数据清理

缺失值处理

- **删除**

对于缺失值比较多的元组，或损失了重要属性的元组，可考虑删除法。比如对于缺失类别的数据，在分类任务中就可以予以删除。该方法比较适合行缺失比例多的情况，不适合列缺失值百分比变化大的情况。

- **人工填充**

这种方法主要来源于专家判断，该方法数据准确行取决于人工，而且比较费时，不适合数据集大、缺失值很多的情况。

- **全局常量**

使用一个固定的值填充所有缺失值，如有0填充所有空值（NULL），不过该方法需要注意在分析时误将该全局值认为是数据具有的某个相同价值。

- **使用属性的集中度量**

比如使用属性的均值、中位数、众数等填充缺失值。对称数据使用均值、倾斜数据使用中位数。

- **同类样本的集中度量**

该方法和和属性的集中度量类似，不过该方法有一个前提条件，采用的不是全部数据，而是在其他某些属性上相同或相似的样本。选择平均数和中位数方法与前面相同。

- **最有可能的值**

该方法需要对数据进行更加复杂的分析，如回归、贝叶斯等数据推断的方式来进行填充。或通过决策树等分类策略来判定缺失数据最有可能存在的类，然后使用同类样本的集中度量。

噪声数据

- **分箱**

噪声数据和缺失数据都属于脏数据，是数据清理的主要对象，不过噪声数据不同的是，它是被测量的变量的随机误差或方差，噪声数据可以转化为缺失数据，反之不成立。

分箱适用于有序数据，它使用数据的近邻来光滑有序数据值。分布到同一个箱中的数据可以进行局部光滑，此处就可以采用缺失值处理的方法，如在箱中使用均值、中位数等。

- **回归**

分箱是通过局部有序数据进行光滑，而回归是通过全局数据进行光滑，使用一个函数拟合来光滑数据。常用的有线性回归和多远线性回归，此处需要注意数据拟合的三种状态，欠拟合、正拟合和过拟合。

- **离群点分析**

可以通过无监督学习的方式检测离群点，如聚类将类似的值聚成簇，落在簇外的就可以被视为离群点。

数据集成

实体识别

数据集成的主要目的是合并多个数据存储的数据时，较少数据冗余和不一致。匹配多个数据源的模式，本质上就是一个实体识别问题。

比如如何确定两份数据的user_id，是同一个属性，每个属性的元数据包含名字、含义、数据类型和属性值域，以及NULL、空值规则等。而且在集中过程中，需要特别注意数据结构。

冗余和相关

这里冗余主要是属性冗余，如果一个属性是可以通过另外的属性“变换”出，那么这个属性就是冗余的，这种冗余可以通过相关分析检出。

- 对于标量数据，可以通过χ^2（卡方）检验。

假设X有m个不同值$x_1, x_2, \cdots, x_m$，Y有n和不同值$y_1, y_2, \cdots, y_n$。令(X_i, Y_j)为属性X取x_i和属性Y取y_j的联合事件，则χ^2（卡方）统计量为：

$\chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{(o_{ij}-e_{ij})^2}{e_{ij}}$

其中o_{ij}表示联合事件(X_i, Y_j)的观测频度，e_{ij}表示(X_i, Y_j)期望频度，为

$e_{ij} = \frac{1}{n} count(X = x_i) * count(Y = y_j)$

- 对于数值数据，可以使用相关系数和协方差检验。

相关系数计算公式为

$r(X, Y) = \frac{\sum_{i=1}^{n}(x_i-\overline{X})(y_i-\overline{Y})}{n\sigma_X\sigma_Y}$

协方差$cov(X, Y)$是方差的泛化形式，定义为

$cov(X, Y) = E(X - \overline{X})(Y - \overline{Y}) = \frac{\sum_{i=1}^{n}(x_i-\overline{X})(y_i-\overline{Y})}{n}$

和相关系数放在一起，就变成了

$r(X, Y) = \frac{cov(X,Y)}{\sigma_X\sigma_Y}$

重复元组

冗余和相关是检查属性，重复是检查元组，这对于设置数据库的相关键非常重要。

数据归约

维归约

维规约是减少需要考虑的属性个数，该方法将源数据投影到更小的空间内。此处主要聊三种方式，分别是属性子集选择、小波变换和主成分分析。

- **属性子集选择**

属性子集选择是通过删除与分析目的不相关或冗余的属性，使得分析目的更容易实现或理解。如何选择出最好的子集，就是此处我们需要重点考虑的问题，“最好”通常使用统计的显著性检验来确定，这里不详细描述假设检验，直接提供4中子集选择方法。

1. 向前选择，该过程从空属性集开始，每次迭代将原属性中最好的属性加入集合中，最终选择最优属性集合。
2. 向后删除， 和向前选择相反，该过程从全集开始，每次迭代从原属性中选择最差的属性从集合中删除，最终留下的就是最优属性集合。
3. 向前选择和向后删除组合， 高过程就是前面两种方法的组合，每一次迭代，选择最优的属性加入集合中，也从集合中删除最差的属性。
4. 决策树归纳， 决策树最开始的目的是用于分类，它可以在每个节点上选择最好的属性，将数据进行分类，所以可以将出现在树中的属性归约为属性子集。

- **小波变换**

离散小波变换（DWT）是一种信号处理技术，他可以用于多维数据变换，它的主要思想是通过一些留存一些最强的小波系数，以保留近似的压缩数据。如用户设定一个阈值，大于这个阈值的小波属性予以保留，小于该属性的值置0，如此可以得到更为系数的数据，如此在小波空间内计算就会变得更高效。该方法不只可用于数据归约，由于它可以光滑数据，所以还可以用于数据噪声处理。

小波变换是一个很大的课题，与其关联的还有傅里叶变化等，此处不展开。

- **主成分分析**

主成分分析（PCA）属于泛因子分析的一种（主成分分析中主成分是原始变量的线性组合，因子分析中原始变量是新因子的线性组合），它是搜索k（$k \leq n$）最能代表数据的n维正交向量，如此，就把原属性投影到来一个更小的属性空间上，使得维规约。该方法与子集选择不同的是，它会创建一个替换原属性集的新属性集，而不是直接在原属性集上选择子集。其主要过程为

1. 规范化输入数据，主要目的是避免较大属性在整个选择过程中权重过大。
2. 计算k个标准正交向量（正交可以理解为低维空间中的垂直），作为规范化输入数据的基，这些向量就是主成分，输入数据是这些主成分的线性组合。
3. 主成分充当了数据的新坐标系，提供了方差信息，理论上当$k = n$时，就能代表全部信息。
4. 对左右成分按照重要性排序，去掉比较弱的成本，保留下来的就是主成本。

与小波变换相比，主成分分析可以更好的处理稀疏数据，而小波变换更适合高维数据。

如下面这种形式

$$a_1 = k_{11} * a_1 + k_{12} * a_2 + \cdots + k_{1n} * a_n$$

$$a_2 = k_{21} * a_1 + k_{22} * a_2 + \cdots + k_{2n} * a_n$$

$$\cdots\cdots$$

$$a_m = k_{m1} * a_1 + k_{m2} * a_2 + \cdots + k_{mn} * a_n$$

数量归约

- **参数方法**

数量规约就是使用较小的数据来替换原数据意义，参数方法就是使用模型来估计数据，使得最终存储只需要存储模型参数，而不是实际数据。一个简单的例子，比如现在有10000个房屋面积、地域、交通及房价的关系数据，在很大程度上，就可以建立一个回归模型来表示这个数据，而不是存储所有数据。如对于x, y两个变量，我们可以建立如下的线性函数以表示它们的关系。

$y = ax + b$

a, b为回归系数，分别是斜率和截距，求解方法常为最小二乘法。

- **非参数方法**

非参数方法不是使用参数来表示原数据，而是通过对数据进行一些特殊的划分以减少原数据。

1. 最常用的就是抽样，抽样方法有很多，比如有放回抽样，无放回抽样，簇抽样，分层抽样等。抽样是一个很高效的方法，它的复杂度只为$O(n)$，而下面介绍的直方图，复杂度却是指数型的。
2. 直方图也是一个很常用的方法，该方法使用对数据分箱的方式来进行数据归约。为来确定箱和属性值的划分，涉及两中规则，分别是等宽（每个箱的宽度区间一致）和等频（每个箱的频度初略估计为一个常数）。
3. 聚类技术也可以用于数据规约，每个簇内的对象相互相似（和直方图中的箱类似），而与其他簇相异。不过用簇代替实际数据，比较依赖数据的性质，比如数据在拓扑结构上就能组织成簇，那该方法就会比较有效。当然，如果数据本身非常离散，不具有局部相似的结构，基本上也就不能进行很好的数量归约。

数据压缩

数据压缩是通过数据变换对原数据进行归约或压缩，前面的维归约和数量归约都可以理解为数据压缩的一种。数据压缩分为无损压缩和有损压缩。

- 无损压缩，原数据能够从压缩过的数据重构而得，同时不损失信息。
- 有损压缩，只能近似重构原数据。

数据变换

规范化

数据变换的主要目的是将数据加工成容易分析挖掘的形式，除规范化外，还有数据光滑，特征工程，数据分组和离散化等。如前面提到的分箱、回归都是在进行数据光滑，特征工程的内容很丰富，此处不展开。

这里介绍两种最常用的规范化方法，分别是归一化和标准化。

- **归一化**

顾名思义，就是把原数据规范到固定的0-1区间内进行分析，就像前面介绍主成分分析时候，为了避免数值较大属性对维规约的影响，就会对原数据进行规范化处理。

归一化计算公式：

$$x_i(new) = \frac{x_i - min(X)}{max(X) - min(X)}$$

- **标准化**

我们实际生活中很多数据都近似服从正态分布（具体原因可参考大数定律），所以我们可以使用正态分布的相关参数对数据进行规范化。即统计学中的z统计量。定义如下

$$x_i(new) = \frac{x_i - \mu}{\sigma}$$

离散化

原数据使用区间标签和概念标签替换，这些标签可以递归的组织成更高维的概念，最终形成概念分成，达到数据分析的目的。

前面提到的分箱、直方图、聚类、决策树归纳及相关分析都属于离散化技术，形成更高的概念分层。

完～