

Machine Learning with Python

2024年2月5日 12:42

Major machine learning techniques

- Regression
- Classification
- Clustering
- Associations
- Anomaly detection
- Sequence mining
- Dimension reduction
- Recommendation systems

Python libraries for machine learning

- Numpy
- Scipy
- Matplotlib
- Pandas
- Scikit-learn

scikit-learn functions

```
from sklearn import preprocessing
X = preprocessing.StandardScaler().fit(X).transform(X)

from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33)

from sklearn import svm
clf = svm.SVC(gamma=0.001, C=100.)

clf.fit(X_train, y_train)

clf.predict(X_test)

from sklearn.metrics import confusion_matrix
print(confusion_matrix(y_test, yhat, labels=[1,0]))

import pickle
s = pickle.dumps(clf)
```

屏幕剪辑的捕获时间: 2024/2/5 13:04

Types of supervised learning

- Classification
The process of predicting a discrete class label, or category
- Regression
The process of predicting a continuous value as opposed to predicting a categorical value in classification

Types of unsupervised learning

- Dimension reduction
- Density estimation

- Market basket analysis
- Clustering

Model evaluation in regression models

Train & Test (where testing set is included in the training set)

Calculate the accuracy of our model

- The error of the model is calculated as the average difference between the predicted and actual values for all the rows

Training accuracy

- High training accuracy isn't necessarily a good thing
- Result of over-fitting: the model is overly trained to the dataset, which may capture noise and produce a non-generalized model

Out-of-sample accuracy

- It's important that our models have a high, out-of-sample accuracy
- How can we improve out-of-sample accuracy?

Answer: use train & test split, mutually exclusive

The issue of train & test split is that it's highly dependent on the datasets on which the data was trained and tested

K-fold cross-validation

- Splits the dataset into 4 folds; e.g. we use the first 25% of the dataset for testing and the rest for training; in the next round, the second 25% is used for testing and the rest for training, we continue for all folds
- Finally, the result of all 4 evaluations are averaged

Evaluation metrics in regression models

- MAE
- MSE
- RMSE
- RAE
- RSE
- $R^2 = 1 - RSE$

K-Nearest Neighbors

- Find one of the closet cases and assign the same class label to our new customer (first-nearest neighbor)
- Rather than choose the first nearest neighbor, we can chose the 5 nearest neighbors and did a majority vote among them to define the class of our new customer

K-Nearest Neighbors algorithm

1. Pick a value for K
2. Calculate the distance of unknown case from all cases
 - Use Euclidean distance: $Dis(x_1, x_2) = \text{the root of } (x_1 - x_2)^2$
3. Select the K-observations in the training data that are "nearest" to the unknown data point
 - How can we find the best value for K?

Answer: reserve a part of the data for testing the accuracy of the model, choose K equals 1 and then use the training part for modeling, and calculate the accuracy of prediction using all samples in your test

set

- Predict the response of the unknown data point using the most popular response value from the K-nearest neighbors

Evaluation metrics in classification

- Jaccard index

y: actual labels, y-hat: predicted labels; Jaccard is the size of the intersection divided by the size of the union of 2 label sets

- F1-score

Assume there's a confusion matrix, where each matrix row shows the Actual labels in the test set, and the columns show the predicted labels by classifier

F1-score

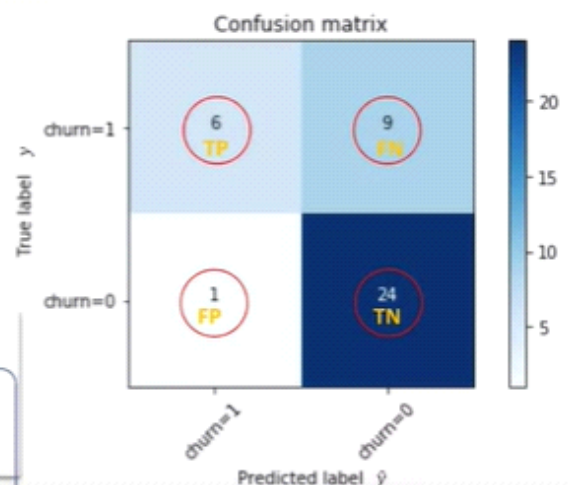
- Precision = $TP / (TP + FP)$
- Recall = $TP / (TP + FN)$
- F1-score = $2 \times (prc \times rec) / (prc + rec)$

F1-score: 0.00 ... 0.20 ... 0.55 ... 0.83 ... 1.00

Higher Accuracy →

	precision	recall	f1-score
Churn = 0	0.73	0.96	0.83
Churn = 1	0.86	0.40	0.55

Avg Accuracy = 0.72



屏幕剪辑的捕获时间: 2024/2/5 15:52

- Log loss

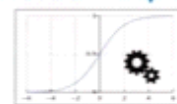
Measures how far the prediction is from the actual label

Log loss

Performance of a classifier where the predicted output is a probability value between 0 and 1.

Test set

	tenure	age	address	income	ed	employ	equip	calcard	wireless	churn
0	11.0	33.0	7.0	136.0	5.0	5.0	0.0	1.0	1.0	1
1	33.0	33.0	12.0	33.0	2.0	0.0	0.0	0.0	0.0	1
2	23.0	30.0	9.0	30.0	1.0	2.0	0.0	0.0	0.0	0
3	38.0	35.0	5.0	76.0	2.0	10.0	1.0	1.0	1.0	0
4	7.0	35.0	14.0	80.0	2.0	15.0	0.0	1.0	0.0	0



Predicted churn	LogLoss
0.91	0.094
0.13	0.89
0.04	0.04
0.23	0.26
0.43	0.56

Actual Labels y

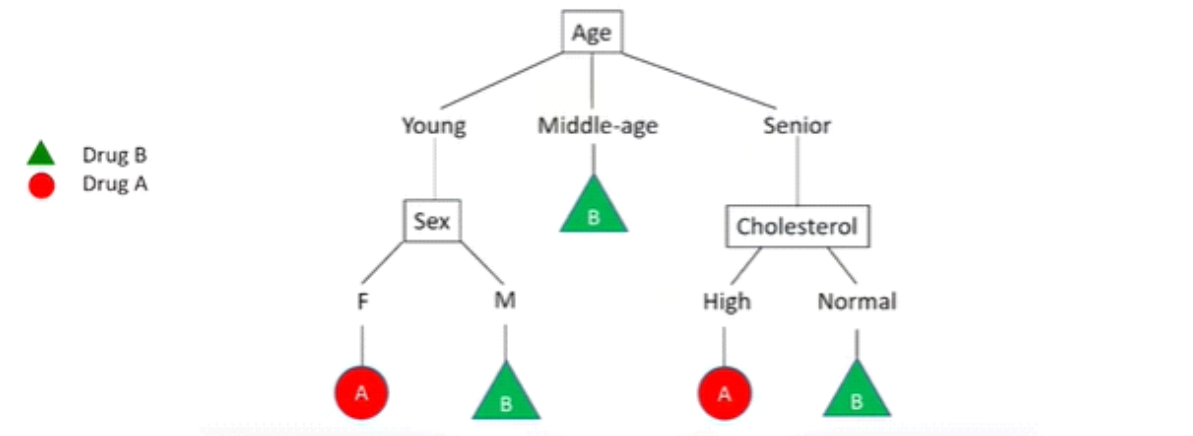
\hat{y} Predicted Probability

$$LogLoss = -\frac{1}{n} \sum (y \times \log(\hat{y}) + (1 - y) \times \log(1 - \hat{y}))$$

屏幕剪辑的捕获时间: 2024/2/5 15:54

Decision tree

Building a decision tree with the training set

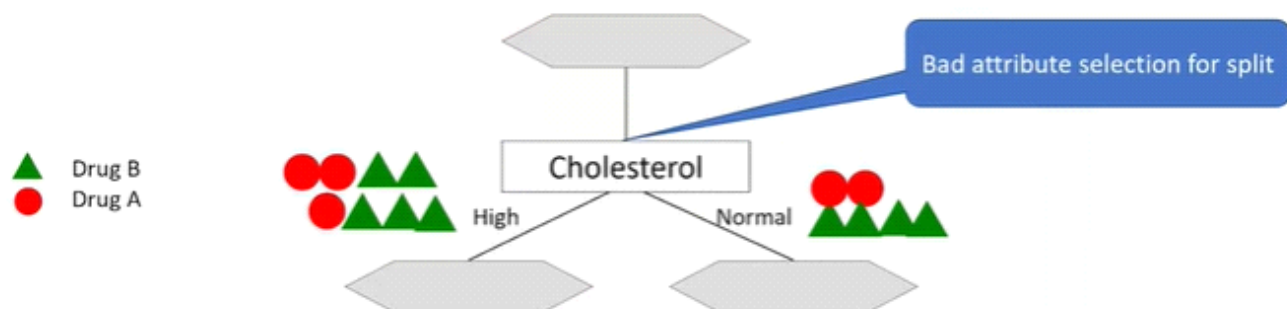


屏幕剪辑的捕获时间: 2024/2/5 15:58

Decision tree learning algorithm

1. Choose an attribute from your dataset

Which attribute is the best ?



屏幕剪辑的捕获时间: 2024/2/5 16:01

Entropy: measure of randomness or uncertainty; the lower the entropy, the less uniform the distribution, the purer the node

2. Calculate the significance of attribute in splitting of data
We should go through all the attributes and calculate the entropy after the split and then choose the best attribute
3. Split data based on the value of the best attribute
Information gain is the information that can increase the level of certainty after splitting
$$\text{Information gain} = \text{entropy before split} - \text{weighted entropy after split}$$
4. Go to step 1

Logistic regression

When is it suitable?

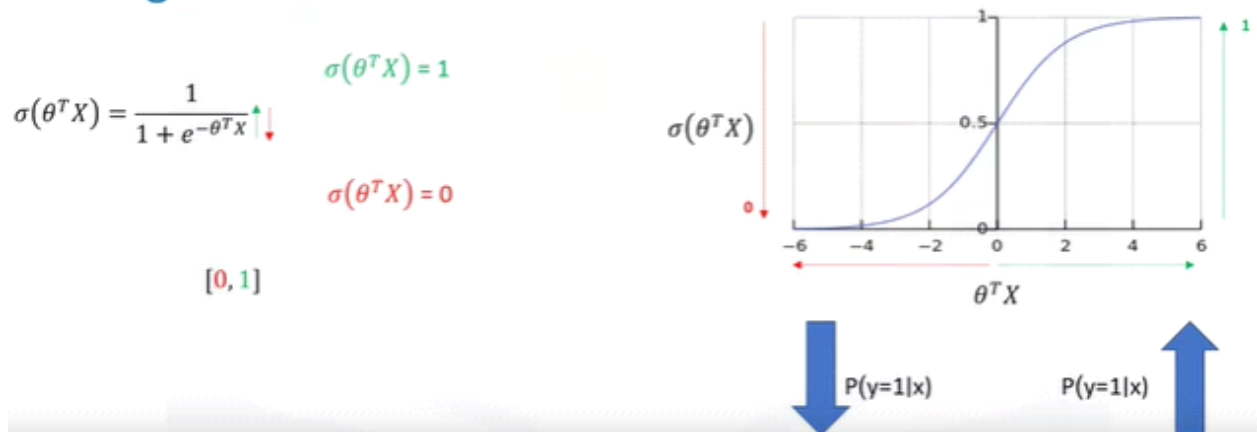
- If your data is binary such as yes/no, true/false
- If you need probabilistic results
- When you need a linear decision boundary

- If you need to understand the impact of a feature

Sigmoid function

Sigmoid function in logistic regression

• Logistic Function



屏幕剪辑的捕获时间: 2024/2/5 17:24

The training process

1. Initialize θ .
2. Calculate $\hat{y} = \sigma(\theta^T X)$ for a customer.
3. Compare the output of \hat{y} with actual output of customer, y , and record it as error.
4. Calculate the error for all customers.
5. Change the θ to reduce the cost.
6. Go back to step 2.

$$\sigma(\theta^T X) \rightarrow P(y=1|x)$$

$$\theta = [-1, 2]$$

$$\hat{y} = \sigma([-1, 2] \times [2, 5]) = 0.7$$

$$\text{Error} = 1 - 0.7 = 0.3$$

$$\text{Cost} = J(\theta)$$

$$\theta_{\text{new}}$$

屏幕剪辑的捕获时间: 2024/2/5 17:28

Logistic Regression Training

Logistic regression cost function

- So, we will replace cost function with:

$$\begin{aligned} \text{Cost}(\hat{y}, y) &= \frac{1}{2} (\sigma(\theta^T X) - y)^2 & \text{Cost}(\hat{y}, y) &= \begin{cases} -\log(\hat{y}) & \text{if } y = 1 \\ -\log(1 - \hat{y}) & \text{if } y = 0 \end{cases} \\ J(\theta) &= \frac{1}{m} \sum_{i=1}^m \text{Cost}(\hat{y}^i, y^i) & J(\theta) &= -\frac{1}{m} \sum_{i=1}^m y^i \log(\hat{y}^i) + (1 - y^i) \log(1 - \hat{y}^i) \end{aligned}$$

屏幕剪辑的捕获时间: 2024/2/5 17:53

How to find the best parameters for our model?

- Minimize the cost function

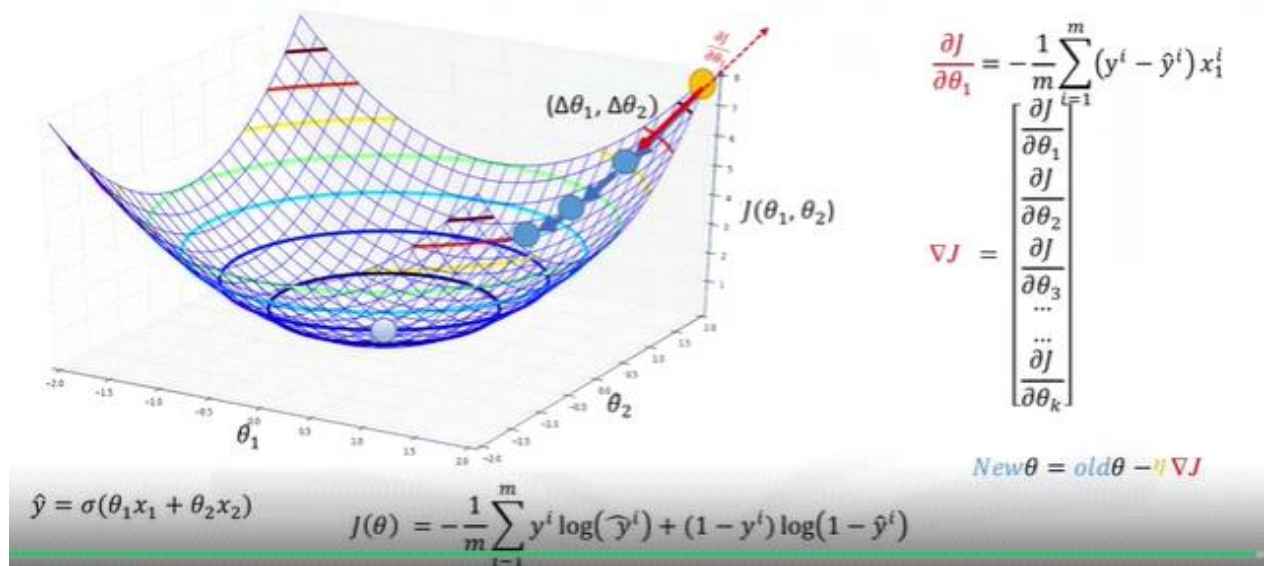
How to minimize the cost function?

- Using Gradient Descent

What is gradient descent?

- A technique to use the derivative of a cost function to change the parameter values, in order to minimize the cost

Using gradient descent to minimize the cost



屏幕剪辑的捕获时间: 2024/2/5 18:01

Support vector machine

SVM is a supervised algorithm that classifies cases by finding a separator

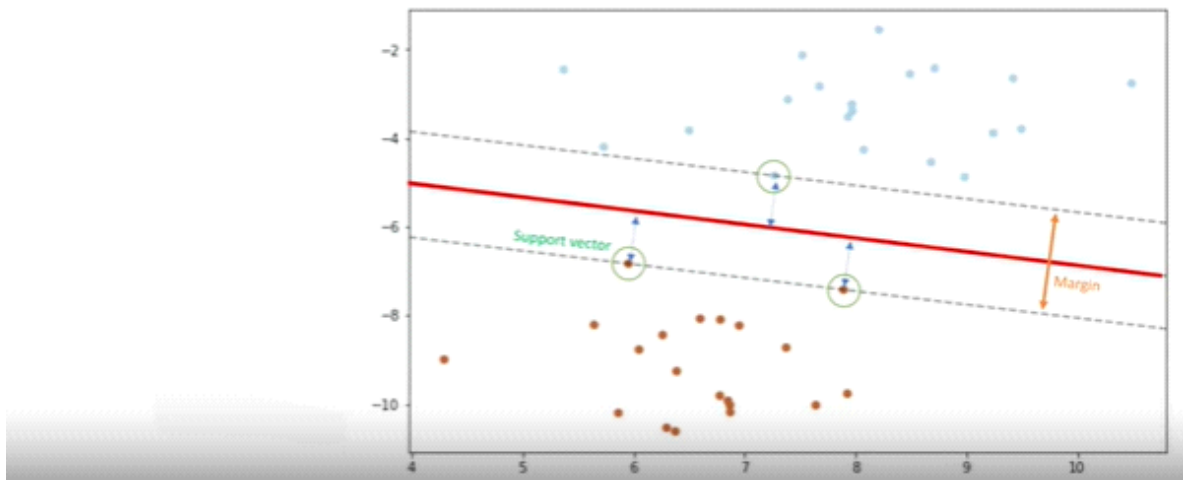
1. Mapping data to a high-dimensional feature space
2. Finding a separator

Data transformation

Kernelling: linear, polynomial, RBF, Sigmoid

Using SVM to find the hyperplane

Using SVM to find the hyperplane



屏幕剪辑的捕获时间: 2024/2/5 18:08

One reasonable choice as the best hyperplane is the one that represents the largest separation or margin between the two classes.

So the goal is to choose a hyperplane with as big a margin as possible.

Examples closest to the hyperplane are support vectors.

It is intuitive that only support vectors matter for achieving our goal.

Pros and cons of SVM

- Advantages:
 - Accurate in high-dimensional spaces
 - Memory efficient
- Disadvantages:
 - Prone to over-fitting
 - No probability estimation
 - Small datasets

屏幕剪辑的捕获时间: 2024/2/5 18:10

SVM applications

- Image recognition
- Text category assignment
- Detecting spam
- Sentiment analysis
- Gene Expression Classification
- Regression, outlier detection and clustering

屏幕剪辑的捕获时间: 2024/2/5 18:11

K-Means clustering

Difference between clustering and classification

Classification is a supervised learning where each training data instance belongs to a particular class, in clustering data is unlabeled and unsupervised

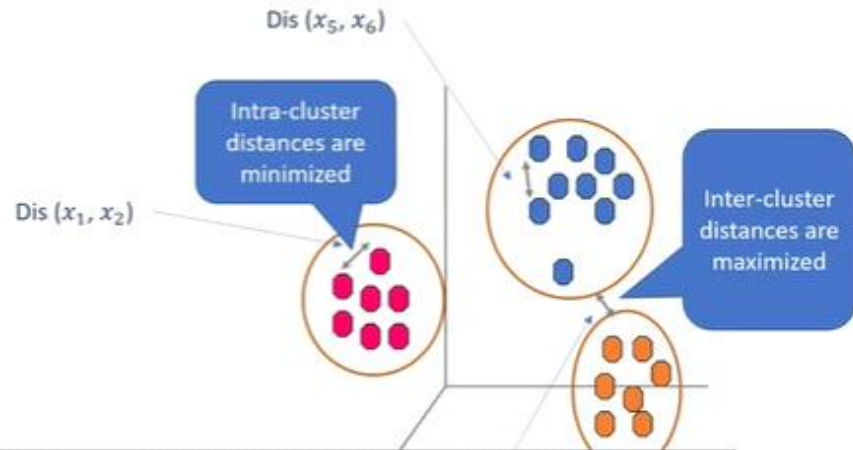
Clustering algorithms

- Partitioned-based clustering
 - Relatively efficient
 - e.g. k-means, k-median, fuzzy c-means
- Hierarchical clustering (intuitive and good for use with small datasets)
 - Produces trees of clusters
 - e.g. agglomerative, divisive
- Density-based clustering
 - Produces arbitrary shaped clusters
 - e.g. DBSCAN

K-Means algorithms

- Partitioned-based clustering
- k-means divides the data into non-overlapping subsets (clusters) without any cluster-internal structure
- Examples within a cluster are very similar, examples across different clusters are very different

Determine the similarity or dissimilarity



屏幕剪辑的捕获时间: 2024/2/6 13:25

1-dimensional similarity/distance



Customer 1

Age

54



Customer 2

Age

50

$$\text{Dis}(x_1, x_2) = \sqrt{\sum_{i=0}^n (x_{1i} - x_{2i})^2}$$

$$\text{Dis}(x_1, x_2) = \sqrt{(54 - 50)^2} = 4$$

屏幕剪辑的捕获时间: 2024/2/6 13:26

Multi-dimensional similarity/distance



Customer 1		
Age	Income	education
54	190	3



Customer 2		
Age	Income	education
50	200	8

$$\text{Dis}(x_1, x_2) = \sqrt{\sum_{i=0}^n (x_{1i} - x_{2i})^2}$$
$$= \sqrt{(54 - 50)^2 + (190 - 200)^2 + (3 - 8)^2} = 11.87$$

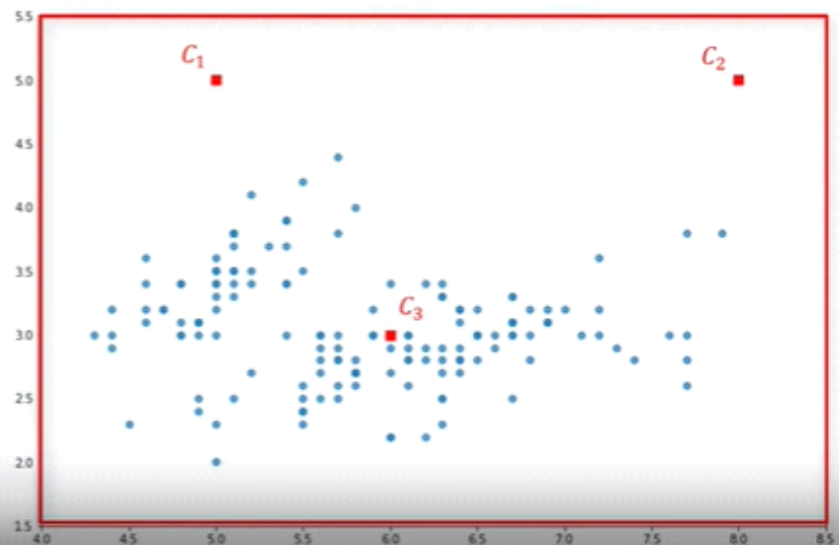
屏幕剪辑的捕获时间: 2024/2/6 13:26

k=3, which means there are 3 clusters

k-Means clustering – initialize k

1) Initialize **k=3**
centroids randomly

$$\begin{aligned} C_1 &= [8., 5.] \\ C_2 &= [5., 5.] \\ C_3 &= [6., 3.] \end{aligned}$$

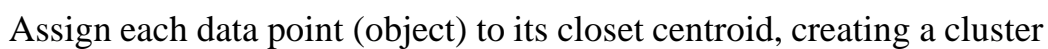


Randomly placing k centroids, one for each cluster

2) Distance calculation



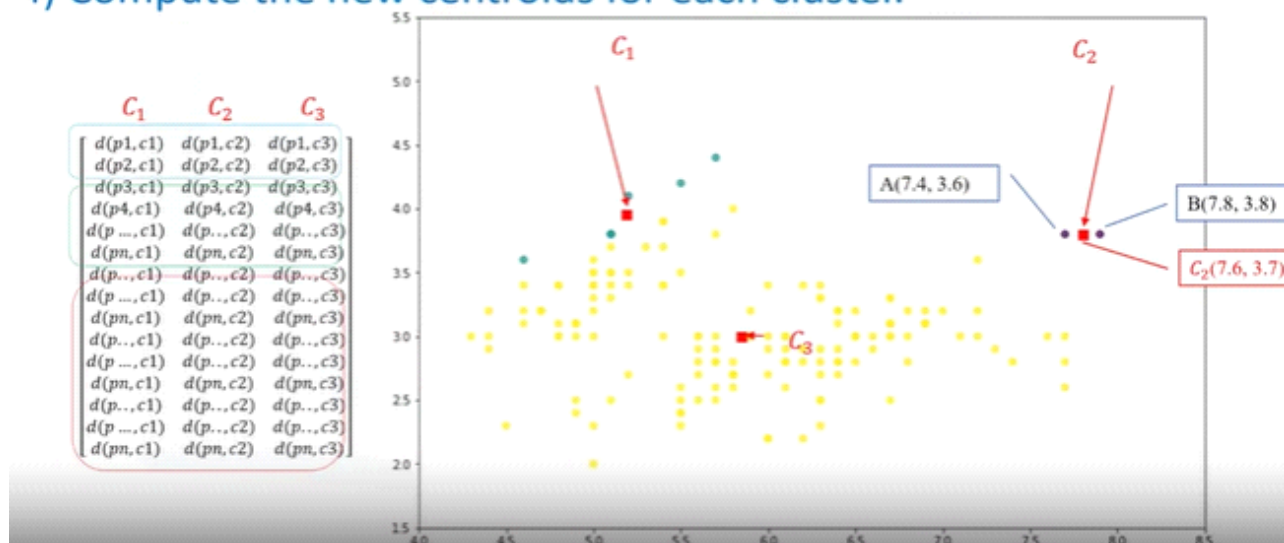
3) Assign each point to the closest centroid



3) Assign each point to the closest centroid



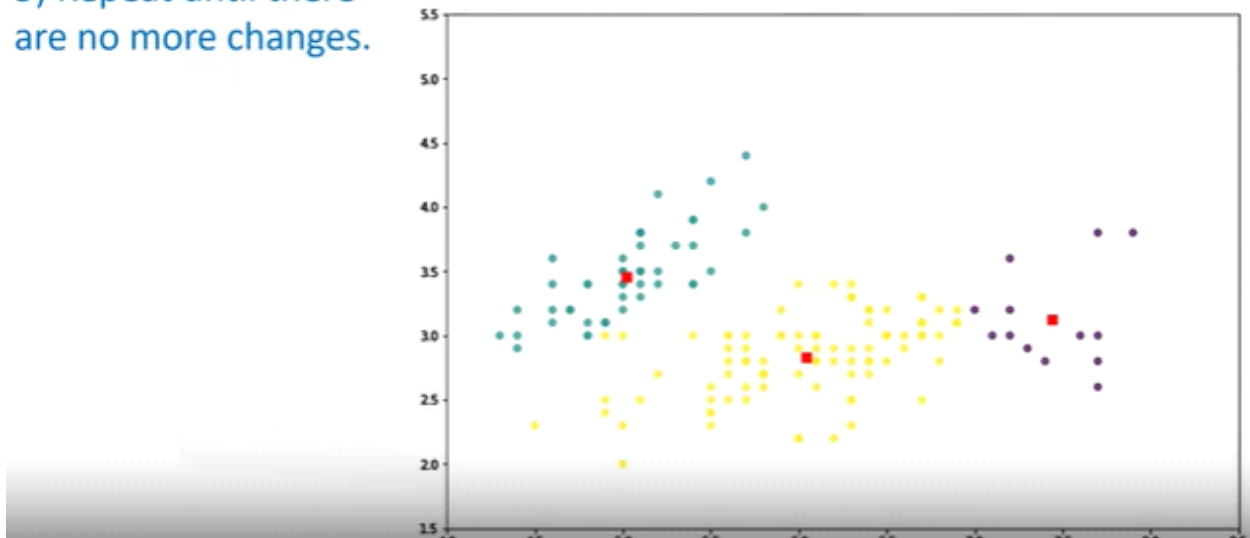
4) Compute the new centroids for each cluster.



Recalculate the position of the k centroids

k-Means clustering – repeat

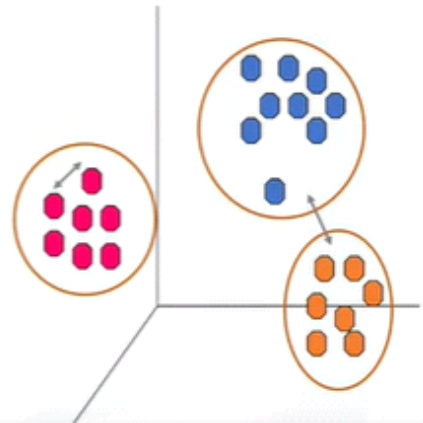
5) Repeat until there are no more changes.



Repeat the steps 2-4, until the centroids no longer move

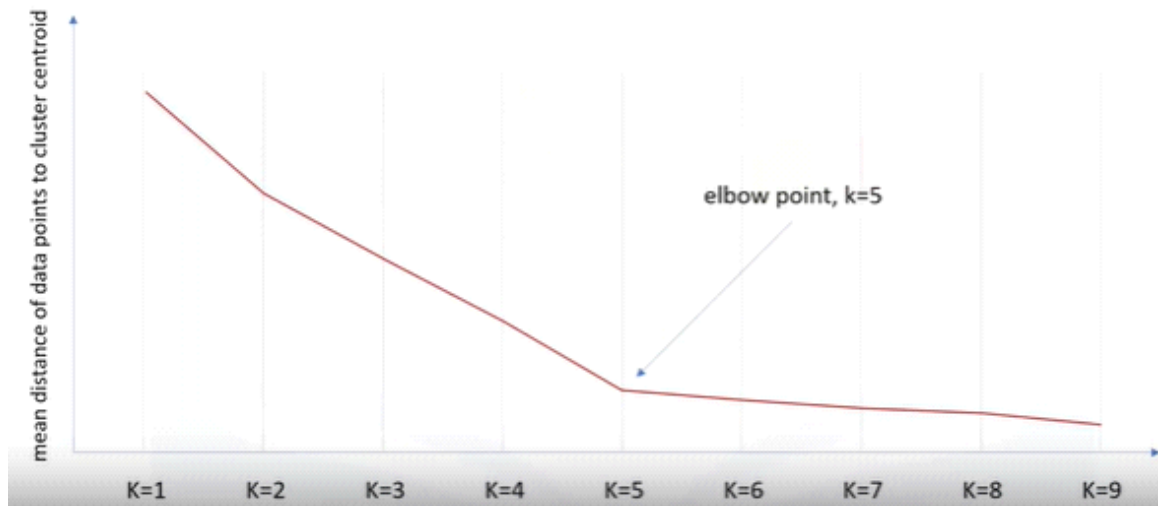
k-Means accuracy

- External approach
 - Compare the clusters with the ground truth, if it is available.
- Internal approach
 - Average the distance between data points within a cluster.



屏幕剪辑的捕获时间: 2024/2/6 13:37

Choosing k



With increasing the number of clusters, the distance of centroids to data points will always reduce, this means increasing k will always decrease the error