

# Exercise 9

## Social Data Science

### 1 Classifying Movie Scores

We revisit the list of movies which has been introduced in exercise 5:

Title	Year	IMDB score	Age rating	Length (min)	Genre
The Lighthouse	2019	7.5	16	109	Drama
High Life	2018	5.8	16	113	Adventure
Damsel	2018	5.5	12	113	Adventure
Good Time	2017	7.4	12	101	Drama
Life	2015	6.1	0	111	Biography
Queen of the Desert	2015	5.7	0	128	Biography
Twilight: Breaking Dawn pt. 2	2012	5.5	12	115	Drama
Twilight: Breaking Dawn pt. 1	2011	4.9	12	117	Adventure
Remember Me	2010	7.1	12	113	Drama
Twilight: New Moon	2009	4.7	12	130	Adventure
Twilight	2008	5.2	12	122	Drama
Harry Potter and the Goblet of Fire	2005	7.7	12	157	Adventure

Like before, we want to make a prediction on the IMDB score, but this time we only want to predict if the IMDB score is bigger than 7.0 or not, i.e. we have a binary prediction task. The features we would like to use are

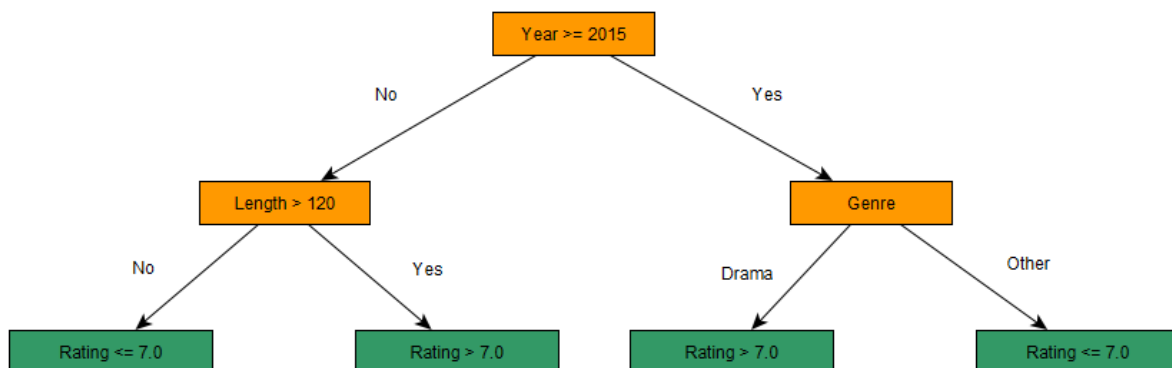
- Year  $\geq 2015$  (binary)
- Age rating (categorical)
- Length  $\geq 2h$  (binary)
- Genre (categorical)

## 1.1 Naive Bayes

Apply the simple count-based Naive Bayes algorithm that was presented in lecture to predict whether the the more recent films *The King* (released 2019, age rating 16 years, 140 minutes, Biography), *The Devil All the Time* (released 2020, age rating 16 years, 138 minutes, Drama), and *Tenet* (released 2020, age rating 12 years, 150 minutes, Drama) will receive a rating over 7.0. Use the full dataset above for training, and give all probabilities that are needed to make the prediction.

## 1.2 Decision Trees

Assume that instead of a Naive Bayes Classifier we have trained a decision tree classifier on the movie data, which yields the following tree structure:



Give the predictions of the three more recent movies!

## 1.3 Evaluation and Diagnostics

Assume that you have trained a classifier that yields the following binary predictions (IMDB score  $\geq 7.0$ ) over the training data:  $\hat{y} = (1, 1, 0, 1, 1, 0, 0, 0, 0, 0, 1)$ . Compute the accuracy and the confusion matrix of these predictions!

## 2 Nearest Neighbor Classification

Consider the following data set:

$P_1 = (0, 4), P_2 = (0, 6), P_3 = (1, 3), P_4 = (1, 5), P_5 = (2, 1), P_6 = (2, 4),$   
 $P_7 = (2, 6), P_8 = (3, 0), P_9 = (3, 3), P_{10} = (3, 5), P_{11} = (5, 2), P_{12} = (5, 4).$

The data set contains the following two classes:

- red =  $\{P_1, P_2, P_4, P_5, P_6, P_8, P_{10}\}$
- blue =  $\{P_3, P_7, P_9, P_{11}, P_{12}\}.$

Classify all data points with the 3-Nearest Neighbor Classifier by ignoring their true class labels. Use the Euclidean distance and the majority voting criteria to determine the classes.

## 3 $k$ -Means Clustering

- Give an example of a dataset consisting of four data vectors where there exist two different optimal (minimum sum of squared errors) 2-means clusterings of the dataset!
- Perform two iterations of the k-means algorithm in order to obtain two clusters for following set of points:

$$P_1(12, 15), P_2(12, 33), P_3(18, 15), P_4(18, 27), P_5(24, 21), P_6(36, 42)$$

Assume that the initial centroids are  $P_1$  and  $P_3$ . Explain if more iterations are needed to get the final clusters!