Workshop 2

BUSA90542 Machine Learning and AI for Business

Entropy

- Entropy is a measure of the *average information* produced by a random process, or contained within a random sample.

$$H(X)=-\sum_x p(x)\log_2 p(x)$$

- The set {A,B,C,A,A,A,A,A} has **low entropy**: low uncertainty and high purity
- The set {A,B,C,D,B,E,A,F} has **high entropy**: high uncertainty and low purity

Entropy

- There are 6 animals, 3 dogs and 3 ducks. Pick one at random, and you can ask a true-false question about it. Then you need to identify whether it is a duck or a dog. What question will you ask?

	Colour	Num of feet	Category
1	Black	4	Dog
2	Yellow	2	Duck
3	White	2	Duck
4	White	2	Duck
5	Yellow	4	Dog
6	Black	4	Dog

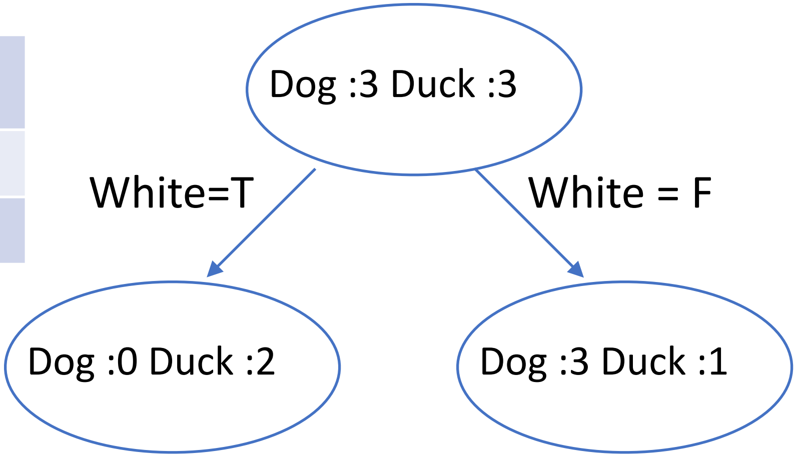
Information gain

- Commonly used split criterion in decision tree
- Based on entropy

$$IG(Y,X) = H(Y) - H(Y|X)$$

$$H(Y|X) = \sum_x p(x) H(Y|X=x)$$

White	Dog	Duck
T	0	2
F	3	1



Dog :3 Duck :3

White=T

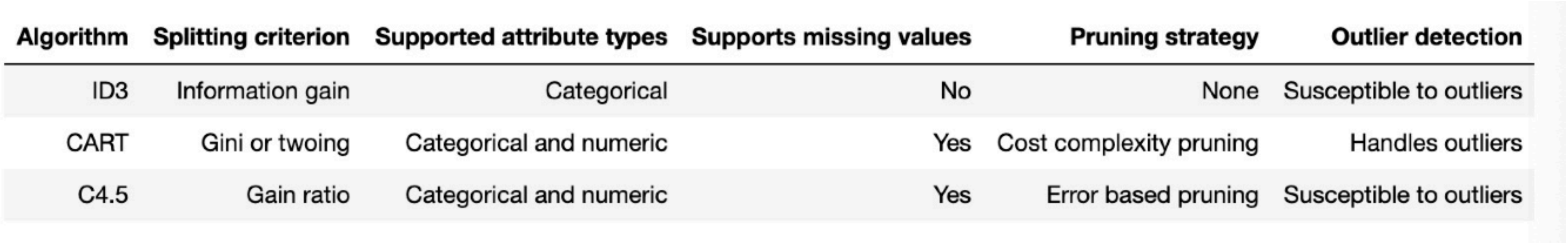
White = F

Dog :0 Duck :2

Dog :3 Duck :1

Decision Tree

- Three most popular decision tree algorithms

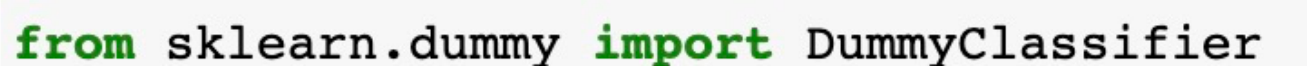


Algorithm	Splitting criterion	Supported attribute types	Supports missing values	Pruning strategy	Outlier detection
ID3	Information gain	Categorical	No	None	Susceptible to outliers
CART	Gini or twoing	Categorical and numeric	Yes	Cost complexity pruning	Handles outliers
C4.5	Gain ratio	Categorical and numeric	Yes	Error based pruning	Susceptible to outliers

- In sklearn, the default decision tree is CART, you can change the criterion to “entropy” to get a ID3.

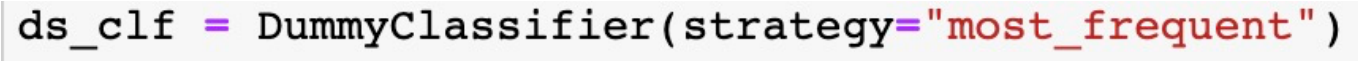
sklearn

- Import training algorithm from sklearn



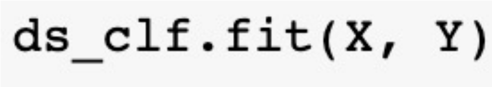
```
from sklearn.dummy import DummyClassifier
```

- Define the model, setup the hyper parameters



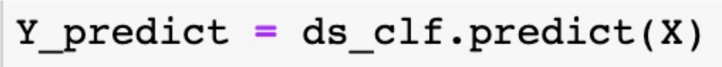
```
ds_clf = DummyClassifier(strategy="most_frequent")
```

- Fit the model with the training data



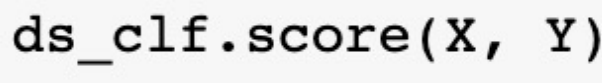
```
ds_clf.fit(X, Y)
```

- Make prediction



```
Y_predict = ds_clf.predict(X)
```

- Evaluate the model



```
ds_clf.score(X, Y)
```

Bank

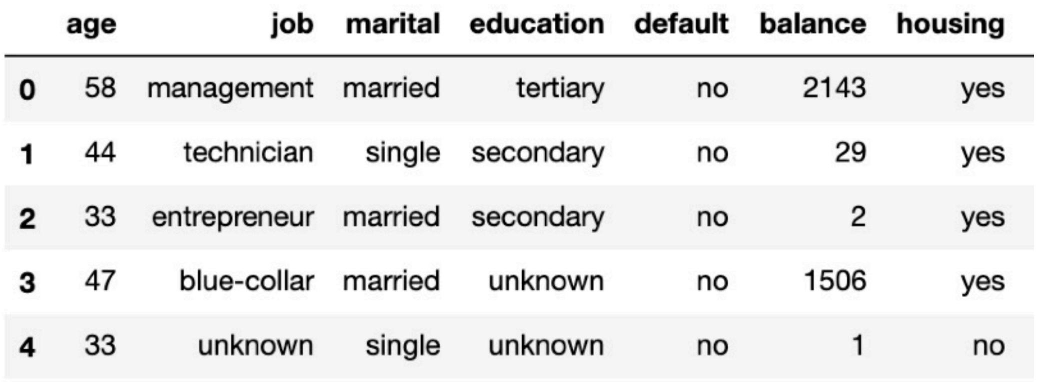
1. Define the goal, understand the task.
   - The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe a term deposit (variable y).[1]
   - Supervised or unsupervised?
   - Classification or Regression?

- [1] https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

Bank

2. Understand and preprocess the dataset.

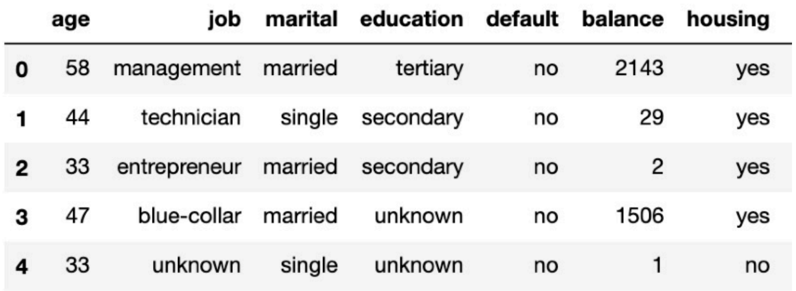
- How many features in this dataset?
- How many instances in this dataset?



	age	job	marital	education	default	balance	housing
0	58	management	married	tertiary	no	2143	yes
1	44	technician	single	secondary	no	29	yes
2	33	entrepreneur	married	secondary	no	2	yes
3	47	blue-collar	married	unknown	no	1506	yes
4	33	unknown	single	unknown	no	1	no

- Preprocess
  - We can't directly feed features like **job** into the model

Bank



	age	job	marital	education	default	balance	housing
0	58	management	married	tertiary	no	2143	yes
1	44	technician	single	secondary	no	29	yes
2	33	entrepreneur	married	secondary	no	2	yes
3	47	blue-collar	married	unknown	no	1506	yes
4	33	unknown	single	unknown	no	1	no

- Preprocess
  - We can't directly feed features like **job** into the model
  - one-hot encode

Size	Colour	Y
10	Red	1
4	Green	0
2	Blue	0
5	Blue	1

⇨

Size	Red	Green	Blue	Y
10	1	0	0	1
4	0	1	0	0
2	0	0	1	0
5	0	0	1	1