## 什么是数据挖掘？

数据挖掘是通过对(大规模)观测数据集的分析,寻找确信的关系,并将数据以一种可理解的且利于使用的新颖方式 概括数据的方法.

## 什么可以挖掘出来？

知识 ；隐藏的关系；不确定的现实；对决策或理解任务性质可能有用的信息。

## 数据挖掘是怎样的？

适用于大数据；观测的数据（不是实验数据，数据收集过程中没有人为控制）；

不受怀疑的的关系（关系应该是正确和重要的）；新颖的（众所周知的结论是无用的）；可理解的（挖掘结果将呈现给用户的决策）；有用处的。

## 什么是机器学习？

计算机程序被要求从经验E中学习某些任务T和绩效测量P，如果它在T中的任务中的表现（由P测量）随经验E而改善，那么我们称之为机器学习。

## 机器学习框架

监督学习；无监督学习；强化学习；半监督学习；主动学习

## 监督学习 

在ML文献中，监督学习问题具有以下特征：我们主要对预测感兴趣。我们只想预测一件事。

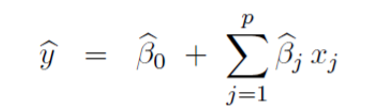
已经指定了我们想要预测的值的可能值，并且我们有一些其值已知的训练案例。 

我们想要预测的事物称为目标或响应变量。通常，我们需要训练数据

## **如何做出预测？**

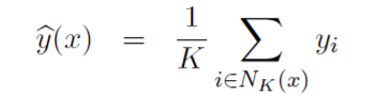
主要方法

我们可以通过使用训练数据来估计模型的参数来训练模型，然后使用这些参数来对测试数据进行预测。 当我们对测试数据进行预测时，这种方法可以节省计算量。因为我们估计一次参数，并多次使用它们。 例如线性回归



其他方法：最近邻方法

基于训练案例的子集对测试数据进行预测，例如，通过近似P（ y | x ） 的均值，中值或模式 。



重要问题：如何选择K？ 如果K太小，我们可能会“过度拟合”，但如果K太大，我们将对与测试用例无关的训练案例进行平均。

## 两种预测方法的比较？

* 这两种方法在计算方面是对立的。

NN近邻方法是基于记忆的方法。我们需要记住所有的训练数据。

线性回归方法在获取参数后，可以忘记训练数据，只需使用参数即可。 

* 在统计特性方面对立。

NN对数据做了一些假设 ，并且具有过度拟合的高可能性

线性回归对数据做出了强有力的假设，因此具有很高的偏差可能

## 无监督学习

对于无监督的学习问题，我们不关注任何特定事物的预测 ，而是试图找到数据的有趣方面。

聚类 示例：我们可能会发现患有类似症状的患者群，我们称之为疾病。我们可能会发现大量图像的集群。

Reinforcement Learning

Learning from interaction(with environment) 

Goal-directedlearning 

Learning what to doand its effect 

Trial-and-errorsearch and delayed reward 

The two most important distinguishing features of reinforcement learning

## 强化学习

从互动中学习（与环境有关） 目标-导向学习 学习去做什么以及它的效果 试错搜索和延迟奖励，是两个最重要的区别特征

计算机？必须利用它已经知道的东西来获得奖励，但它也必须进行探索，以便在将来做出更好的行动选择。 

困境-无论是开发还是探索都不能完全追求，而不会失败。

## 半监督学习

顾名思义，它对于需要培训所需的标记和未标记数据量处于监督和非监督学习技术之间， 与监督学习相比，目标是减少所需的监督量。同时将无监督聚类的结果提升到用户的期望值。

## SSL（半监督学习）概述

约束聚类（Constrained Clustering）；距离度量学习（Distance Metric Learning）

基于流形的学习（Manifold based Learning）；基于学习（压缩感知）[based Learning (Compressed Sensing)]; 主动学习

## 约束聚类

当我们有以下任何一项时：数据子集的类标签。有关集群的领域知识。有关对象之间“相似性”的信息。用户偏好。 可以是成对约束或标记子集。 必须链接或不能链接约束。标签总是可以转换为成对关系。 

可以通过搜索尊重约束的分区来进行聚类。[Can be clustered by searching for partitionings that respect the constraints.]

最近趋势是基于相似性的方法。

## 主动学习

基本理念： 

传统的监督学习算法被动地接受训练数据。而主动学习是从未标记的数据中查询有关信息图像的注释。理论结果表明，通过主动学习可以大大减少训练规模！但是如何找到信息量最大的图像呢？

一个想法是使用不确定性抽样 我们不确定分类的图像可能会提供信息！

什么是不确定性的概念？想法：在训练集上训练像SVM这样的分类器。 对于每个未标记的图像，输出指示类成员资格的概率。估计概率可用于推断不确定性。可以使用一对一SVM方法来处理多个类。

## 其他框架

多标签学习 多实例学习 多实例多标签学习 深度学习

## 机器学习&数据挖掘问题与传统统计

动机：

预测----理解----因果关系 

许多传统统计数据主要是通过表明一个因素导致另一个因素而产生的。 接下来是理解，最后预测。 这些方法通常假设线性关系（可能在简单转换之后），简单分布（例如正态分布）。

在机器学习和数据挖掘中，顺序通常是相反的：预测是最重要的。例如有很多变量的问题。

## 机器学习&数据挖掘和传统统计学的态度

ML＆DM

1.没有解决哲学或广泛接受的理论框架。

2.如果它们似乎运作良好，愿意使用临时方法（虽然外观可能会产生误导）。

3.强调自动方法，很少或没有人为干预。

4.适用于许多问题的方法。

5.大量使用计算。

传统统计

1.经典（频率论）和贝叶斯哲学竞争。

2.不愿意使用没有一些理论依据的方法（甚至没有意义）

3。强调通过情节和诊断来辅助人类判断。

4.基于科学知识的模型。

5.最初设计用于手动计算，但现在计算机计算非常重要。

## 机器学习的的挑战

处理的复杂性 

涉及许多变量时，我们如何处理这种复杂性而不会遇到麻烦。

优化和整合

通常涉及为某些参数（优化问题）找到最佳值，或者在许多合理值上求平均值（积分问题）。 当有很多参数时，我们怎样才能有效地做到这一点。

可视化

了解发生的事情很难，2D？ 3D？

当存在许多变量或参数时，所有这些挑战都会变得更大—所谓的 “维度诅咒”。

但更多变量也提供了更多信息

这是一个祝福？还是诅咒？

## 如何处理复杂性？

正确处理复杂性是机器学习的关键问题。 

限制复杂性是一种方法 

使用足够复杂的模型来表示问题的基本方面，但这并不复杂到以至于发生过度拟合。 

当我们选择适合我们已经很好的数据的模型的参数，但是对新数据做得不好（泛化能力差）时，就表明发生了过度拟合。

交叉验证，正规化，减少维度是另一种可能性。 

很明显，如果能够找到如何将很多变量减少到少量，事情就会变得更简单。 

对复杂性求平均值是贝叶斯方法。可能需要使用复杂的模型，但不要选择单个参数值。相反，使用适合数据的所有参数值对发现的预测进行平均，这对于问题是合理的

## 复杂性应该受到限制吗？

如果我们使用模型的“最佳拟合”参数进行预测，我们必须限制参数的数量以避免过度拟合。

对于此示例，M = 3 的 模型似乎很好。 我们可以使用“交叉验证”方法为M选择一个好的值，该方法寻找在预测来自其余数据的数据的一部分时最有效的值。 

但是我们知道sin（2πx）不是多项式函数，它有一个无穷大的系列表示，其任意高阶。

怎么会是很好用的模型，我们知道是假的？ 

贝叶斯回答：这不好。我们应该放弃使用最佳参数的概念，而是平均参数的所有合理值。 然后我们可以使用一个尽可能接近正确的模型（可能是一个非常复杂的模型）。

## 减少维度

假设输入数据的维数为1000，我们可以用较少的维度替换它们，而不会丢失信息。 

简单的方法是使用PCA（主成分分析） 

假设所有数据都在一个空间中，我们首先找到这些数据点的最高方差的方向，然后找到与第一个方差正交的第二高方差的方向， 依此类推...... 

通过某些方向上的输入投影替换每个训练样本。 

可能会丢弃有用的信息 ，但要保留大部分信息 。

## ML和DM之间的关系

数据挖掘（DM）是从数据中提取模式的过程。主要目标是了解关系，验证模型或识别意外不到的关系。

机器学习（ML）算法允许计算机从数据中学习。学习过程包括提取模式，但最终目标是使用知识对新数据进行预测。

## ML和DM之间的差异

它们通常可以互换，但是数据挖掘更常用于存在大量数据的问题，其中计算效率比统计复杂性（通常是业务应用程序）更重要。机器学习更常用于人工智能的问题 - 例如在视觉场景中识别物体。

**三、简答题**

## 应用贝叶斯决策需要满足的三个前提条件

（1）分类的类别数；（2）先验概率；（3）各类的类条件概率密度。

## 您对先验概率和后验概率理解

* 先验概率：预先已知的或者可以估计的模式识别系统位于某种类型的概率。根据大数定律，当训练集包含充足的独立同分布样本时，先验概率就可以通过各类样本出现的频率来进行估计。
* 后验概率：后验概率是通过贝叶斯公式对先验概率进行修正，计算而得出的概率。表示系统在某个具体的模式样本X条件下位于某种类型的概率。

## Fisher线性判别的基本思想

解决多维到一维的投影问题，利用线性判别方法解决一维分类问题。

从k个总体中抽取具有p个指标的样品观测数据，借助方差分析的思想构造一个线性判别函数，构造原则是使得总体之间区别最大，而使各总体内部的离差最小。有了线性判别函数之后，对于一个新的样品，将它的p个指标值带入线性判别函数式中求出结果值，然后根据一定的规则，就能判别新的样品属于哪个总体。

* 投影降维：将多维空间的样本投影到一维空间，根据实际情况找到一条最好的、易于分类的投影线。寻找合适的投影方向，即寻找好的变换向量。
* 一维分类：当维数和样本数都很大时，可用贝叶斯决策规则；上述条件都不符合，可用先验知识选定分界阈值点y0，再有决策规则判断x属于何类别。

## 何为k-近邻法

（1）K-近邻学习是一种常用的监督学习方法，

（2）k-近邻法的基本思想：给定一个训练数据集，对新的输入实例，基于某种距离度量找出训练集中与其最靠近的 K 个实例（邻居）。然后基于这 K 个实例的信息来进行预测，使各点邻域体积为数据的函数，而不是样本数的函数，实现各点密度估计。看这K个近邻中多数属于哪一类，就把输入实例归为哪一类；

## 您对非线性支持向量机（SVM）理解

对于线性支持向量机，选择一个合适的惩罚参数C >0 ，并构造凸二次函数线性规划问题，求得原始问题的对偶问题的最优解\alpha ^\ast ,由此可以求出原始问题的最优解；

在处理非线性问题时，可以通过将非线性问题转化成线性问题，并通过已经构建的线性支持向量机来处理。在线性不可分的情况下，SVM首先在低维空间中完成计算，然后通过选择合适的核函数来将样本从原始空间映射到高维特征空间，最终在高维特征空间中构造出最优分离超平面，从而分离非线性数据。

（1）选取适当的核函数K(x,y)和适当的参数C>0，构造原始问题的对偶问题,求得对应的最优解\alpha ^\ast；

（2）选择\alpha ^\ast的一个满足0\leq \alpha _j\leq C的分量，求b^\ast；

（3）构造决策函数；

## 何为度量学习

在机器学习中，对高维数据进行降维的主要目的是希望找到一个合适的低维空间，在此空间中学习比原始空间更好。事实上，每个空间对应了在样本属性上定义的一个距离度量。度量学习可以根据不同的任务来自主学习出针对某个特定任务的度量距离函数，找出一个合适的距离度量。

度量学习方法可以分为通过线性变换的度量学习和度量学习的非线性模型。

## 何为半监督学习（SSL）

半监督学习是监督学习与[无监督学习](http://baike.baidu.com/view/3552442.htm)相结合的一种学[习](http://baike.baidu.com/subview/361686/20348304.htm)方法，就是让学习器不依赖外界交互，自动地利用未标记样本来提升学习性能。它主要考虑如何利用少量的标注样本和大量的未标注样本进行训练和分类的问题。

SSL的成立依赖于模型假设，当模型假设正确时，无类标签的样例能够帮助改进学习性能。包括：1)平滑假设2)聚类假设3)流形假设。

## 何为聚类

聚类试图将数据集中的样本划分为若干个通常是不相交的子集称为一个『簇』，通过这样的划分，每个簇可能对应于一些潜在的概念（类别），并且这些概念对于聚类算法而言事先是未知的，聚类过程仅能自动地形成簇结构，簇所对应的概念语义需要使用者来把握和定义。

## 对稀疏表达的理解

稀疏表达的目的就是在给定的超完备字典中用尽可能少的原子来表示信号，可以获得信号更为简洁的表示方式，从而使我们更容易地获取信号中所蕴含的信息，更方便进一步对信号进行加工处理，如压缩、编码等。两大主要任务就是字典的生成和信号的稀疏分解。

假设一个样本数据 D，D 对应的矩阵中存在很多零元素，并且它们不是以整行整列的形式出现的，那么这种表达是稀疏的。这样的稀疏表达形式对学习任务会有不少好处。例如，SVM 在文本上有很好的性能。

若给定数据集 D 是稠密的，即普通非稀疏数据，我们可以通过『字典学习』（『稀疏编码』）来将样本转化为合适的稀疏表示。

## 对流型学习的理解

假设数据是均匀采样于一个高维欧氏空间中的低维[流形](http://baike.baidu.com/view/159731.htm)，流形学习就是从高维采样数据中恢复低维流形结构，即找到高维空间中的低维流形，并求出相应的嵌入映射，以实现维数约简或者数据可视化。它是从观测到的现象中去寻找事物的本质，找到产生数据的内在规律。

流形学习也通常被用于可视化，因为当维数被降至二维或三维时，能进行可视化。等度量映射和局部线性嵌入是两种著名的流形学习方法。

## 对同分布问题的理解

我们在设计一个模式识别时，是有一些假设或前提条件的，其中之一就是：假定用于训练或建模的有标记样本集的分布，是和全集的分布相同的，这就是同分布。

如果满足同分布的预期模型将有很好的泛化能力。

## 对模型泛化能力的理解

经过训练的网络对于不是样本集的输入也能给出合适的输出，该性质称为泛化能力，意味着从训练样本数据得到的模型也能够很好的适应测试样本数据。

通常期望经训练样本训练的网络具有较强的泛化能力，但并非训练的次数越多越能得到正确的输入输出映射关系。训练过程中过分追求分类器将训练样本尽量分类正确或完全分类正确反倒使泛化能力下降。

**四、论述题**

## 何为集成学习

集成学习是使用一系列学习器进行学习，并使用某种规则把各个学习结果进行整合从而获得比单个学习器更好的学习效果的一种机器学习方法。集成学习的思路是在对新的实例进行分类的时候，把若干个单个分类器集成起来，通过对多个分类器的分类结果进行某种组合来决定最终的分类，以取得比单个分类器更好的性能。如果把单个分类器比作一个决策者的话，集成学习的方法就相当于多个决策者共同进行一项决策。

## 简述模式的概念及其直观特性，模式识别的分类，有哪几种方法。

答（1）：什么是模式？广义地说，存在于时间和空间中可观察的物体，如果我们可以

区别它们是否相同或是否相似，都可以称之为模式。

模式所指的不是事物本身，而是从事物获得的信息，因此，模式往往表现为具有时间和空间分布的信息。

模式的直观特性：可观察性；可区分性；相似性。

答（2）：模式识别的分类：

假说的两种获得方法（模式识别进行学习的两种方法）：

监督学习、概念驱动或归纳假说；非监督学习、数据驱动或演绎假说。

模式分类的主要方法：

数据聚类：用某种相似性度量的方法将原始数据组织成有意义的和有用的各种数据集。是一种非监督学习的方法，解决方案是数据驱动的。

统计分类：基于概率统计模型得到各类别的特征向量的分布，以取得分类的方法。特征向量分布的获得是基于一个类别已知的训练样本集。是一种监督分类的方法， 分类器是概念驱动的。

结构模式识别：该方法通过考虑识别对象的各部分之间的联系来达到识别分类的目的。（句法模式识别）

神经网络：由一系列互相联系的、相同的单元（神经元）组成。相互间的联系可以在不同的神经元之间传递增强或抑制信号。增强或抑制是通过调整神经元相互间联系的权重系数来（weight ）实现。神经网络可以实现监督和非监督学习条件下的分类。

## 什么是神经网络？有什么主要特点？选择神经网络模式应该考虑什么因素？

答（1）：所谓人工神经网络就是基于模仿生物大脑的结构和功能而构成的一种信息处理系统（计算机）。由于我们建立的信息处理系统实际上是模仿生理神经网络，因此称它为人工神经网络。 这种网络依靠系统的复杂程度，通过调整内部大量节点之间相互连接的关系，从而达到处理信息的目的。

人工神经网络的两种操作过程：训练学习、正常操作（回忆操作）。

答（2）：**人工神经网络的特点：**

固有的并行结构和并行处理；知识的分布存储；有较强的容错性； 有一定的自适应性；

**人工神经网络的局限性：**

人工神经网络不适于高精度的计算；人工神经网络不适于做类似顺序计数的工作；人工神经网络的学习和训练往往是一个艰难的过程；人工神经网络必须克服时间域顺序处理方面的困难；硬件限制；正确的训练数据的收集。

答（3）：选取人工神经网络模型，要基于应用的要求和人工神经网络模型的能力间的匹配，主要考虑

**Svm** 优点： 1.可用于线性/非线性分类，也可以用于回归，泛化错误率低，计算开销不大，结果容易解释； 2.可以解决小样本情况下的机器学习问题，可以解决高维问题 可以避免神经网络结构选择和局部极小点问题。 3.SVM是最好的现成的分类器，现成是指不加修改可直接使用。并且能够得到较低的错误率，SVM可以对训练集之外的数据点做很好的分类决策。

缺点：对参数调节和和函数的选择敏感，原始分类器不加修改仅适用于处理二分类问题。

**Logistic回归**：根据现有数据对分类边界线建立回归公式，依次进行分类。

优点：实现简单，易于理解和实现；计算代价不高，速度很快，存储资源低；

缺点：容易欠拟合，分类精度可能不高

**EM 期望最大化算法**-上帝算法

只要有一些训练数据，再定义一个最大化函数，采用EM算法，利用计算机经过若干次迭代，就可以得到所需的模型。EM算法是自收敛的分类算法，既不需要事先设定类别也不需要数据见的两两比较合并等操作。缺点是当所要优化的函数不是凸函数时，EM算法容易给出局部最佳解，而不是最优解。