Machine learning

What is machine learning?

Informally: the ability of a computer program to learn (find patterns, discover relationships) without being explicitly programmed*

*but, of course, we have to program the computer to learn without being explicitly programmed!

What is machine learning?

Machine learning works through induction:
study lots of examples and generalize
or induce a rule from those examples

Machine learning is data driven!

Types of machine learning

1. Supervised learning: The program learns by studying paired ‘input – output’ observations

   - Classification: predict which category a data point belongs to
   - Regression: predict the value (continuous) of the output for given inputs

2. Unsupervised learning: The program learns by studying untagged data – for example by dividing the dataset into ‘similar’ groups

   - Clustering: group similar observations together
   - Dimensionality reduction: eliminate unimportant explanatory variables

Training and Testing

1. Training dataset: a subset of the data that the program uses to ‘learn’

2. Testing dataset: a subset of the data that the program uses to test what it has ‘learned’

3. Validation dataset: a subset of the data that the program uses to tune its learning parameters

4. Cross validation: the dataset is partitioned into n datasets and the program trains on n-1 datasets, tests on one set, and then repeats the process with different training and testing subsets

Performance measures

1. supervised learning:
    1. prediction errors
    2. bias vs. variance

2. unsupervised learning:
    1. accuracy: fraction of cases that are correctly classified.
    2. precision: fraction of cases that are classified as type x to the cases that are actually of type x
    3. recall: fraction of cases that are correctly classified as type x to the total number of cases that are of type x in the data

Estimators

Estimators are models used for prediction.

First, fit the model to the data

then use the model to predict

cost functions measure the error

residuals measure in-sample error

Preparation Steps

1. Preprocessing: Ensure – if necessary – that different data elements have similar distributions or lie in the same range
2. standardization: Transform the data to have mean zero and std one.

```
from sklearn import preprocessing
preprocessing.scale(x)
```

3. min-max: Transform the data so that all variables lie in the same min-max range

```
from sklearn import preprocessing
min_max_scaler = preprocessing.MinMaxScaler(feature_range=(10,100))
x_data = min_max_scaler.fit_transform(x_train)
```

4. Other preprocessing: norms, binarization (see http://scikit-learn.org/stable/modules/preprocessing.html)

Example: Linear Regression

```
from sklearn.linear_model import LinearRegression


model = LinearRegression()
model.fit(x,y)                              #Fit
model.predict(1000)                         #Predict
np.mean((model.predict(x) - y) ** 2)        #Residuals
model.score(x,y)                            #R-squared
```

Decision trees

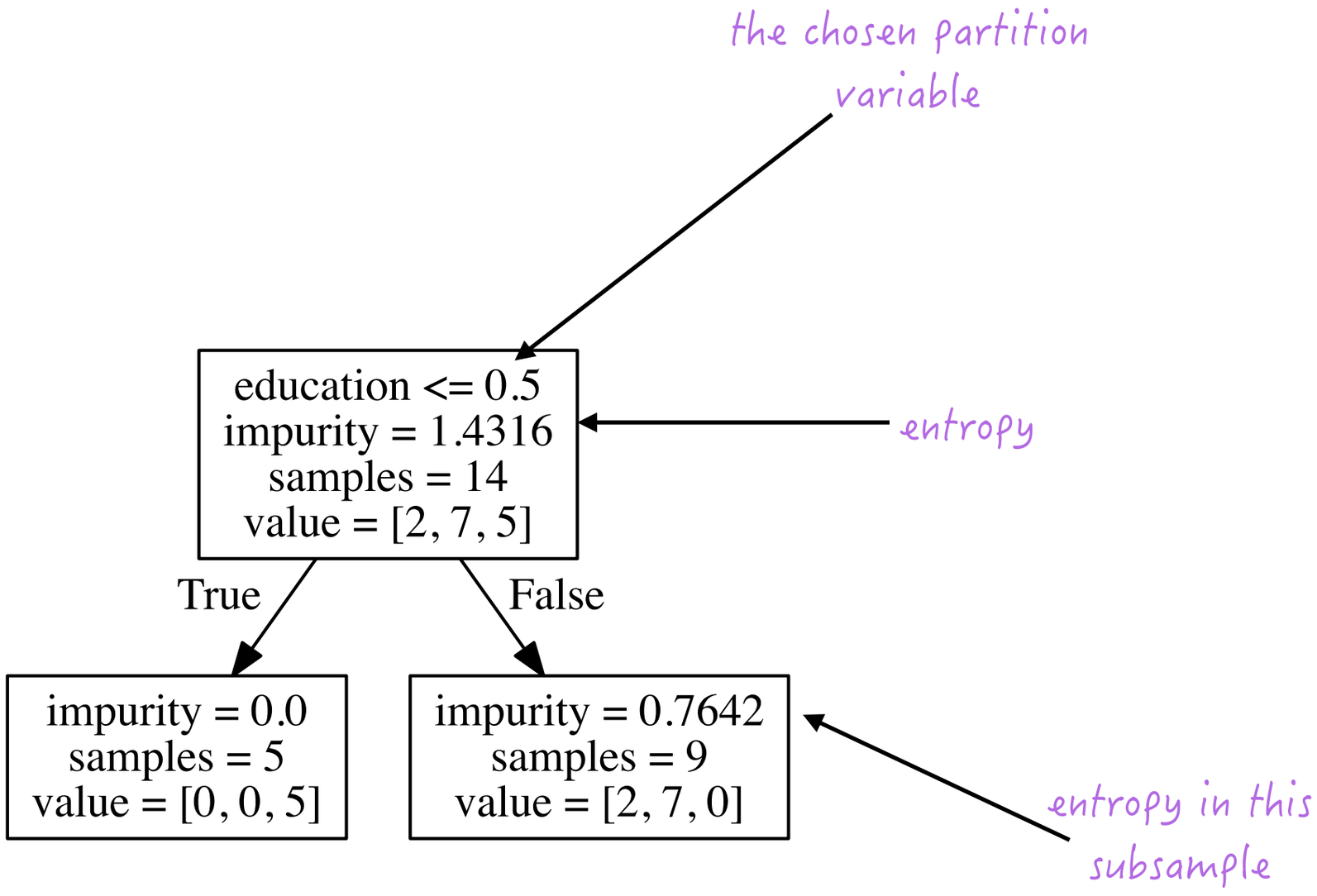
1. Entropy: a measure of uncertainty in the data
    1. what is the uncertainty in color when you draw a marble from a box of 100 blue marbles?
    2. what is the uncertainty when you draw a marble from a box with 50 blue and 50 red marbles?
2. Entropy minimization: decision tree algorithms seek to partition the data on features in the way that total entropy is minimized

Decision trees with scikit-learn

```
from sklearn.tree import DecisionTreeClassifier
from sklearn import tree
```

```
clf = DecisionTreeClassifier(max_depth=1,criterion="entropy")
clf.fit(x_train,y_train)
clf.score(x_train,y_train)
clf.predict(x_test)
```

Tree of depth 1



Tree of depth 2