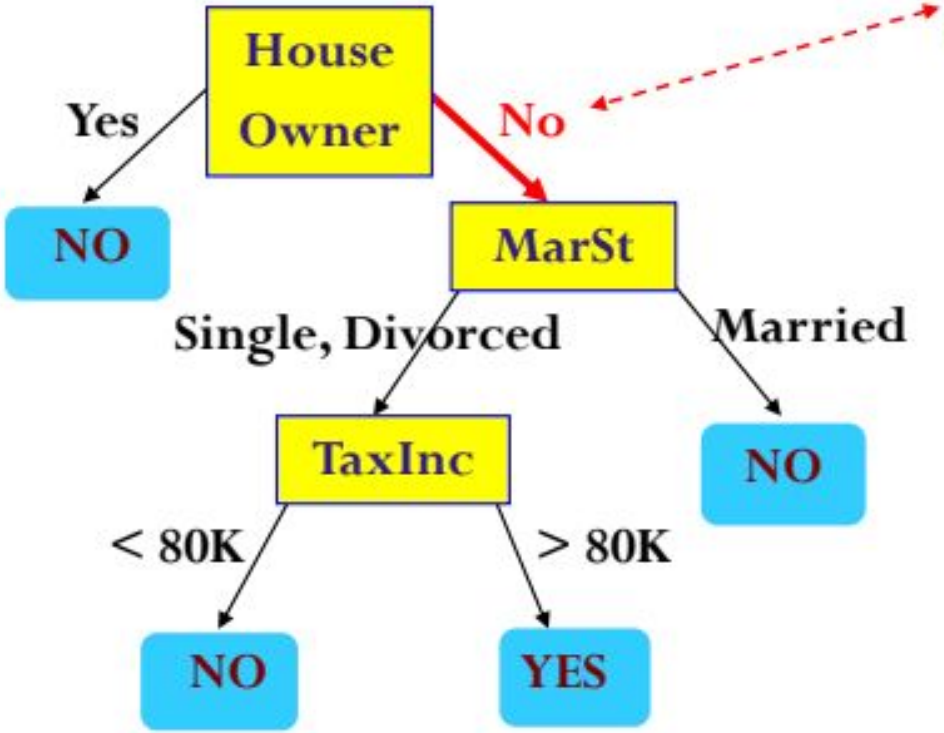
House Owner	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Apply Model to Test Data

Test Data

House Owner	Marital Status	Taxable Income	Cheat
No	Married	80K	?

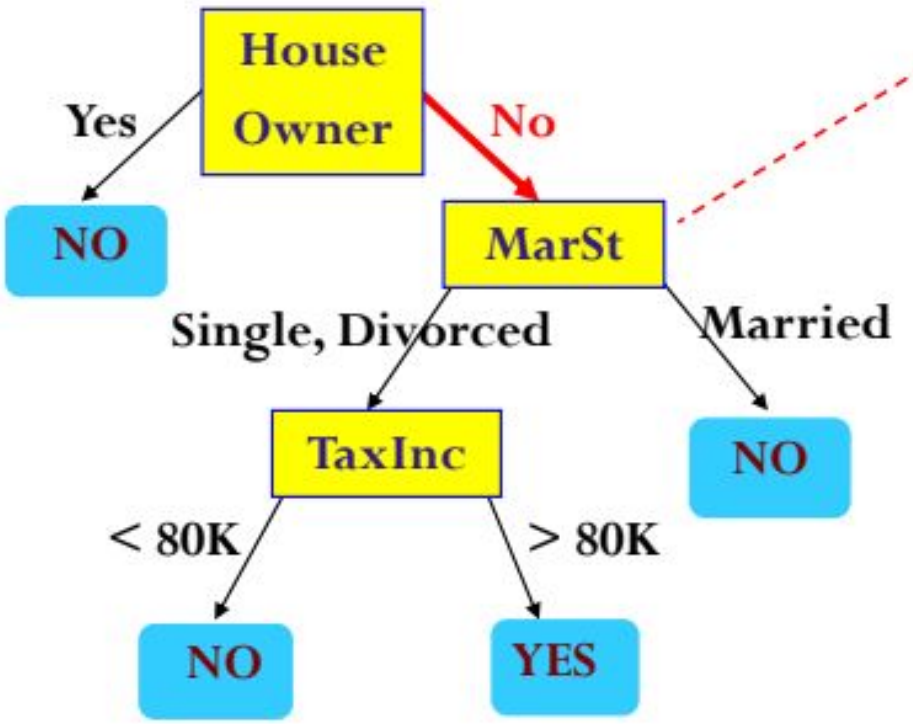


upGrad

Apply Model to Test Data

Test Data

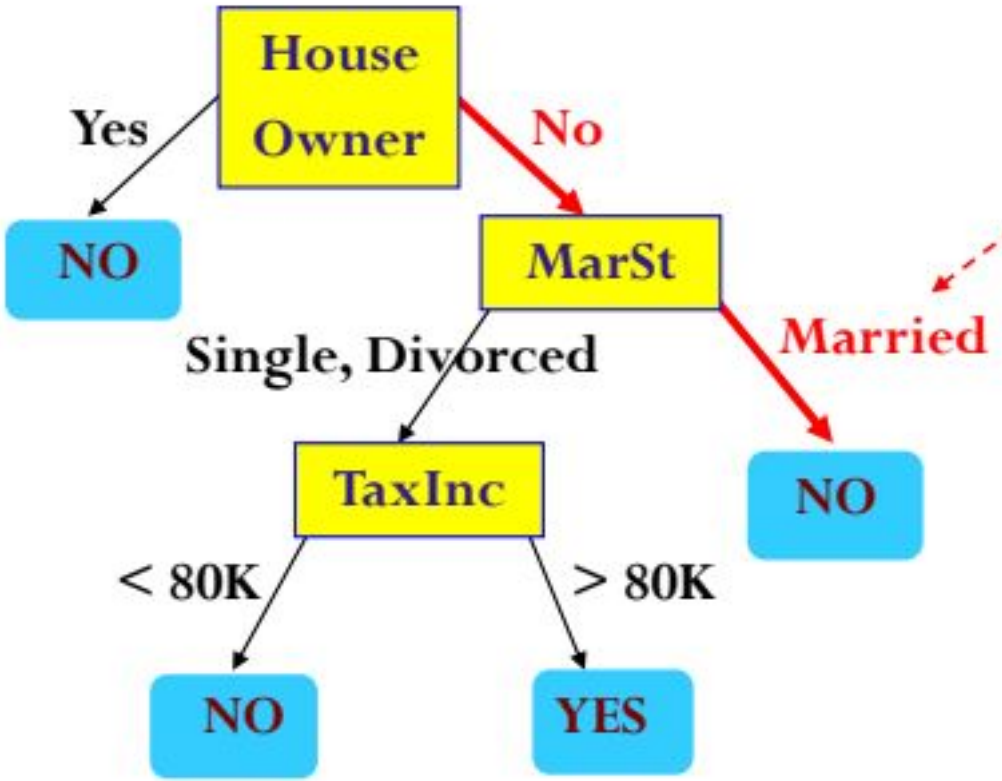
House Owner	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Apply Model to Test Data

Test Data

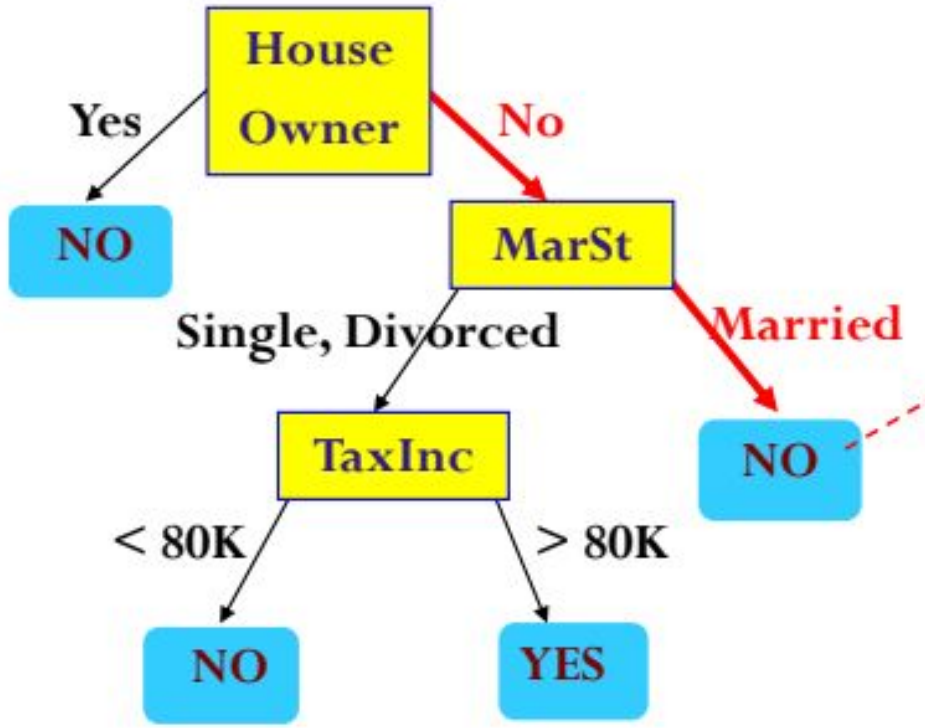
House Owner	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Apply Model to Test Data

Test Data

House Owner	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Assign Cheat to "No"

Model Performance Evaluation

upGrad

Confusion Matrix:

Actual class\Predicted class	C_1	$\neg C_1$
C_1	**True Positives (TP)**	**False Negatives (FN)**
$\neg C_1$	**False Positives (FP)**	**True Negatives (TN)**

Example of Confusion Matrix:

Actual class\Predicted class	buy_computer = yes	buy_computer = no	Total
buy_computer = yes	**6954**	**46**	7000
buy_computer = no	**412**	**2588**	3000
Total	7366	2634	10000

Model Performance Evaluation

upGrad

Measure	*Formula*
accuracy, recognition rate	$\frac{TP+TN}{P+N}$
error rate, misclassification rate	$\frac{FP+FN}{P+N}$
sensitivity, true positive rate, recall	$\frac{TP}{P}$
specificity, true negative rate	$\frac{TN}{N}$
precision	$\frac{TP}{TP+FP}$
F, F_1, F-score, harmonic mean of precision and recall	$\frac{2 \times precision \times recall}{precision + recall}$

Model Performance Evaluation - Example

upGrad

Actual Class\Predicted class	cancer = yes	cancer = no	Total	Recognition(%)
cancer = yes	**90**	**210**	300	30.00 (*sensitivity*
cancer = no	**140**	**9560**	9700	98.56 (*specificity)*
Total	230	9770	10000	96.40 (*accuracy*)

- *Precision* = 90/230 = 39.13% *Recall* = 90/300 = 30.00%

Model Performance Evaluation - Question

upGrad

Sr. No.	Email (Spam/Not Spam)	Prediction
1	Spam	Spam
2	Spam	Not Spam
3	Not Spam	Not Spam
4	Spam	Not Spam
5	Not Spam	Not Spam
6	Not Spam	Spam
7	Spam	Spam
8	Not Spam	Not Spam
9	Not Spam	Spam
10	Spam	Not Spam

Find

1. Accuracy
2. Precision
3. Recall
4. True Positive Rate
5. Sensitivity

Model Performance Evaluation - Question

upGrad

Sr. No.	Email (Spam/Not Spam)	Prediction	
1	Spam	Spam	TP
2	Spam	Not Spam	FN
3	Not Spam	Not Spam	TN
4	Spam	Not Spam	FN
5	Not Spam	Not Spam	TN
6	Not Spam	Spam	FP
7	Spam	Spam	TP
8	Not Spam	Not Spam	TN
9	Not Spam	Spam	FP
10	Spam	Not Spam	FN

Model Performance Evaluation - Question

upGrad

Confusion Matrix:

Actual class\Predicted class	C_1	$\neg C_1$
C_1	**True Positives (TP)**	**False Negatives (FN)**
$\neg C_1$	**False Positives (FP)**	**True Negatives (TN)**

Actual \ Predicted	**Email = Spam**	**Email = Not Spam**	**Total**
Email = Spam	2	3	5
Email = Not Spam	2	3	5
Total	4	6	10

Model Performance Evaluation - Question

upGrad

Actual \ Predicted	Email = Spam	Email = Not Spam	Total
Email = Spam	2 (TP)	3 (FN)	5 (P)
Email = Not Spam	2 (FP)	3 (TN)	5 (N)
Total	4	6	10

1. Accuracy = 0.5
2. Precision = 0.5
3. Recall = 0.4
4. True Positive Rate = 0.4
5. Sensitivity = 0.4

THANK YOU