第五节　数据科学与大数据

无监督学习	无监督学习的数据集中，每个观测单位只有自变量(特征xi)，没有因变量（标签yi）。 无监督学习的主要任务是探索数据之间的内在联系和结构。 无监督学习中有两大类典型任务：聚类和降维。

聚类	把一组数据按照差异性和相似性分为几个类别使得同类数据相似性尽可能大，不同类数据相似性尽可能小，跨类的数据关联性尽可能低。 聚类分析常用于客户细分，文本归类，结构分组，行为跟踪。 常用的聚类方法包括：基于划分的方法（例如k-均值算法）、基于分层的方法、基于密度的方法、基于网格的方法和基于模型的方法。
降维	在不损失过多信息的前提下将N个相关特征降为K个不相关特征，使其具有更好的解释性，也称为特征提取。 例如：根据客户的能力，品格，担保，资本，环境等特征评价客户的信用等级。 常用的降维方法包括：主成分析法，因子分析法。

第五节　数据科学与大数据

半监督学习	半监督学习是监督学习与无监督学习相结合的一种学习方法。 半监督学习的数据集中，一部分观测单位既有自变量(特征xi)，又有因变量（标签yi），另一部分观测单位只有自变量(特征xi)，没有因变量（标签yi），而且没有标签的观测单位数量远大于有标签的观测单位数量。
常见的半监督学习有半监督分类、半监督回归、半监督聚类。	

第五节　数据科学与大数据

【多选题】数据挖掘包含了几种不同的含义，具体有（　）。

A. 数据源必须是真实的、大量的、有噪声的

B. 挖掘完成后必须能获得反馈

C. 发现的是用户感兴趣的知识

D. 发现的知识是可接受、可理解、可运用的

E. 并不要求发现放之四海而皆准的知识

网校答案：ACDE

第五节　数据科学与大数据

【多选题】下列属于聚类分析的是（　）。

A. 结构分组

B. 客户细分

C. 行为跟踪

D. 主成分析法

E. 因子分析法

网校答案：ABC

【单选题】下列关于数据挖掘说法正确的是（ ）。

A. 每个观测单位只有自变量(特征xi)，没有因变量（标签yi）属于监督学习

B. 每个观测单位既有自变量(特征xi)，又有因变量（标签yi）属于无监督学习

C. 根据客户的能力，品格，担保，资本，环境等特征评价客户的信用等级属于聚类技术

D. 把一组数据按照差异性和相似性分为几个类别使得同类数据相似性尽可能大，不同类数据相似性尽可能小，跨类的数据关联性尽可能低

网校答案：D

谢谢观看

THANK YOU