Prof. Dr. Markus Strohmaier
Dr. Ivan Smirnov
Tobias Schumacher

Exercise 9

Social Data Science

1 Classifying Movie Scores

We revisit the list of movies which has been introduced in exercise 7:

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

Prof. Dr. Markus Strohmaier
Dr. Ivan Smirnov
Tobias Schumacher

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.

Naive Bayes: We want to predict/model

$$P(Y=c \mid X=x) = P(Y=c \mid X_1=x_1, \ldots, X_n=x_n)$$

↳ "Naive" Assumption, namely that all features are stochastically independent:

$$P(X_1=x_1, \ldots, X_n=x_n) = P(X_1=x_1) \cdot P(X_2=x_2) \cdot \ldots \cdot P(X_n=x_n).$$

Then, we can use Bayes Theorem to predict:

$$P(Y=c \mid X=x) = \frac{P(X=x \mid Y=c) \cdot P(Y=c)}{\cancel{P(X=x)}} = \frac{P(Y=c) \cdot P(X_1=x_1 \mid Y=c) \cdot \ldots \cdot P(X_n=x_n \mid Y=c)}{\cancel{P(X=x)}}$$

- denominator not known, but also not relevant, because if want to predict the class c, all conditional Probs. will have the same denominator
- prior probabilities $P(Y=c)$ and cond. probabilities $P(x_i \mid c)$ can be estimated from training data

· Prior Probabilities:

$Y = [\text{IMDB score} \geq 7.0]$

$P(Y=0) = \frac{8}{12} = \frac{2}{3}$

$P(Y=1) = \frac{1}{3}$

- Compute Conditional Probabilities $P(x_i = x \mid Y = c)$ for all x_i that occur in the test data, and all classes $c \in \{0,1\}$

- Year: only ≥ 2015 in all test data

$P(Year \geq 2015 \mid Y=0) = \frac{4}{8} = \frac{1}{2}$

$P(Year \geq 2015 \mid Y=1) = \frac{2}{4} = \frac{1}{2}$

- length: all movies $\geq 2h$ in test data

$P(length \geq 2h \mid Y=0) = \frac{3}{8}$

$P(length \geq 2h \mid Y=1) = \frac{1}{4}$

- age rating: 12 and 16 in test data

$P(age\ 12 \mid Y=0) = \frac{5}{8}$

$P(age\ 12 \mid Y=1) = \frac{3}{4}$

$P(age\ 16 \mid Y=0) = \frac{1}{8}$

$P(age\ 16 \mid Y=1) = \frac{1}{4}$

- genre: either Biography or Drama

$P(bio \mid Y=0) = \frac{1}{4}$

$P(bio \mid Y=1) = 0$

$P(drama \mid Y=0) = \frac{1}{4}$

$P(drama \mid Y=1) = \frac{3}{4}$

Now Predict $\hat{Y}$

- The King (2019, age 16, $\geq 2h$, bio)

$\hookrightarrow P(Y=0 \mid King) \sim P(Y=0) \cdot P(\geq 2015 \mid Y=0) \cdot P(age\ 16 \mid Y=0) \cdot P(bio \mid Y=0) \cdot P(\geq 2h \mid Y=0)$

$= \frac{2}{3} \cdot \frac{1}{2} \cdot \frac{1}{8} \cdot \frac{1}{4} \cdot \frac{3}{8} = \frac{1}{256}$

$P(Y=1 \mid King) \sim 0 < P(Y=0 \mid King) \rightarrow$ predict $\hat{Y} = 0$

- Devil all the time:

$P(Y=0 \mid Devil) \sim \frac{1}{256}$

$P(Y=1 \mid Devil) \sim \frac{1}{128}$

} predict $\hat{Y} = 1$

- Tenet:

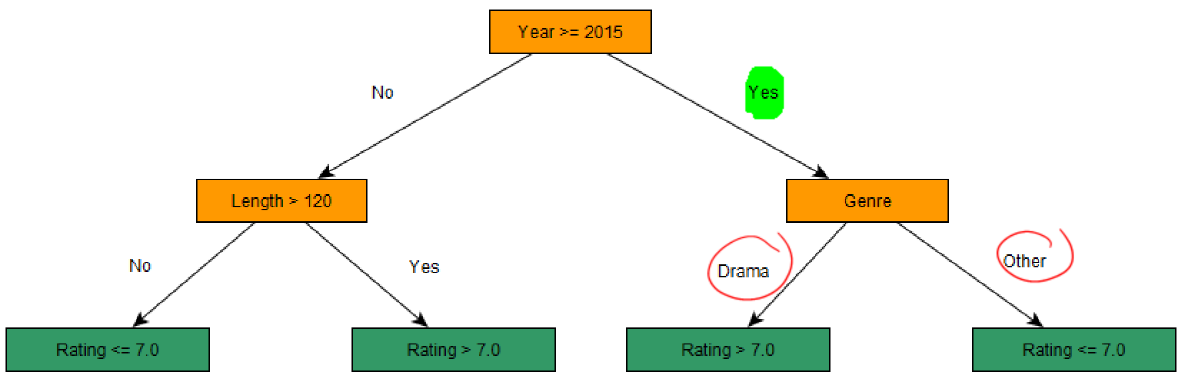
$P(Y=0 \mid Tenet) \sim \frac{5}{256}$

$P(Y=1 \mid Tenet) \sim \frac{3}{128}$

} predict $\hat{Y} = 1$

1.2 Decision Trees

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



Give the predictions of this tree on each of the movies in the test dataset, as well as the three more recent movies!

- all "test" movies after 2015
- Now we have that the King is Biography → predict rating ≤ 7.0
- Devil all the Time and Tenet are Drama → predict rating > 7.0

Prof. Dr. Markus Strohmaier
Dr. Ivan Smirnov
Tobias Schumacher

1.3 Evaluation and Diagnostics

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

True Y	1	0	0	1	0	0	0	0	1	0	0	1
Predictions $\hat{Y}$	1	1	0	1	1	0	0	0	0	0	0	1

accuracy : $\frac{\#[\hat{y} = y]}{N} = \frac{9}{12} = \frac{3}{4}$

Confusion Matrix

Predictions \ True	0	1
0	6	1
1	2	3

Prof. Dr. Markus Strohmaier
Dr. Ivan Smirnov
Tobias Schumacher

2 Nearest Neighbor Classification

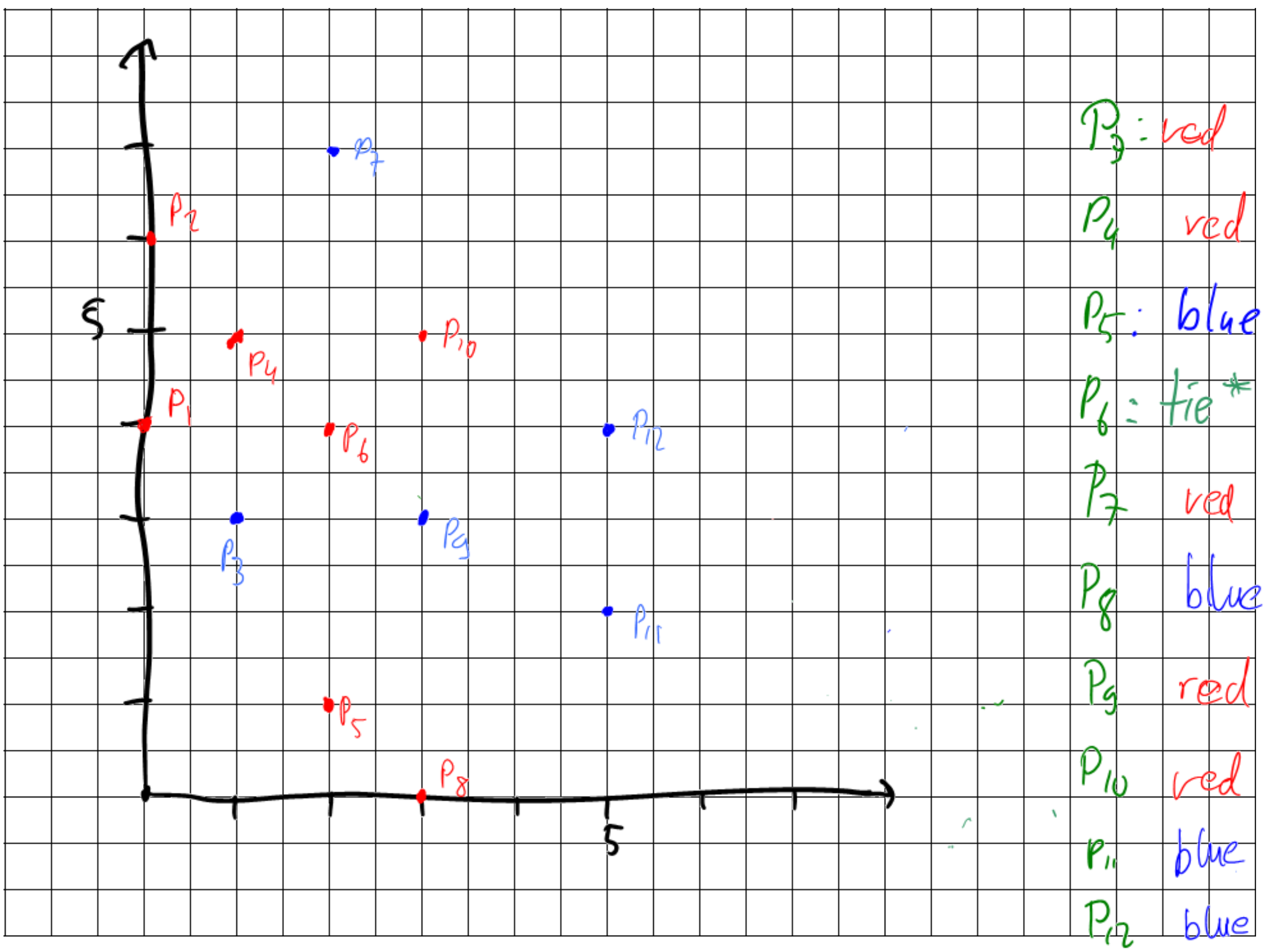
Consider the following data set:

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

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.



P_1 has 3 nearest neighbors P_3, P_4, P_6 → classify as red

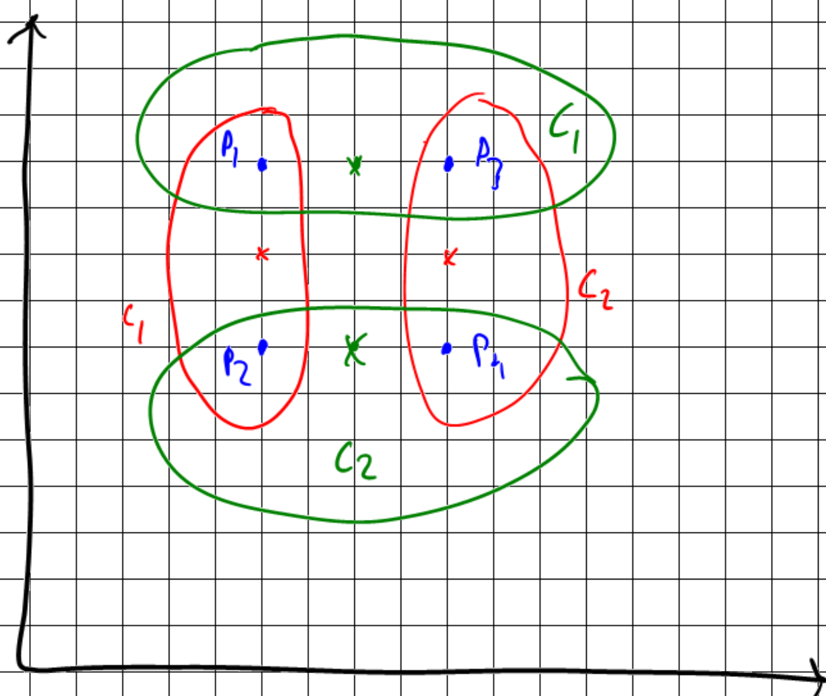
P_2 —— " —— P_1, P_4, P_7 → classify as red

* we have that P_4, P_{10}, P_7, P_9 are all at the same distance → break tie randomly
→ random prediction

Prof. Dr. Markus Strohmaier
Dr. Ivan Smirnov
Tobias Schumacher

3 k-Means Clustering

a) 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!



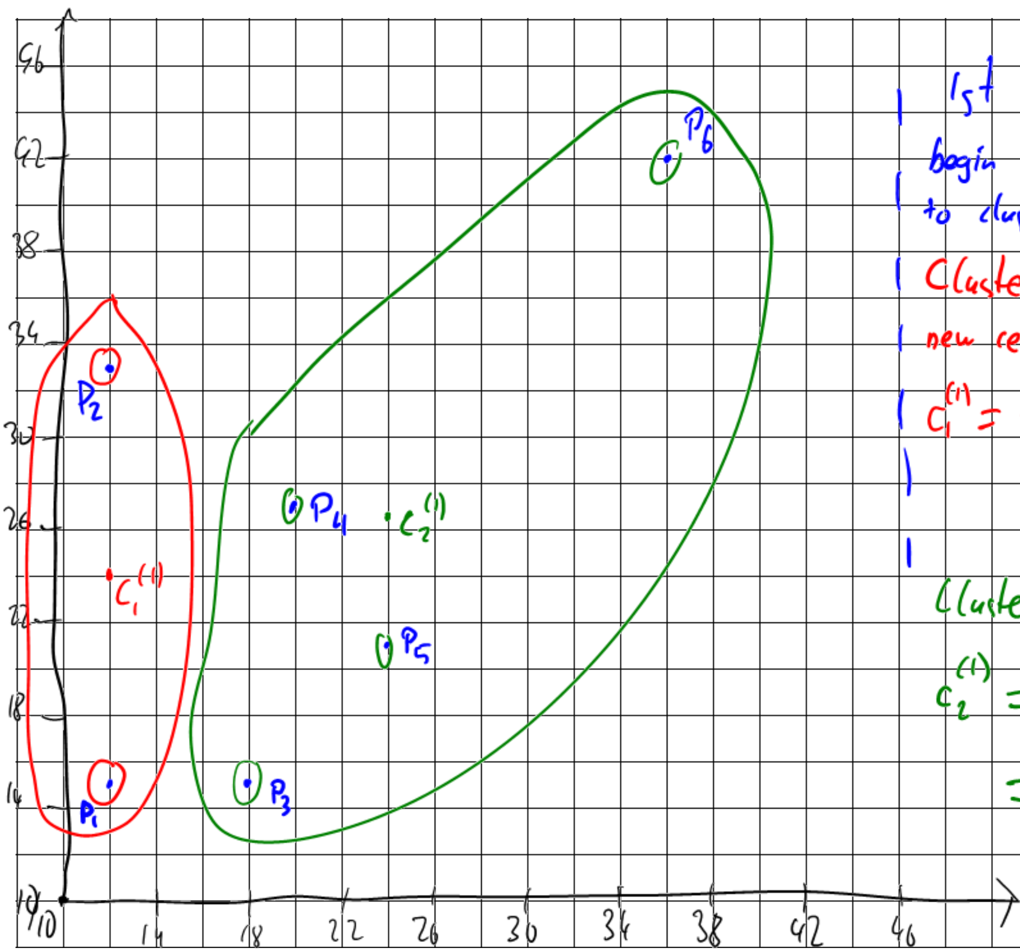
In both cases, the cluster centers of C_1 and C_2 would all have ditances of 2 to the point in their clusters

Prof. Dr. Markus Strohmaier
Dr. Ivan Smirnov
Tobias Schumacher

b) 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!



1st iteration:

begin with assigning points to clusters

Cluster 1 : P_1, P_2

new centroid $c_1^{(1)}$:

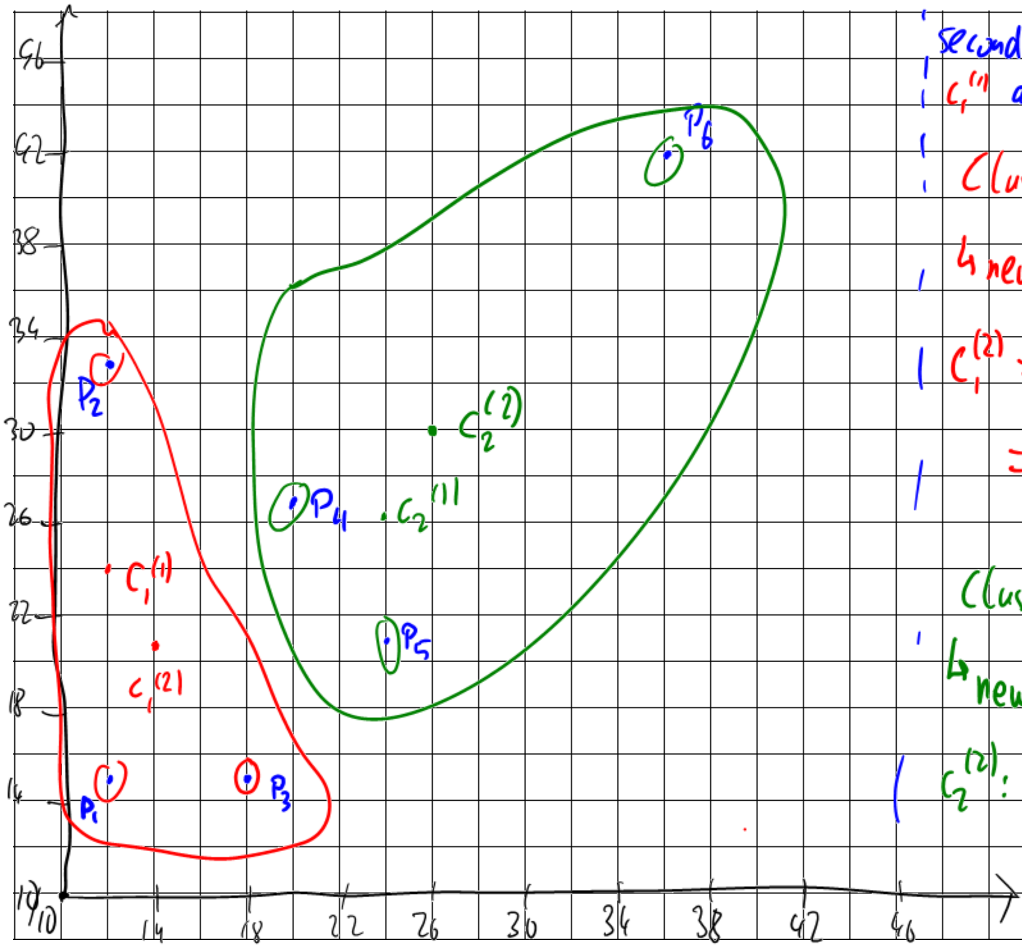
$c_1^{(1)} = \frac{1}{2}(P_1 + P_2) = (12, 24)$

Cluster 2 : P_3, P_4, P_5, P_6

$c_2^{(1)} = \frac{1}{4}(P_3 + P_4 + P_5 + P_6)$

$= (24, 26.25)$

Prof. Dr. Markus Strohmaier
Dr. Ivan Smirnov
Tobias Schumacher



second iteration, $c_1^{(1)}$ and $c_2^{(1)}$ cluster centers

Cluster 1: P_1, P_2, P_3

↳ new centroid:

$c_1^{(2)} = \frac{1}{3}(P_1 + P_2 + P_3)$

$= (14, 21)$

Cluster 2: P_4, P_5, P_6

↳ new cluster center

$c_2^{(2)}: \frac{1}{3}(P_4, P_5, P_6)$

$= (26, 30)$

We expect more iterations, as P_2 seems to be closer to $c_1^{(2)}$ than to $c_2^{(2)}$.