Machine Learning with Python

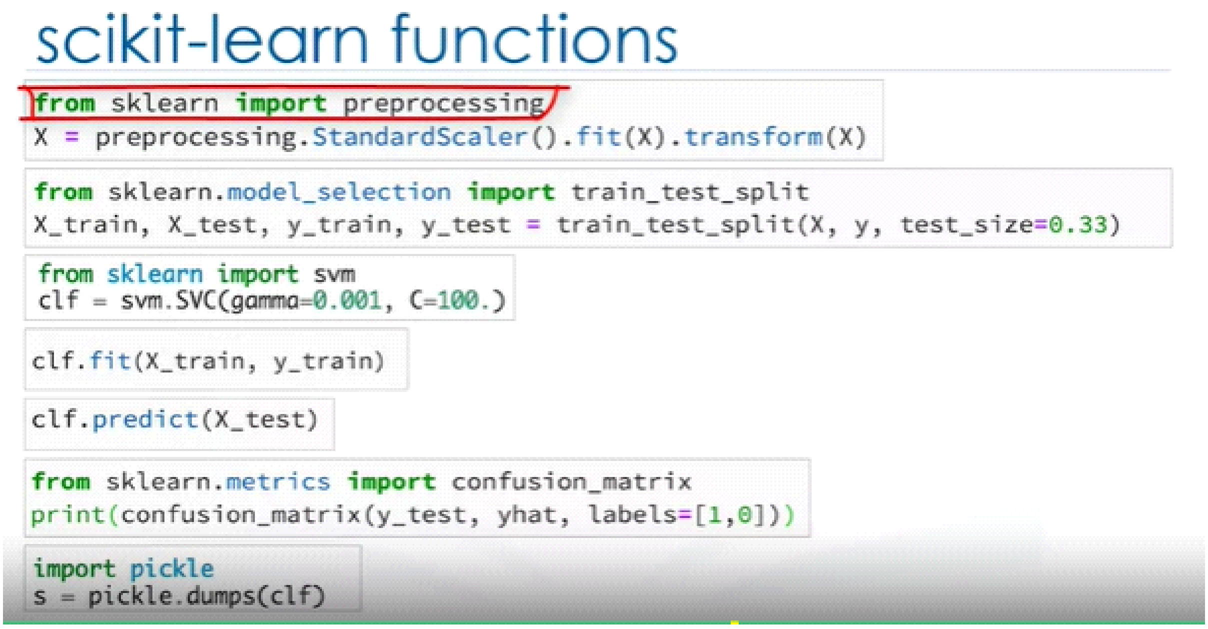
2024年2月5日 12:42

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



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

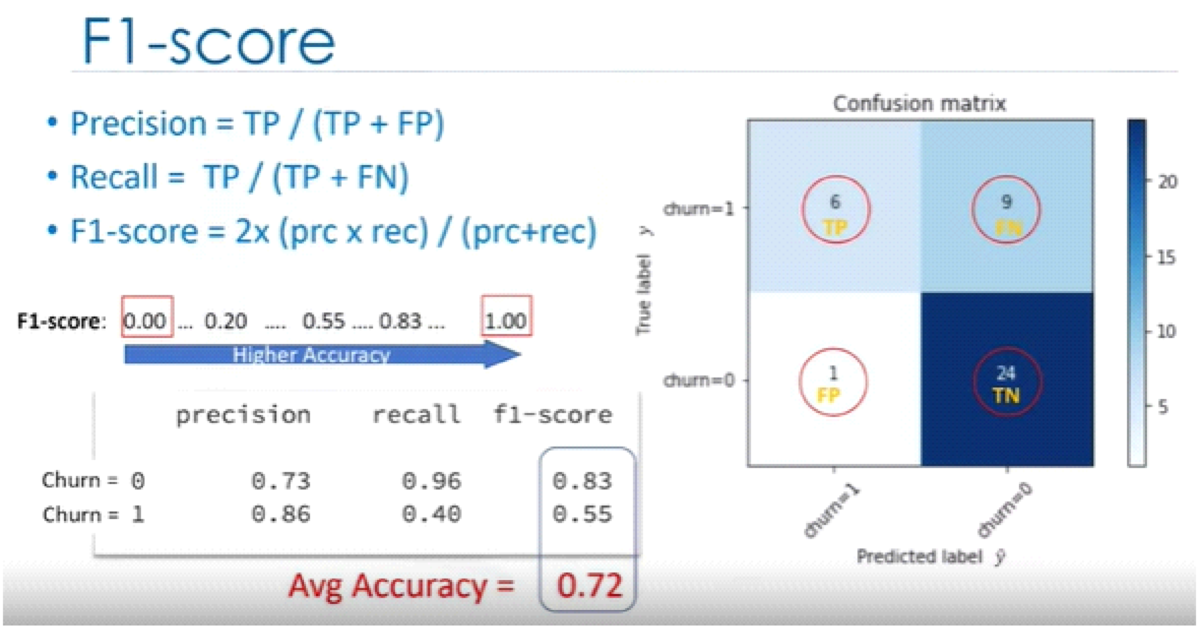
1. Pick a value for K
2. Calculate the distance of unknown case from all cases
   - Use Euclidean distance: Dis (x1, x2) = the root of (x1-x2)^2
3. Select the K-observations in the training data that are "nearest" to the unknown data point
   - How can we find the best value for K?

     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

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

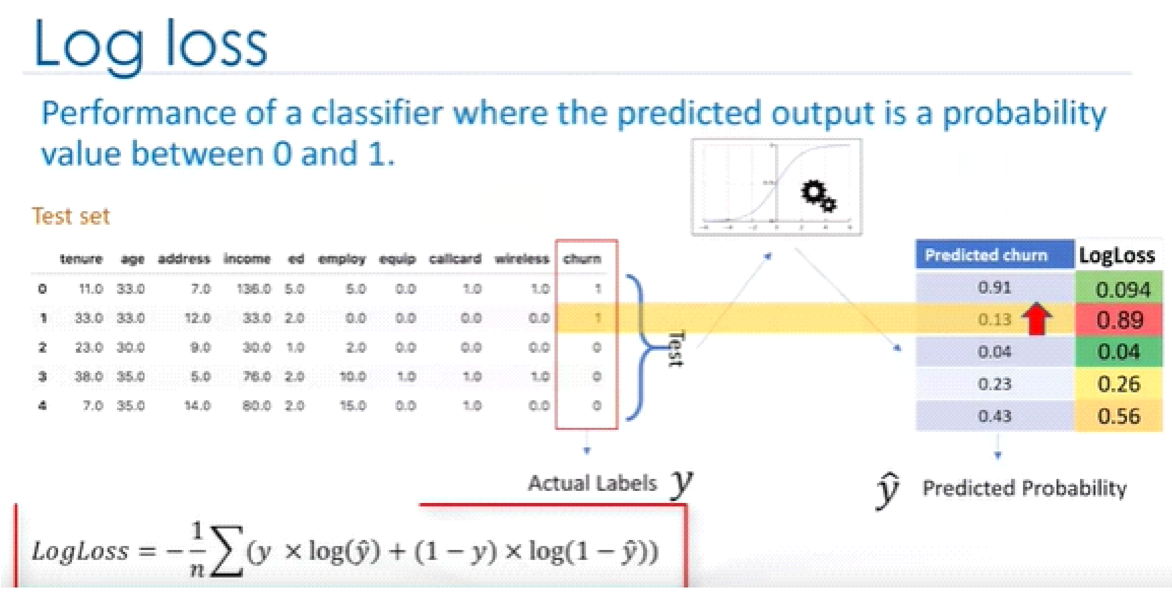
Evaluation metrics in classification

- 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



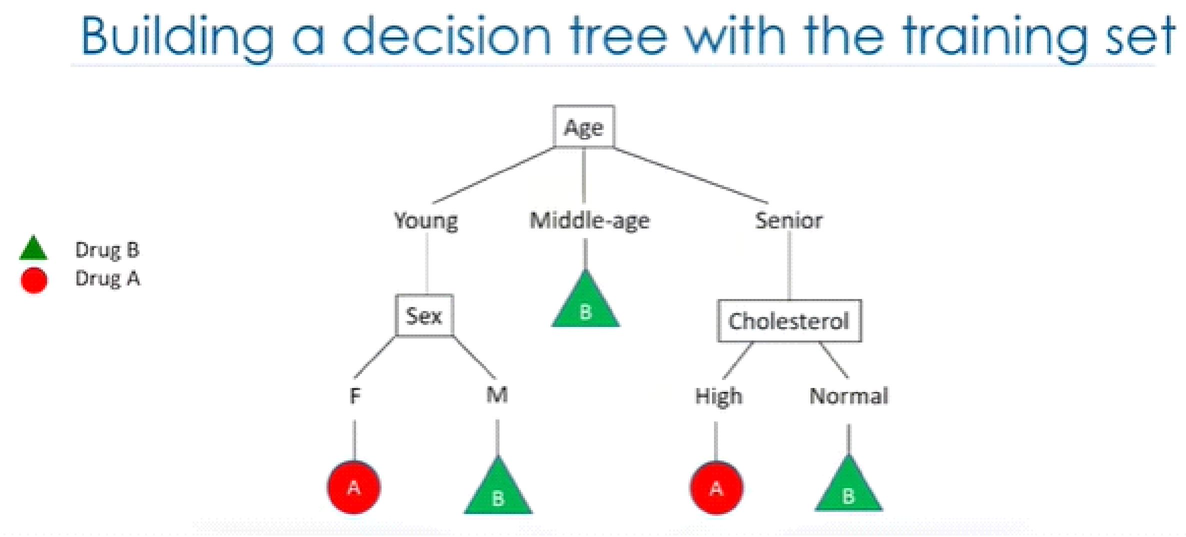
屏幕剪辑的捕获时间: 2024/2/5 15:52

- Log loss
  Measures how far the prediction is from the actual label



屏幕剪辑的捕获时间: 2024/2/5 15:54

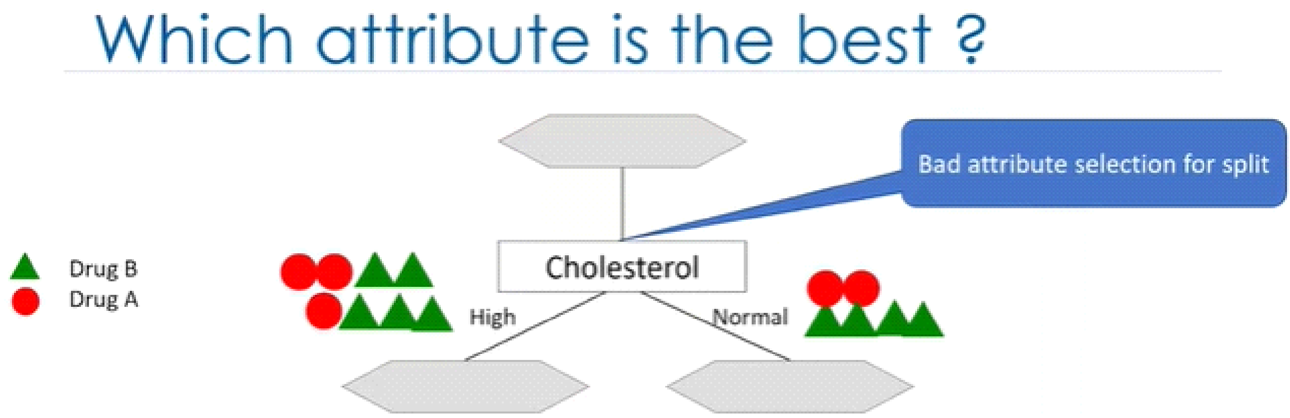
Decision tree



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



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

   Information gain = entropy before split - weighted entropy after split

4. Go to step 1

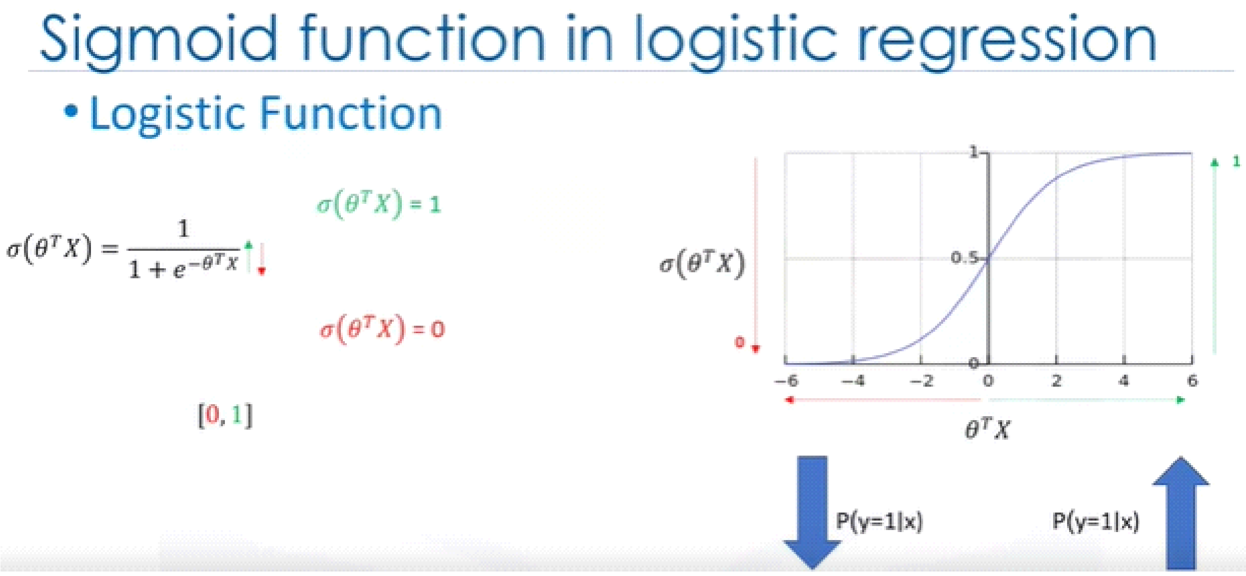
Logistic regression

When is it suitable?

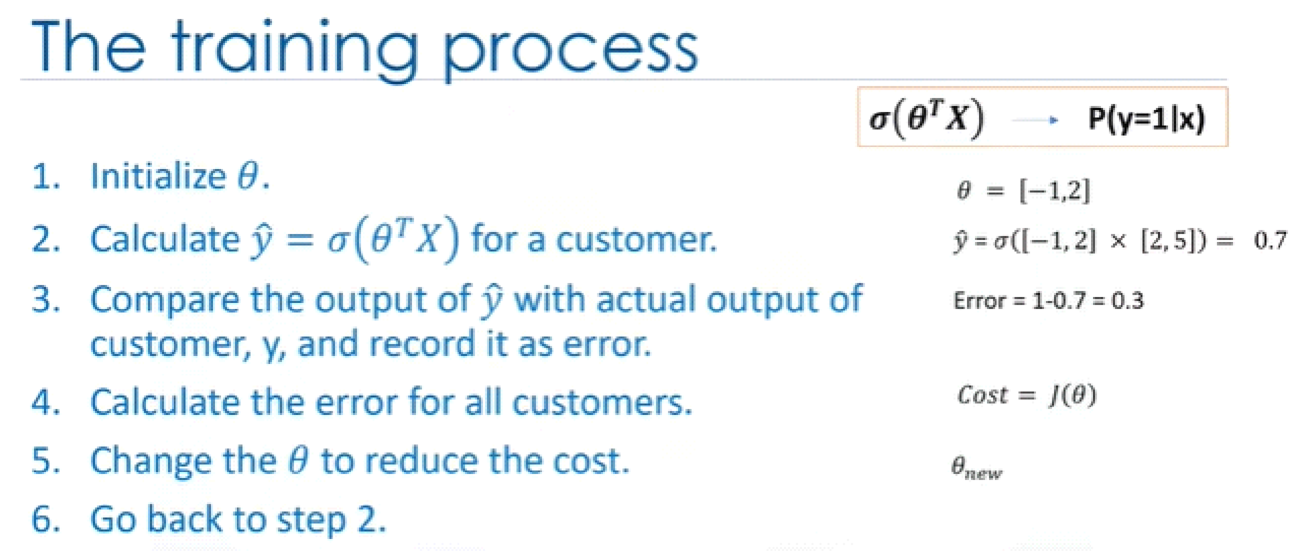
- If your data is binary such as yes/no, true/false
- If you need probabilistic results
- When you need a linear decision boundary

- If you need to understand the impact of a feature

Sigmoid function



屏幕剪辑的捕获时间: 2024/2/5 17:24



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:

$$Cost(\hat{y}, y) = \frac{1}{2}\left(\sigma(\theta^T X) - y\right)^2 \qquad Cost(\hat{y}, y) = \begin{cases} -\log(\hat{y}) & if\ y = 1 \\ -\log(1-\hat{y}) & if\ y = 0 \end{cases}$$

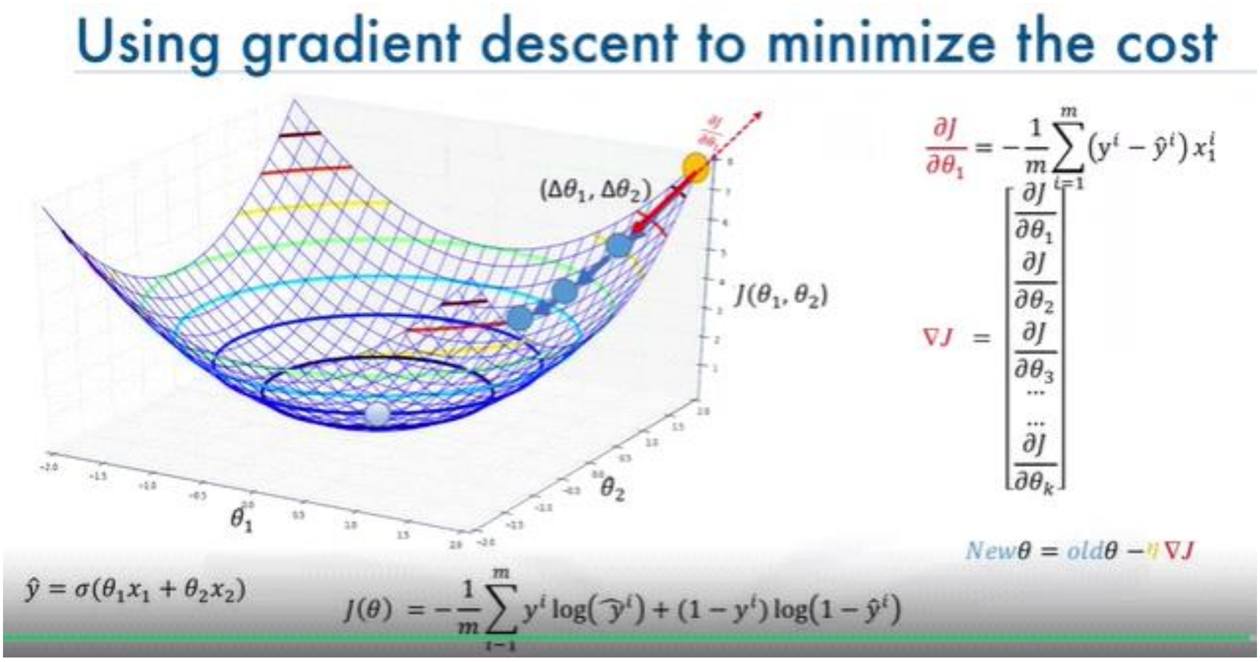
$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} Cost(\hat{y}, y) \qquad J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} y^i \log(\hat{y}^i) + (1-y^i)\log(1-\hat{y}^i)$$

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



屏幕剪辑的捕获时间: 2024/2/5 18:01

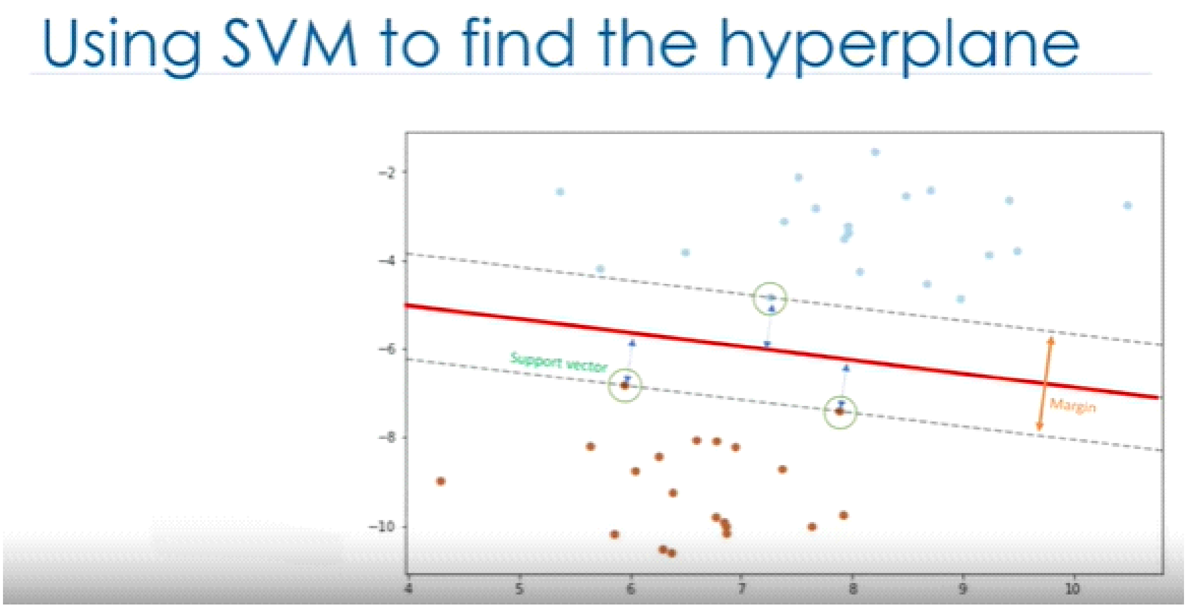
Support vector machine

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



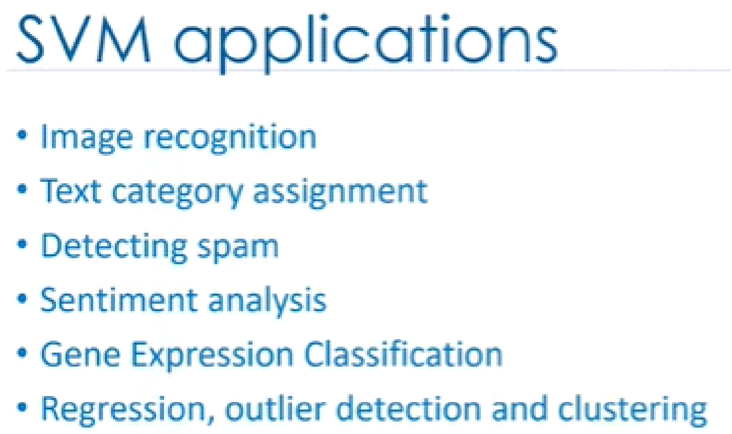
屏幕剪辑的捕获时间: 2024/2/5 18:08

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



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

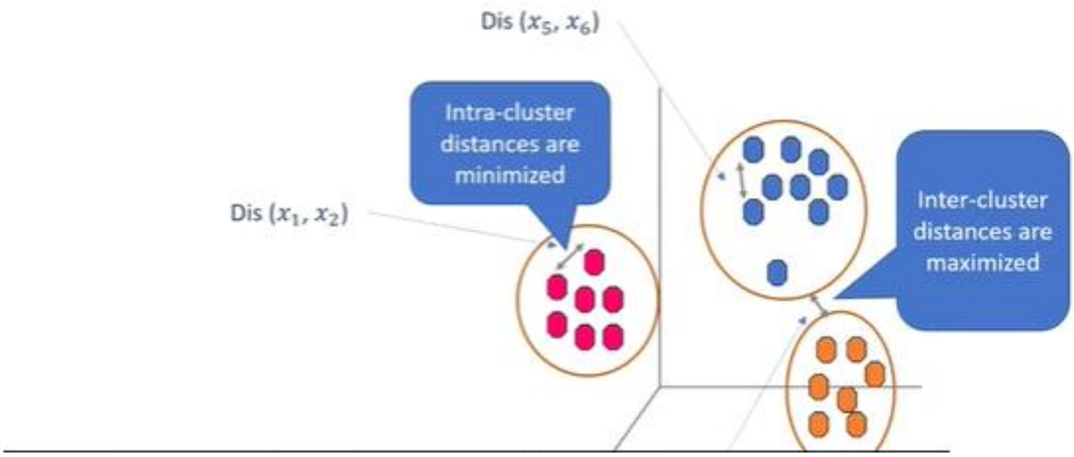
Clustering algorithms

- 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

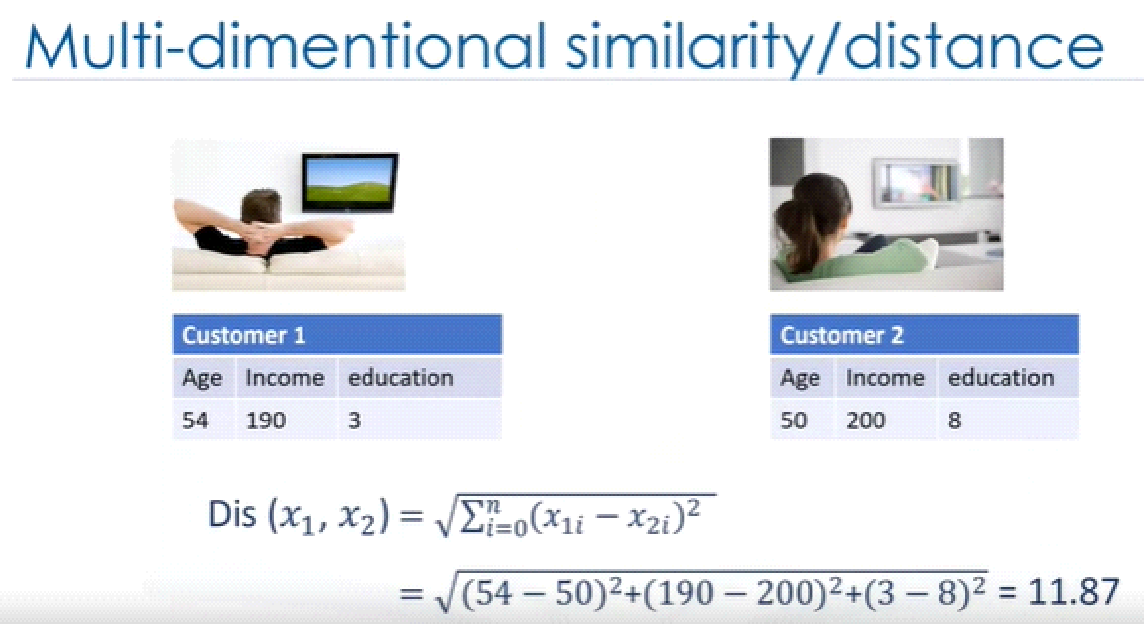
Customer 1
Age
54

Customer 2
Age
50

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

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

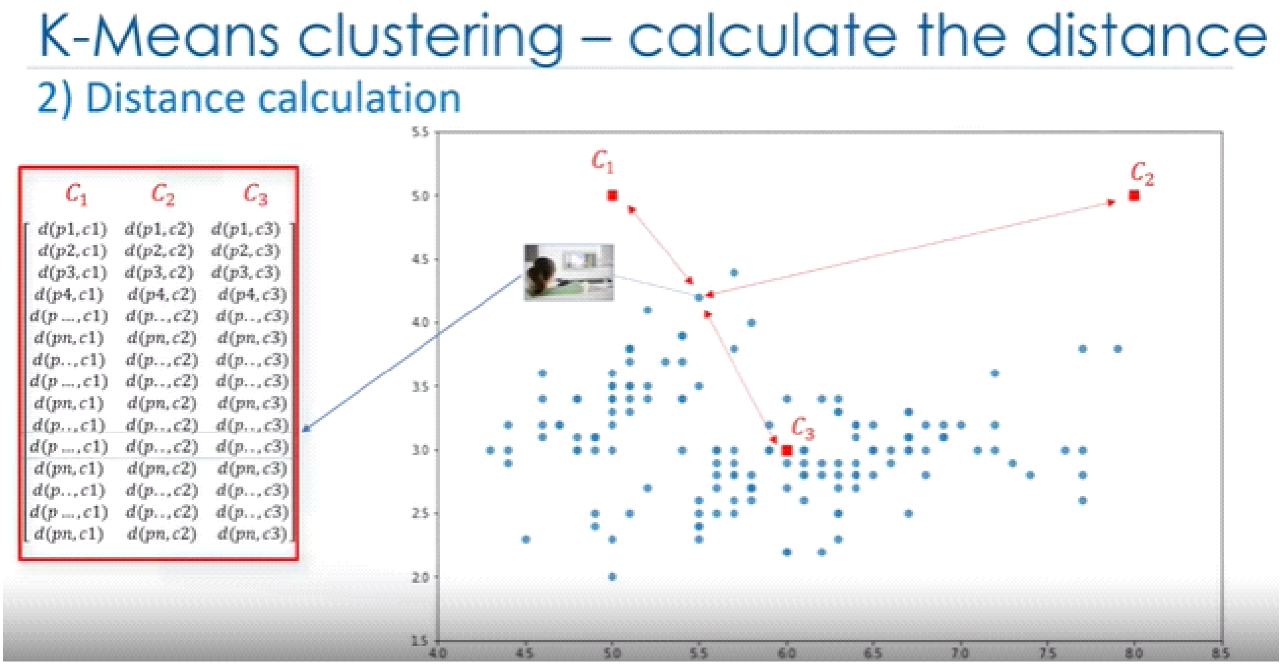
屏幕剪辑的捕获时间: 2024/2/6 13:26



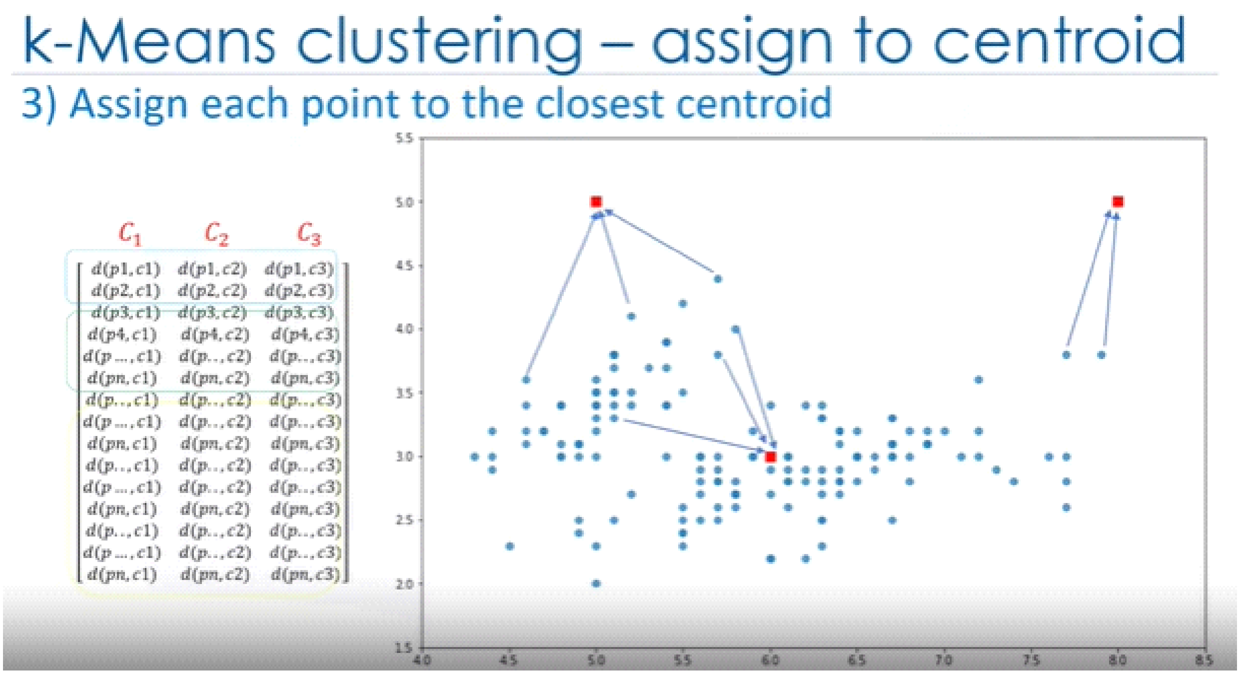
屏幕剪辑的捕获时间: 2024/2/6 13:26

k=3, which means there are 3 clusters

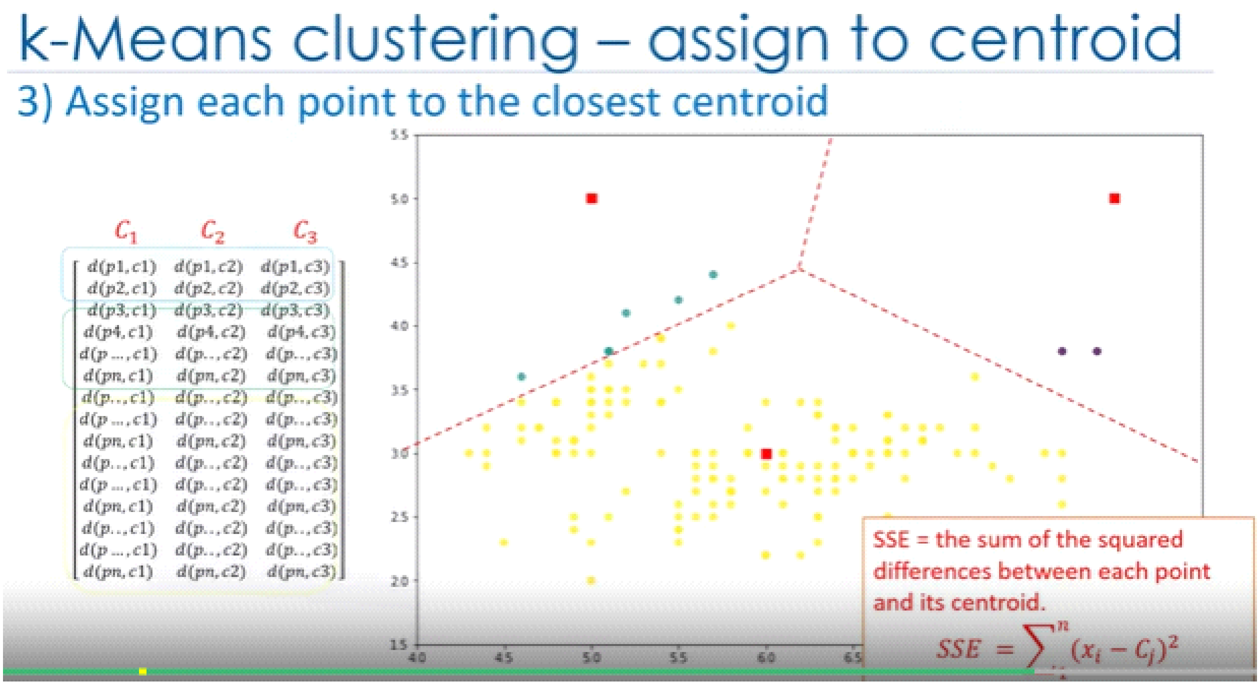
Randomly placing k centroids, one for each cluster



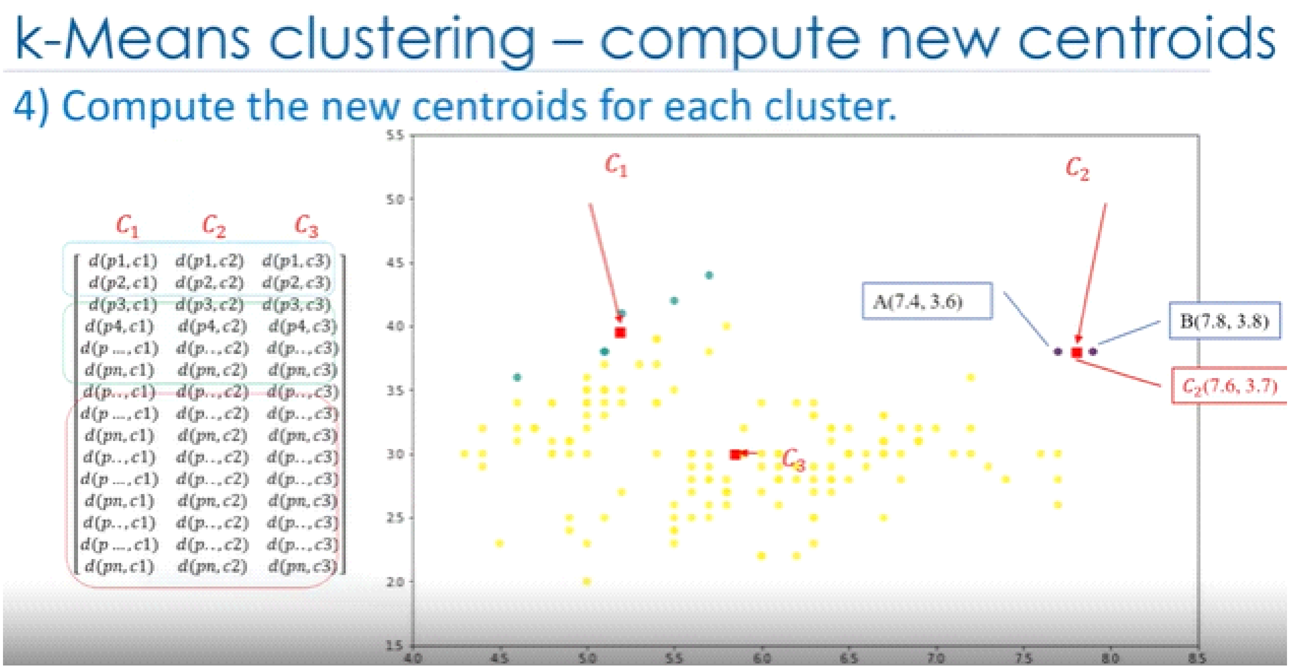
Calculate the distance of each point from each centroid



Assign each data point (object) to its closet centroid, creating a cluster

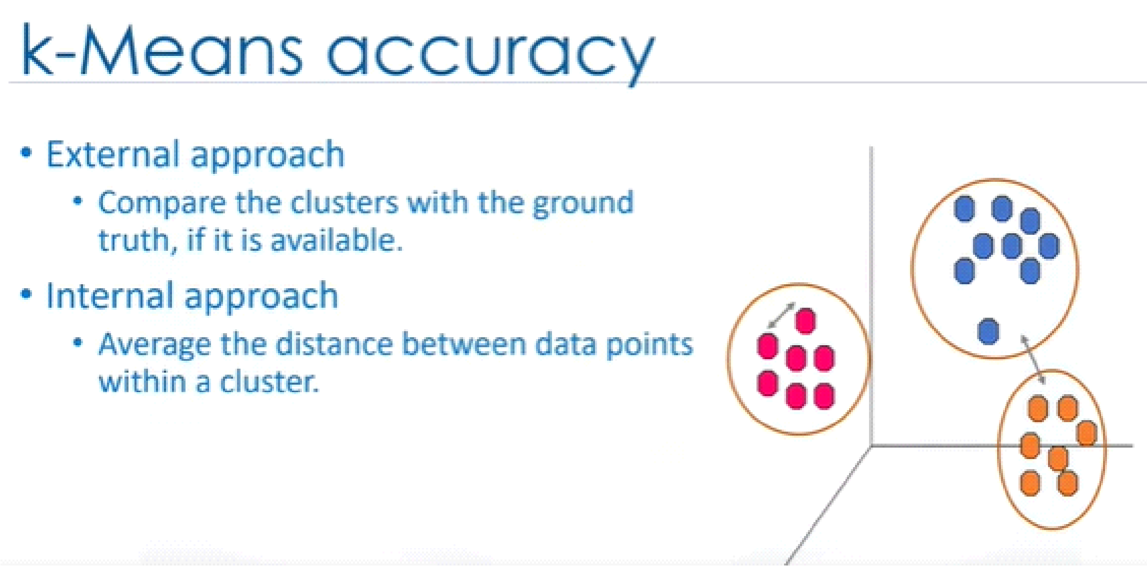


屏幕剪辑的捕获时间: 2024/2/6 13:31



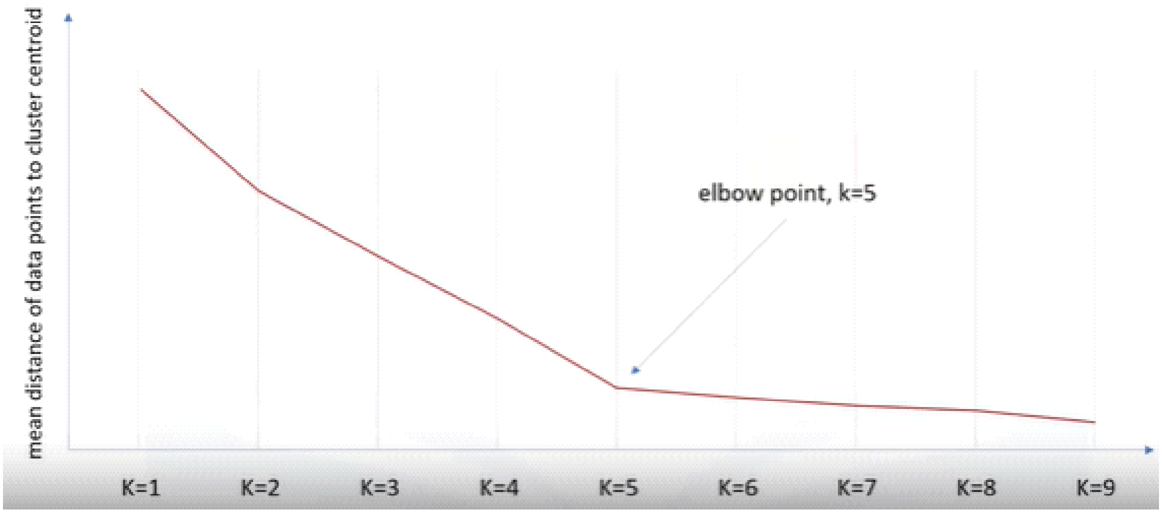
Recalculate the position of the k centroids

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



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