**Thinking1 什么是监督学习，无监督学习，半监督学习**

监督学习的训练数据是有标签的，训练目标是能够给新数据（测试数据）以正确的标签。经典算法有：支持向量机、线性判别、决策树、朴素贝叶斯。

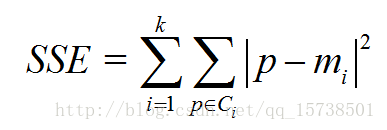
无监督学习常常被用于数据挖掘，用于在大量无标签数据中发现些什么。训练数据没有标签，训练目标是能对观察值进行分类或者区分等。机器会主动学习数据的特征，并将它们分为若干类别，相当于形成“未知的标签”。算法：K-Means等。

半监督学习（Semi-Supervised learning）：

* 半监督学习介于监督学习与无监督学习之间
* 通常半监督学习的任务与监督学习一致，即任务中包含有明确的目标（如分类），采用的数据既包括有标签的数据，也包括无标签的数据
* 作用：只有少量的数据有Label，利用没有标签的数据来学习整个数据的潜在分布
* 半监督学习算法有很多，标签传播算法label propagation是其中之一

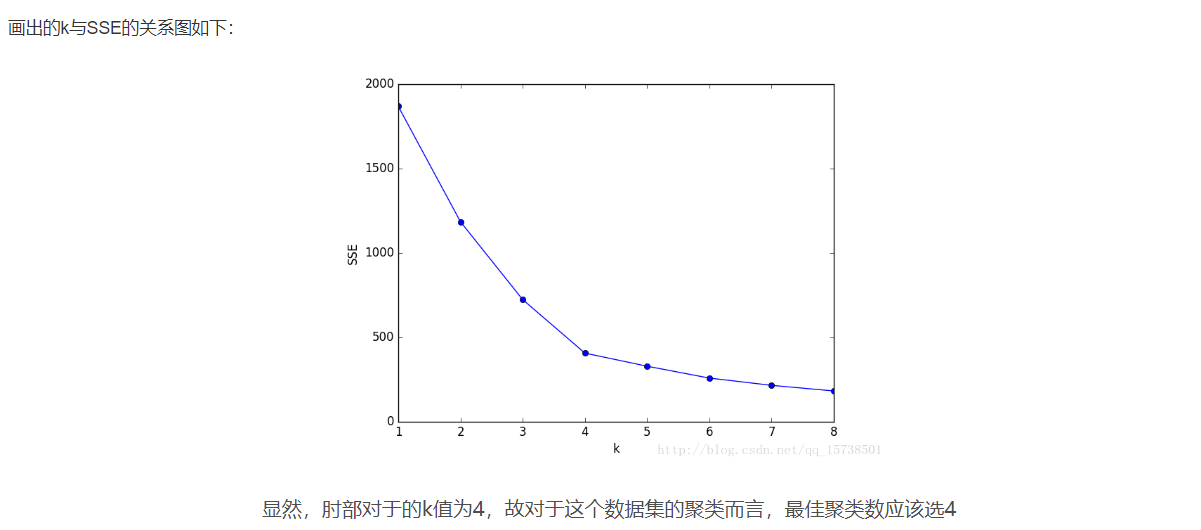
**Thinking2 K-means中的k值如何选取**

采用手肘法。手肘法的核心指标是SSE(sum of the squared errors，误差平方和)，



其中，Ci是第i个簇，p是Ci中的样本点，mi是Ci的质心（Ci中所有样本的均值），SSE是所有样本的聚类误差，代表了聚类效果的好坏。

       手肘法的核心思想是：随着聚类数k的增大，样本划分会更加精细，每个簇的聚合程度会逐渐提高，那么误差平方和SSE自然会逐渐变小。并且，当k小于真实聚类数时，由于k的增大会大幅增加每个簇的聚合程度，故SSE的下降幅度会很大，而当k到达真实聚类数时，再增加k所得到的聚合程度回报会迅速变小，所以SSE的下降幅度会骤减，然后随着k值的继续增大而趋于平缓，也就是说SSE和k的关系图是一个手肘的形状，而这个肘部对应的k值就是数据的真实聚类数。当然，这也是该方法被称为手肘法的原因。



**Thinking3 随机森林采用了bagging集成学习，bagging指的是什么**

随机森林的生成：

* 森林中每棵树都是独立的
* bagging思想是将若干个弱分类器的分类结果进行投票选择，从而组成一个强分类器。少数服从多数，获得票数最多的类别就是森林的分类结果
* bagging不用单棵决策树来做预测，增加了预测准确率，即不容易过拟合，减少了预测方差
* 如果训练集大小为N，对于每棵树而言，随机且有放回地从训练集中的抽取N个训练样本（采样方式称为bootstrap sample），作为该树的训练集。每棵树的训练集都是不同的，而且里面包含重复的训练样本
* 如果每个样本的特征维度为M，指定一个常数m<<M，随机地从M个特征中选取m个特征子集，每次树进行分裂时，从这m个特征中选择最优的
* 每棵树都尽最大程度的生长，没有剪枝过程

**Thinking4 主动学习和半监督学习的区别是什么**

对于主动学习：

* 如果机器可以自己选择学习的样本，它可以使用较少的训练取得更好的效果
* 需要人工介入，模型主动向worker提供数据

对于半监督学习：

* 指在训练数据十分稀少的情况下，利用没有标签的数据，提高模型效果的方法