THINKING 1 什么是监督学习，无监督学习，半监督学习?

1 监督学习：数据有确定的标签，很多算法属于监督学习：线性回归，逻辑回归，随机森林等等

2 无监督学习：数据没有确定的标签，算法直接对输入的数据集进行建模，如聚类方法

3 半监督学习：部分数据有标签，部分数据没有标签，综合利用两部分数据来构建合适的算法

THINKING 2 K-means中的k值如何选取

1 手肘法:画出k和误差（sse）之间的关系图，选取随着k的增加，误差减小急剧变换的点（肘部），作为合适的k值

2 轮廓系数法，求出平均轮廓系数，选取平均轮廓系数最小的k值。

$$S = \frac{b-a}{\max(a,b)}$$

$$C_j = arg\ \min_{C_k} \frac{1}{n} \sum_{p \in C_k} |p - X_i|^2$$

其中，a为聚合度，即某个样本距离距离同簇其他样本的平均距离。b为分离度，即某个样本距离最近簇的样本的平均距离

3 通常需要综合考虑，手肘法和轮廓系数的结果

THINKING 3 随机森林采用了bagging集成学习，bagging指的是什么

1 **Bagging算法** （英语：**B**ootstrap **agg**regat**ing**）又称**装袋算法**，是机器学习领域的一种团体学习算法。

2 **Bagging的重点是获得一个比其组成部分方差更小的模型**

3 步骤：给定一个大小为n的训练集D，Bagging算法从中均匀、有放回地（即使用自助抽样法）选出m个大小为n'的子集Di，作为新的训练集。在这m个训练集上使用分类、回归等算法，则可得到m个模型，再通过取平均值、取多数票等方法，即可得到Bagging的结果 。

THINKING 4 表征学习和半监督学习的区别是什么

1 表征学习是为了学习到特征的计算方法而半监督学习通常还是为了最后的结果（标签或者数值结果）

2 特征学习的方法包括有监督和无监督的各种方法：树模型，pca，神经网路等等

3 半监督学习的方法主要包括标签传播算法等