《数据挖掘》复习题

（初稿，2017年）

一、名词解释

**1、数据挖掘**

从数据集中识别出有效的、新颖的、潜在有用的，以及最终可理解的模式的非平凡过程。

**2、孤立点分析**

专门进行孤立点研究的方法与技术。统计方法占主流，考察数据的分布，用距离来度量。

**3、数据清理**

主要通过填写空缺的值，平滑噪声数据，识别、删除孤立点，解决不一致性来清理数据。

**4、数据集成**

将多个数据源中的数据整合到一个一致的存储中。

**5、数据归约**

可以用来得到数据集的归约（压缩）表示，它小得多，但仍保持数据的完整性。对归约后的数据集挖掘将更有效，并产生相同（或几乎相同）的分析结果。

**6、数据离散化**

通过将属性域划分为区间，减少给定连续属性值的个数。区间的标号可以代替实际的数据值。

**7、数据规范化**

将数据按比例缩放，使之落入一个小的特定区间。

**8、关联规则**

关联规则是所有形如X->Y蕴涵式，其中X和Y是不相交的项集。

1. **K-均值聚类**
2. 任意选择k个对象作为初始的类的中心

Repeat

2. 根据类中样本的平均值,将每个样本(重新)赋给最相近的类

3. 更新类的平均值

until

4.不再发生变化,即没有对象进行被重新分配时过程结束。

该算法试图找出使平方误差值最小的k个划分。

**10、文本自动分类**

文档自动分类（Automatic Text Categorization，ATC）, 是指在给定的分类体系下，根据文本的内容用计算机程序确定文本所属类别的过程。

**11、因子分析**

二、简答题

1. **互联网数据有哪些主要特征？**

数量巨大；数据类型繁多：表格、结构与半结构数据、多媒体数据等；异构性；动态性；互联性；包含大量噪音；虚拟社会：包括人、组织之间的交互。

1. **简述什么是数据挖掘**

从大量数据中提取或发现（挖掘）知识的过程。从数据集中识别出有效的、新颖的、潜在有用的，以及最终可理解的模式的非平凡过程。

非平凡；有效性；新颖性；潜在有用性；最终可理解性；有趣性；

1. **简述数据挖掘的主要步骤**

数据准备：数据选择、数据预处理、数据变换、数据规约；数据挖掘算法的选择：明确任务、考虑用户、结合数据集；结果的解释评估。

1. **什么是孤立点分析**

专门进行孤立点研究的方法和技术。孤立点与数据的一般行为或模式不一致。多数为噪声或异常数据，常被剔除。在某些应用中，孤立点更为有趣。统计方法占主流，考察数据的分布，用距离来度量。

1. **简述什么是数据清理**

主要通过填写空缺的值，平滑噪声数据，识别、删除孤立点，解决不一致性来清理数据

1. **简述什么是数据集成**

将多个数据源中的数据整合到一个一致的存储中。

1. **简述什么是数据归约**

可以用来得到数据集的归约（压缩）表示，它小得多，但仍保持数据的完整性。对归约后的数据集挖掘将更有效，并产生相同（或几乎相同）的分析结果。

1. **简述什么是关联规则挖掘**

关联规则挖掘就是发现大量数据中项集之间有趣的关联。具体地,关联规则挖掘是指找出所有满足最小支持度阈值与最小置信度阈值的规则。

1. **简述什么是聚类分析**

一个“无指导的学习”过程,它将数据对象分组成多个类或簇，使得在同一个簇中个体的具有较高的相似度，而不同簇中的个体差别较大。

**10、简述什么是文本自动分类**

文本自动分类（Automatic Text Categorization，ATC）, 是指在给定的分类体系下，根据文本的内容用计算机程序确定文本所属类别的过程。

**11、举例说明什么是K-均值聚类算法。**

1. 任意选择k个对象作为初始的类的中心

Repeat

2. 根据类中样本的平均值,将每个样本(重新)赋给最相近的类

3. 更新类的平均值

until

4.不再发生变化,即没有对象进行被重新分配时过程结束。

举例：略

**12、简述什么是层次聚类法。**

层次聚类法（Hierarchicalmethods）将样本集合进行层次分解，组成一颗凝聚树。根据层次的形成方式可以分为两类：凝聚的方法（自底向上）、分裂的方法（自顶向下）。

以凝聚法为例：

它第一步先把最近的两类（点）合并成一类，然后再把剩下的最近的两类合并成一类；这样下去，每次都少一类，直到最后只有一大类为止。越是后来合并的类，距离就越远。

**13、聚类分析有哪些主要方法？**

K-means 聚类；层次聚类（凝聚法、分裂法）；基于密度的聚类（DBSCAN）

**14、什么是KNN-分类算法？**

对于一个待分类样本，在训练集中找到k个最相近的邻居，使用这k个邻居的类别为该样本的候选类别。该样本与k个邻居之间的相似度按类别分别求和，减去一个预先得到的截尾阈值，就得到该样本的类别测度。

**15、举例说明什么是决策树**

决策树（decision tree）: 是一个类似于流程图的树结构，其中，每个内部节点表示在一个属性上的测试，每个分支代表一个测试输出，而每个叶节点代表类，树的最顶层为根节点。

举例：略。

**16、简述数据挖掘的主要软件工具**

SPSS modeler、R（arules）、Python（sklearn）.

**17、聚类分析与自动分类有何区别与联系？**

聚类分析是一个无监督的过程，自动分类属于有监督学习。两者都是要在一个离散变量上对目标进行分类。对于聚类来讲我们不知道数据应当分为多少类，在聚类之前也不清楚每一类的具体含义，需要解释；而自动分类中所有的分类已经实现给定，有明确的语义，我们要做的只是利用其它数据，训练出分类器，将对象映射到某一类中。

**18、因子分析法的主要思想是什么？**

三、论述题

1. **详述数据挖掘的主要步骤？**

数据准备：数据选择、数据预处理、数据变换、数据规约；数据挖掘算法的选择：明确任务、考虑用户、结合数据集；结果的解释评估。

1. **在进行数据挖掘或分析之前为什么要进行数据预处理？有哪些主要方法？**

为了提高数据挖掘的精度，降低实际挖掘所需要的时间。

数据清理、数据集成、数据变换、数据归约、数据离散化与概念分层。

1. **举例说明数据预处理的主要方法？**

举例：

数据变换：可以有平滑、聚集、概化、规范化等。其中规范化又可以采用min-max规范化、z-score规范化、小数定标规范化等。

1. **举例说明什么是关联规则挖掘？它有哪些典型应用？**

关联规则挖掘就是发现大量数据中项集之间有趣的关联.具体地,关联规则挖掘是指找出所有满足最小支持度阈值与最小置信度阈值的规则。

比如我们可以从购物篮数据中挖掘出哪些商品经常被一起购买。

可以应用于高校图书馆管理系统、文献借阅历史数据分析、发现学科间的相关性等。

1. **举例说明什么是聚类分析？它有哪些典型应用？**

有100个学生的数学成绩，我们可以进行聚类，并对产生的簇进行解释；

浏览文本、显示文本集合、组织搜索引擎的返回结果、帮助市场分析人员刻画不同用户群体的特征。

1. **举例说明什么是自动文本分类？它有哪些主要应用？**

文本自动分类（Automatic Text Categorization，ATC）, 是指在给定的分类体系下，根据文本的内容用计算机程序确定文本所属类别的过程。

举例：天网搜索。

应用：搜索引擎可以对新闻分类，方便读者检索。论文抄袭检测。舆情分析等。

**7、什么是数据标准化 ？举例说明有哪些主要方法？**

即为规范化。min-max规范化、z-score规范化、小数定标规范化。

**8、什么是层次聚类法？请举例说明层次聚类法的基本过程。**

层次聚类法（Hierarchicalmethods）将样本集合进行层次分解，组成一颗凝聚树。根据层次的形成方式可以分为两类：凝聚的方法（自底向上）、分裂的方法（自顶向下）。

以凝聚法为例：

它第一步先把最近的两类（点）合并成一类，然后再把剩下的最近的两类合并成一类；这样下去，每次都少一类，直到最后只有一大类为止。越是后来合并的类，距离就越远。

**9、自动文本分类的一般过程？有哪些主要方法？**

首先我们有labeled集合和unlabeled集合，先对两个集合进行必要的预处理使之便于分类。然后训练集又分为train与test，选定合适的特征选择算法和分类算法作用在train上，根据train的预测值与实际值的差距来调整参数，比如阈值策略。之后用test测试这些算法的效果。选取较好的模型作为最终的模型。对于待分类样本应用确定的特征选择算法、分类算法以及阈值选取，就可以得到预测分类。

决策树、KNN、朴素贝叶斯、支持向量机、神经网络。