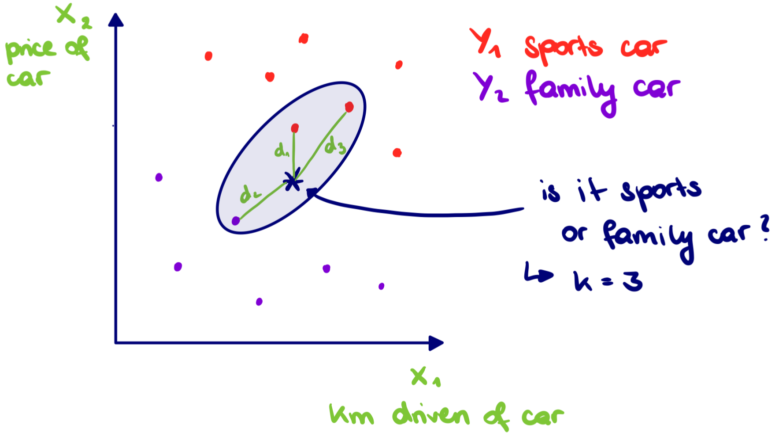
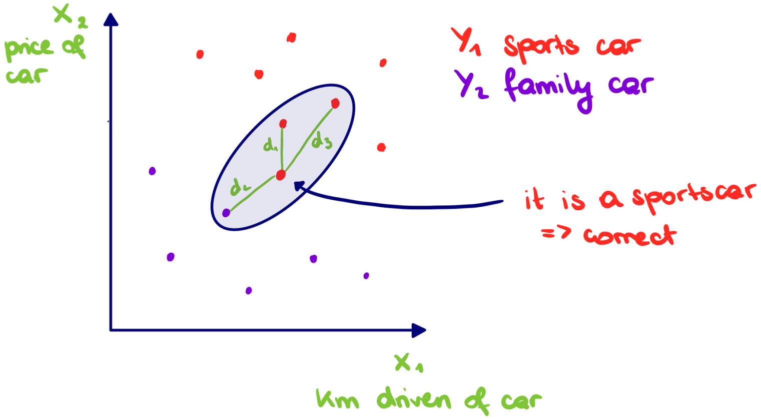
k nearest neighbours

- Belongs to the group of supervised machine learning algorithms
- Can be used for classification and regression
  - Classification: classes of the k-nearest neighbors are considered, the most often occurring class among those is the predicted class
  - Regression: values of the k-nearest neighbors are considered, the average value of those is the predicted value
- Relies on the distance between data points, so if the ranges of feature values are vastly different, a scaling can be helpful, e.g. standardization, mean normalization, min-max normalization
- Additionally, one can put different weights on the neighboring values, so that e.g. nearest values have a higher impact on the prediction



k nearest neighbours

- Process:
  - Split data into test- and training-sample using 10-fold cross validation process
  - Predict class for each data point of test-sample by using training sample:
    1. Calculate Euclidean distance of first data point in test sample to each row of training sample and sort by ascending distance
    2. Most occurring class among k nearest data points is predicted value
    3. Repeat step 1 and 2 for each row of test sample and compare each predicted value to actual value
    4. Calculate ratio where prediction was correct in relation to total sample
    5. Repeat step 1 – 4 for each cross validation run
    6. Final performance evaluation is average of all cross validation runs
  - To evaluate the effect of:
    - hyperparameter k, the value was varied (3, 5, 7)
    - feature scaling, the features were adapted beforehand (standardization: $\frac{x_i - \mu}{\sigma}$)


X_2
price of car
Y_1 sports car
Y_2 family car
d_1
d_2
d_3
it is a sportscar
=> correct
X_1
km driven of car