**Machine learning review**:

**Chapter 2:**

1. Steps to train a single neuron:
2. Initialize the weights to 0 or small number.
3. For each training sample x(i) perform the following steps:
   1. Compute the output value y^.
   2. Update the weights.
4. Difference between Adaline rule and perceptron:

Weights in Adaline rule are updated based on a **linear activation function** instead of a **unit step function** ( )

1. Adaline rule is always accompanied with “batch gradient descent” method in **small scale**.

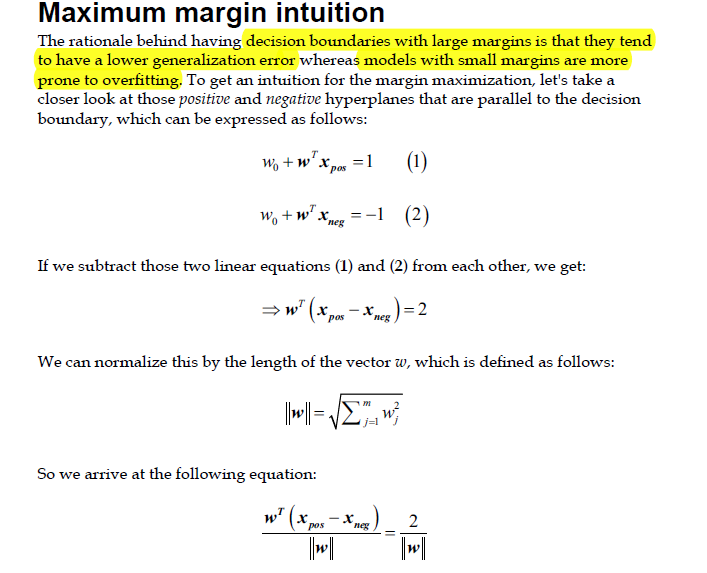
When the scale of data size becomes large, we should use **stochastic gradient descent** or **mini-batch gradient descent.**

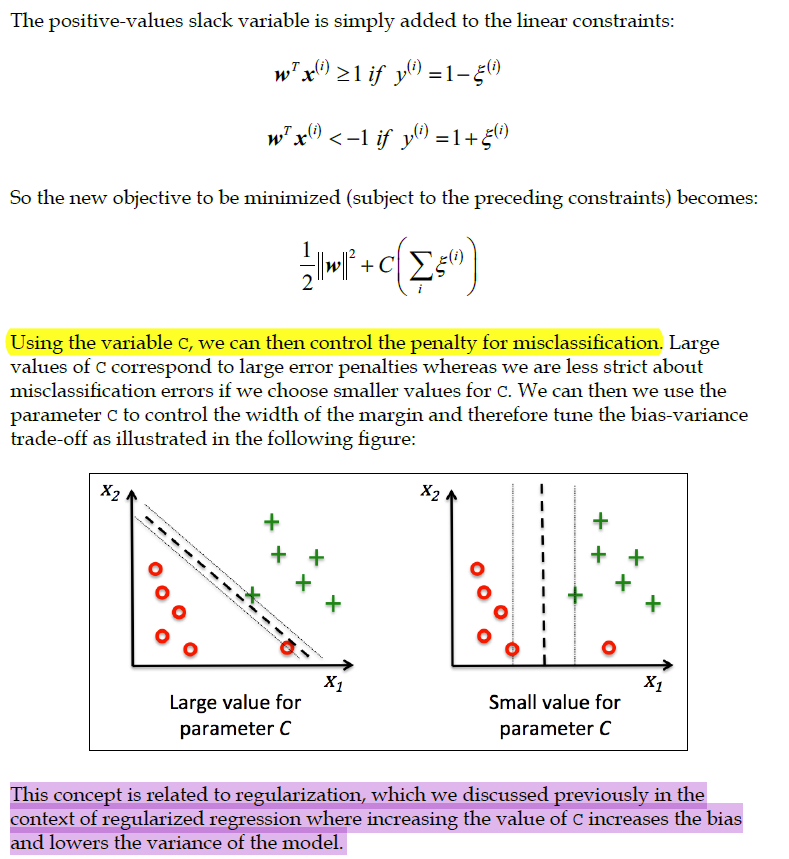
**Chapter 3:**

* + 1. **支持向量机soft margin:**

Soft margin旨在maximize 然而转换为minimize 更便于计算。

当C->0时，边界变宽，限制变小。 反之边界变窄，趋近于hard margin。

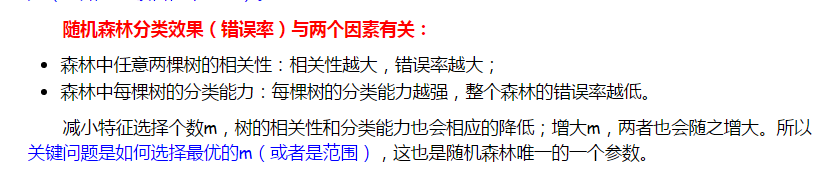
****



* + 1. **Random forest：**

**两个Random的含义**：

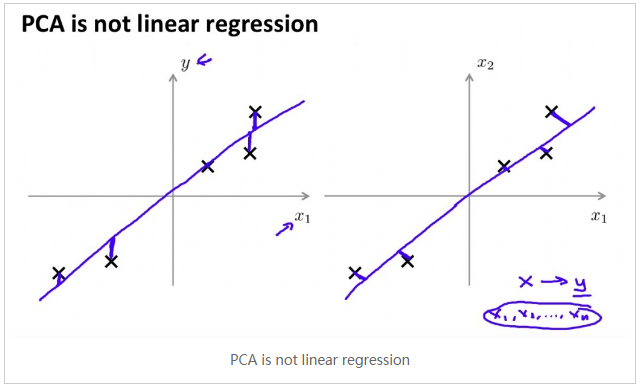
1. 如果有N个训练集，那么我们每次从中随机有放回的选择一个训练集训练一个决策树。 训练N次。  
   2.训练一棵决策树时，我们在每个节点分裂时随机从全部M个特征中，选择m(<<<<M)个特征。然后从m个特征中选择一个最优特征进行分裂。



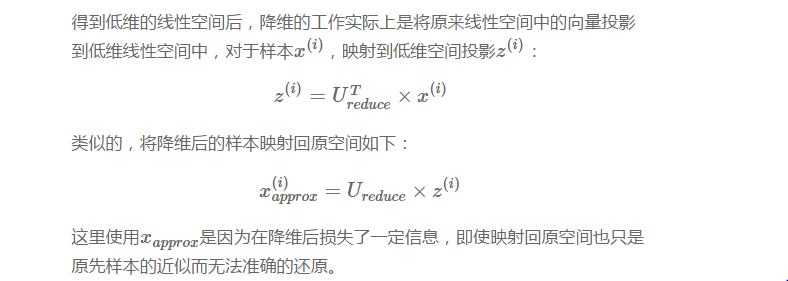
**3. PCA, LDA and kernel PCA:**

## 1. 主成分分析（PCA）

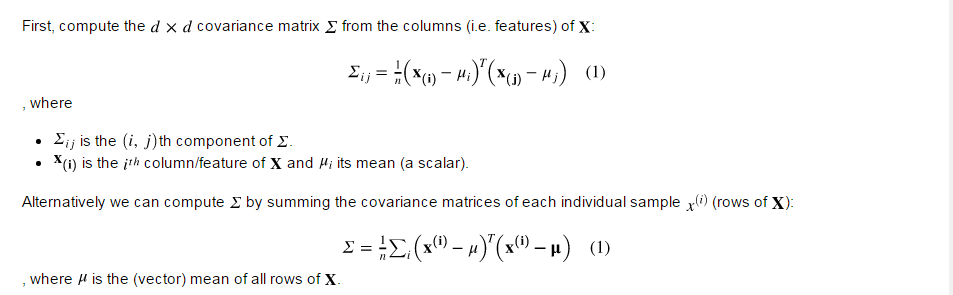
PCA是一种常见的数据降维算法，其核心思想非常的简单：寻找一个低维的超平面，使所有样本和其在该低维超平面上的投影间的距离和最小，如下图：



首先，做数据预处理：normalize，或者 standard scaler。

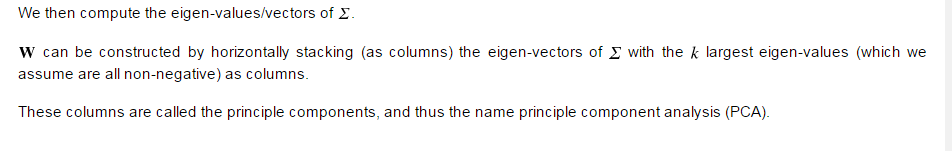


步骤：1.首先计算出协方差矩阵 ∑ d\*d：



2.然后将协方差矩阵奇异分解，前K个特征向量进行堆积形成的

矩阵就是所要求的转换矩阵W(d\*k)

****

**未完待续**

1. **LDA :**

**与PCA差别：（1）LDA是supervised algorithm，需要class information.**

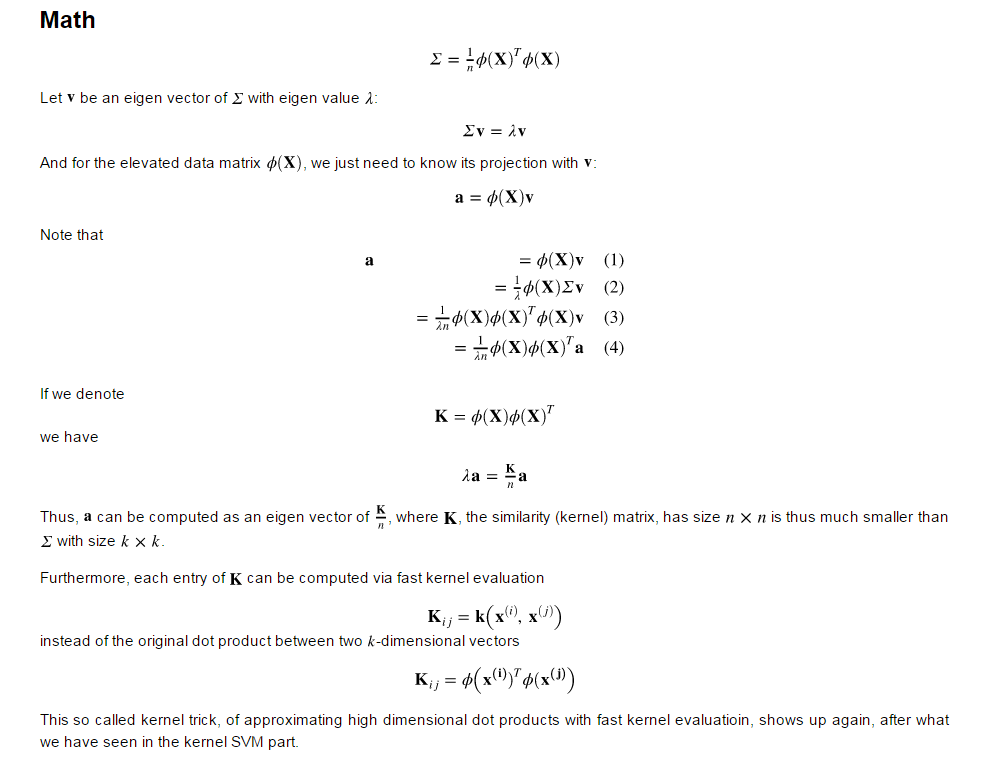
**（2）LDA将数据投影到（a）最大类间离散度（b）最小类内离散度(variance)。**

**PCA投影到整体最大离散度（variance）.**

1. **Kernel-PCA**

**将不线性可分的数据先投影到高维，然后在利用PCA进行降维，从而使得可以使用linear classifier。在数据不线性可分时使用。**

**考点：chapter 5 kernel math**

****

## Question

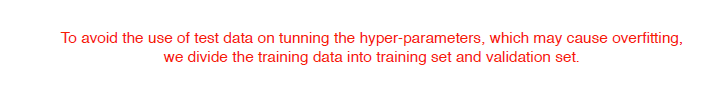
How should we choose between the following two?

1. Linear PCA  kernel SVM
2. kernel PCA  linear SVM

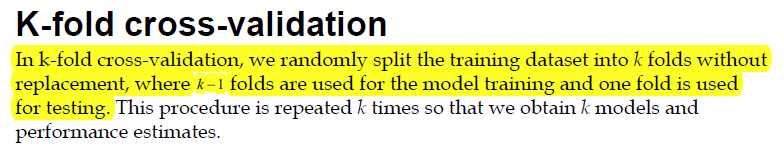
Answer: If the data contains lots of duplicated features and not linearly seperable, we choose to use kernel PCA to project the data into higher dimensions and then apply pca to find the most important features i.e. the principle components to do classification. If there are not many duplicated features and it’s not linearly seperable, we use kernel SVM directly.

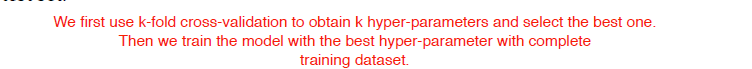
1. **Cross validation and Performance evaluation metrics:**

**1)holdout method:**

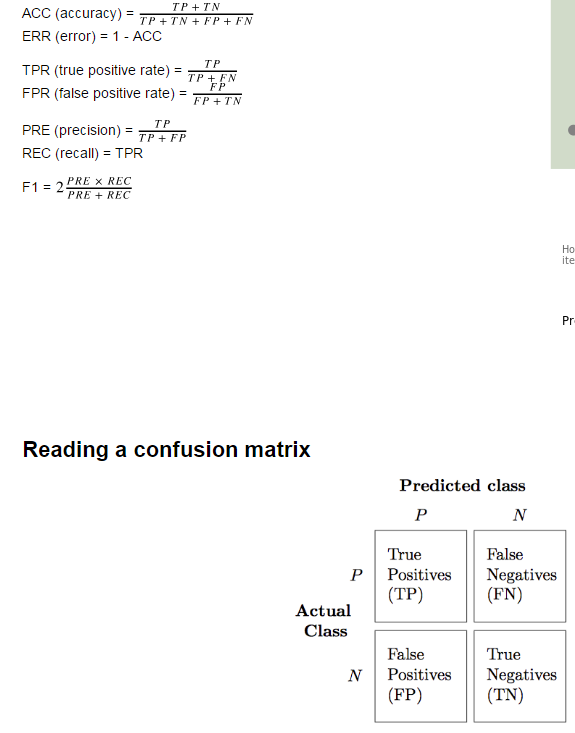
****

**2)K-fold cross-validation:**

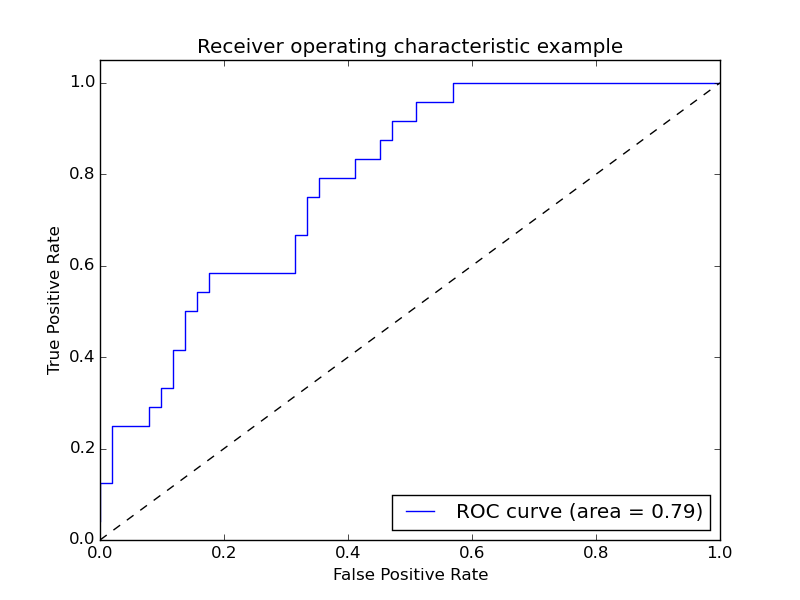
****

****

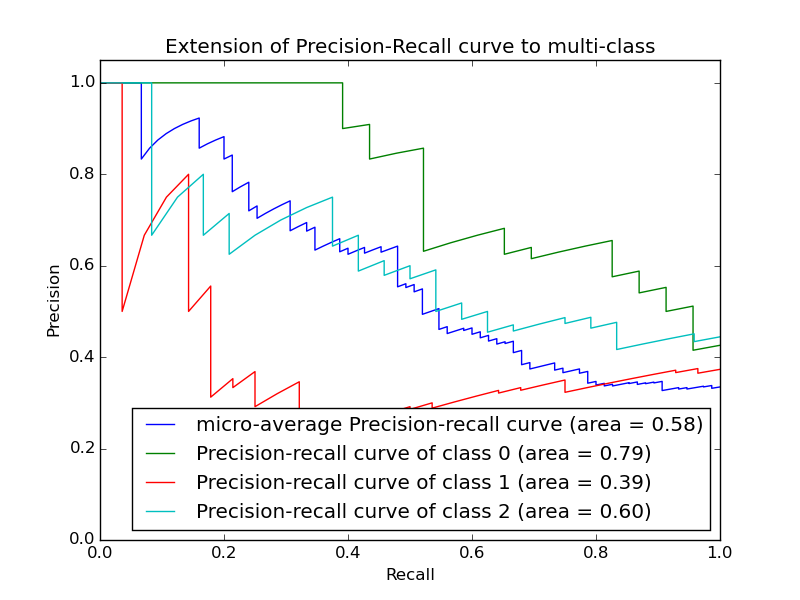
**3) Performance evaluation metrics**

****

* 1. **如果ROC是光滑的，那么说明没有太多overfitting。并且ROC图形围成的面积越大，模型越好。**

****

* 1. **Precision recall curve:曲线越平滑，overfitting越小。同一个测试集上，上面的线比下面的线好一些（面积越大）。**

****

1. **Ensemble learning:**
2. **Achieve diversification:**

**(1) Different hyper-parameters for the same algorithm/model**

**(2) Different input representations of the same event**

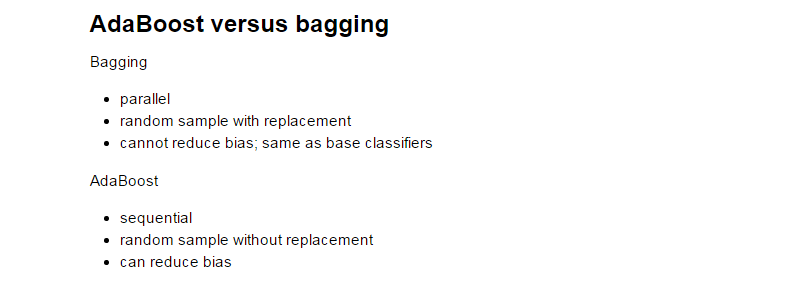
**(3)** **Different training sets(e.g. bagging and boosting).**

**2. How to combine multiple classifiers:**

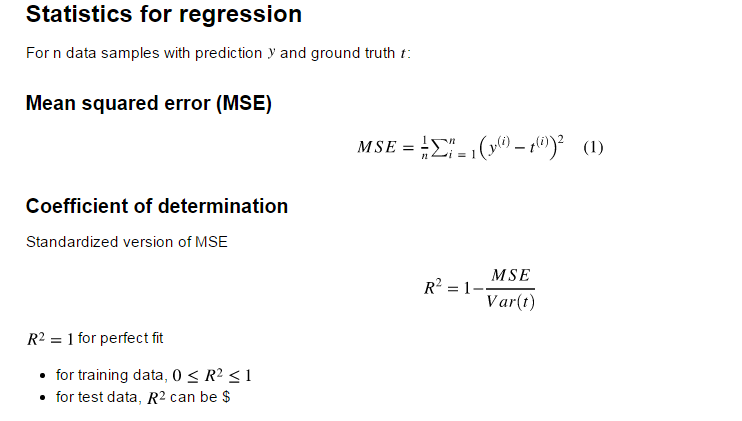
**(1)Parallel: multi-expert combination**

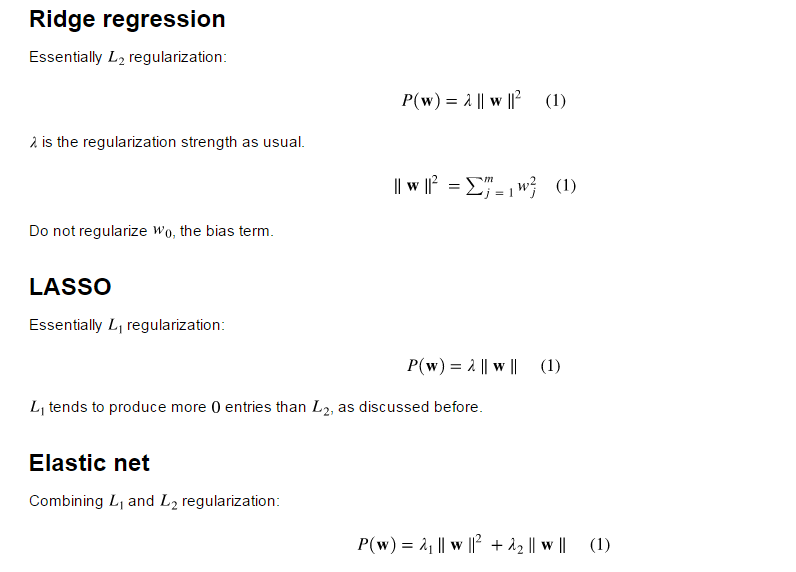
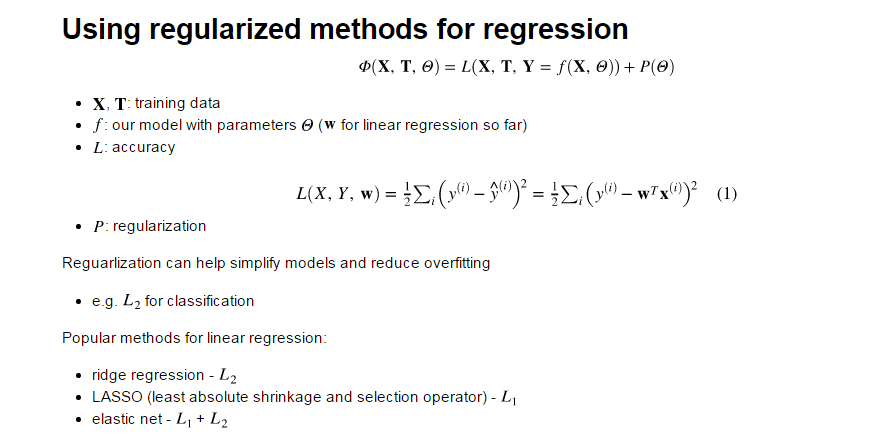
**(2) Sequential: multi-stage combination**

**3.Comparison between AdaBoost and bagging**

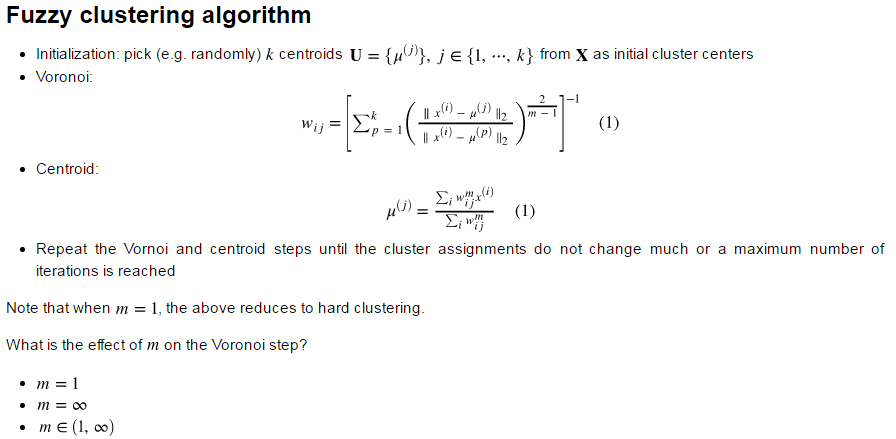
****

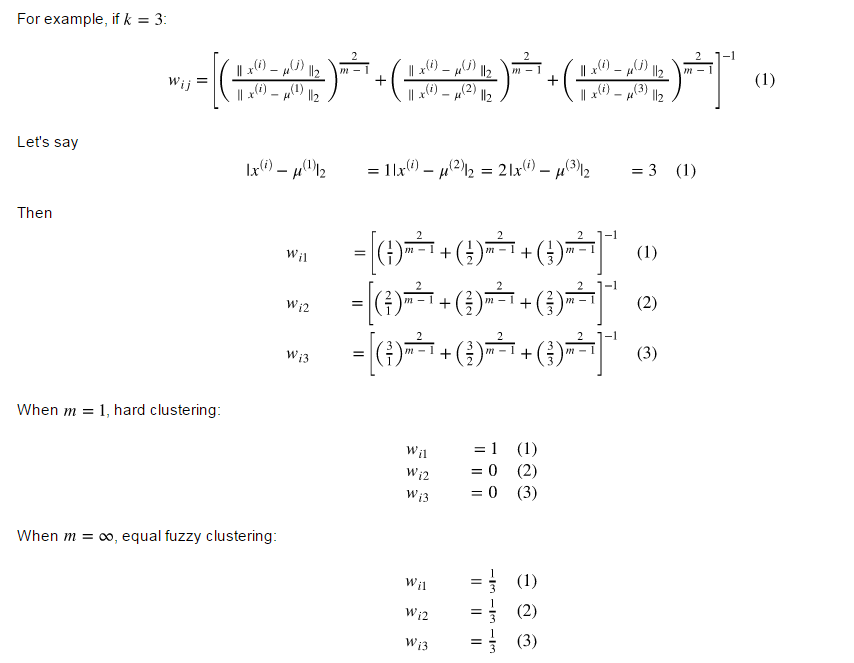
1. **Regression Analysis:**
2. **Find the least square parameters for a linear regression model:**

****

****

1. **Regression 的输出是连续值而分类输出的数值是离散值，用于分类。**
2. **Clustering；**
3. **Unsupervised learning algorithm, no class information is given.**
4. **Use distance measures: d(xi,xj) to decide the class classification. L2(Euclidean distance ) is often used.**
5. **Analysis of fuzzy clustering:**

****

****

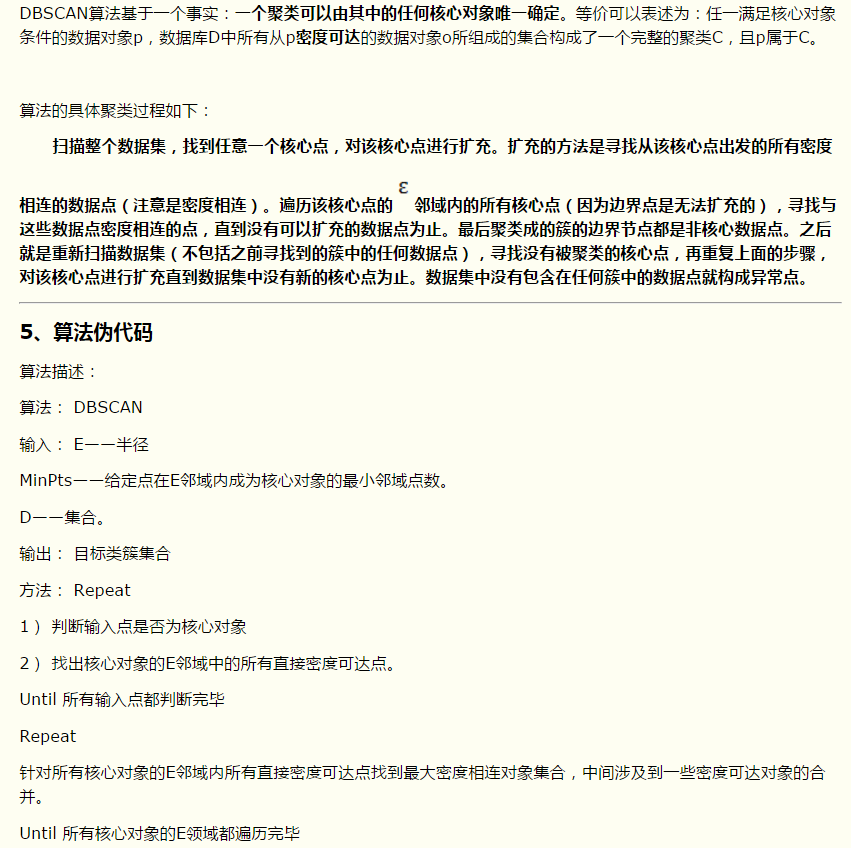
1. **Silhouette for clustering；**

**Compare the mean distances of a example to all elements in another clusters and the mean distance of the example to all other examples(except itself) in its own cluster.**

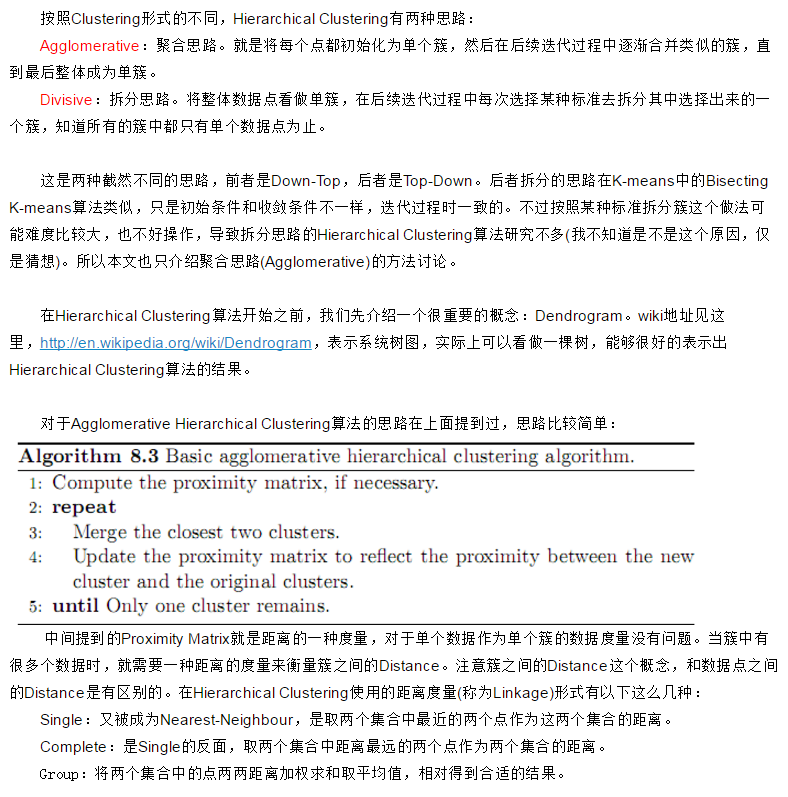
**Remember the formula is . -1<S(i)<1 and S(i) is close to 1🡪good clustering**

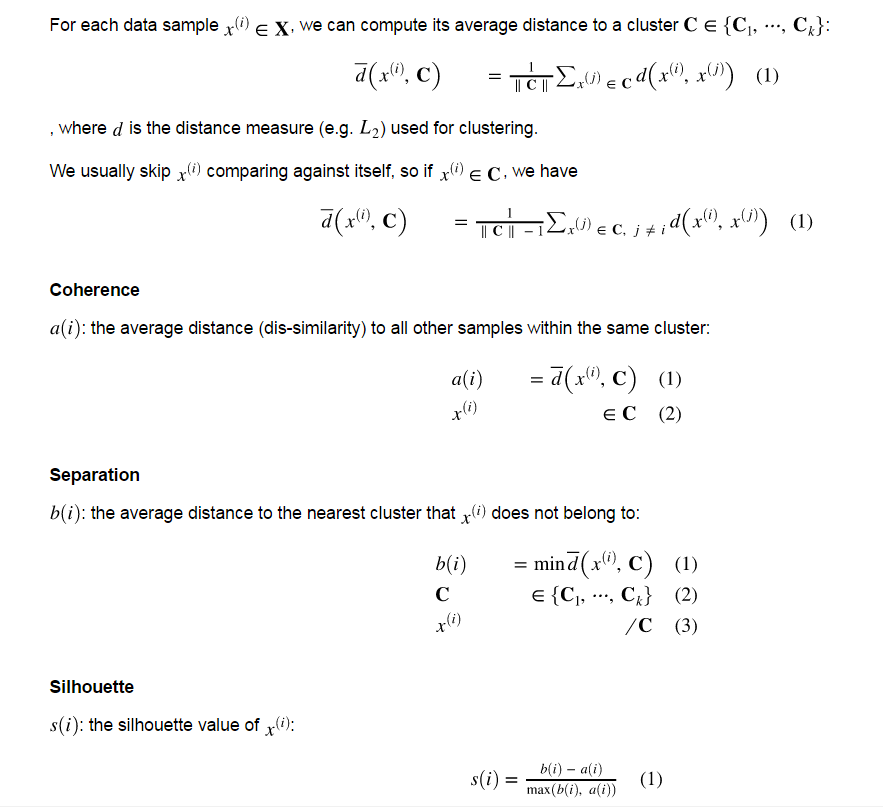
**S(i) is close to -1 🡪bad clustering.**

1. **Density based clustering:**

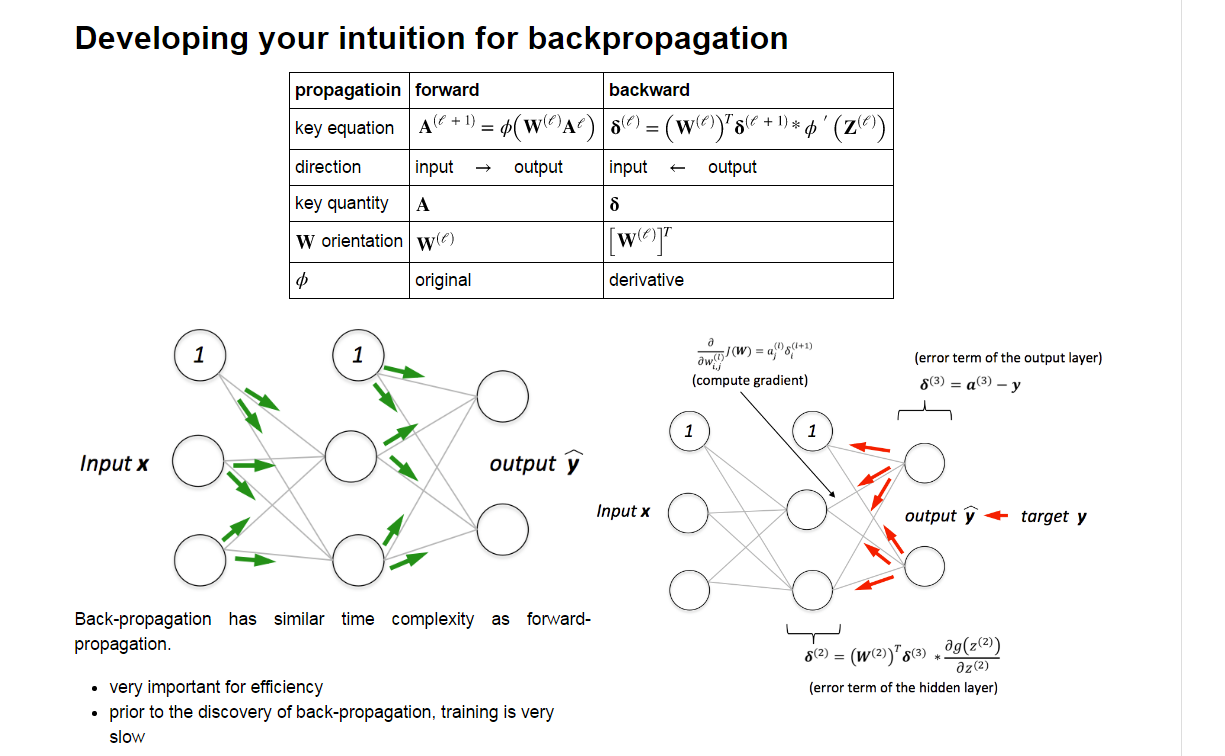
****

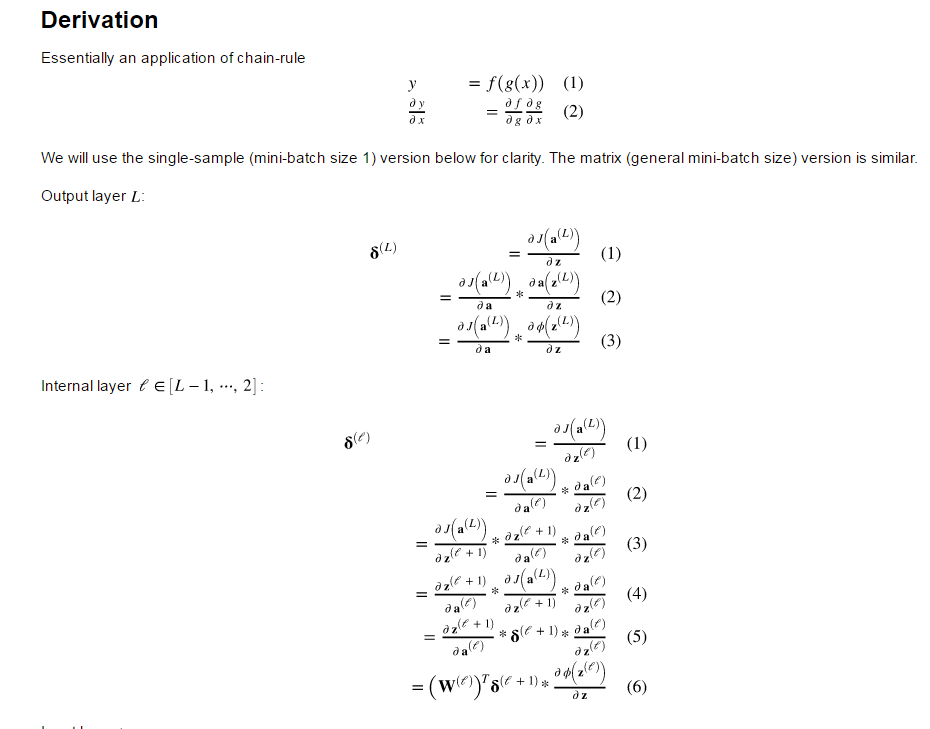
1. **层聚类：**

****

****

1. **Neural network:**

****

****