Data Types

- Tabular data : matrices, vectors, objects, relations, etc.
  - Data objects : also called samples, examples, instances, data points, objects, tuples, vectors
  - Attributes : each row of a table, also called dimensions, features, variables
- Graphical data : networks, graphs, etc.
- Multi-media data : texts, images, videos, audios, etc.

表格型数据

图数据

多媒体数据.

Types of Attributes

- Discrete : $\mathbf{x} \in$ some countable sets, e.g., $\mathbb{N}$
  - Nominal : Countries={China, US, UK, France, Germany}, Universities={Peking U, Tsinghua U, SUSTech, Shenzhen U, HIT}, not comparable
  - Boolean : 0 or 1, male or female, spam or non-spam, etc.
  - Ordinal : Heights={tall, short}, Scores={A+, A, A-, B+, B, B-, C, C-, D, F}, can be compared, but cannot operated arithmetically
- Continuous : $\mathbf{x} \in$ some subset in $\mathbb{R}^n$
  - Numerical : Income, exact marks, weights, etc., can be operated arithmetically

离散：$x \in$ 可数集

名义型

称型

等级型

连续：$x \in R^n$ 的子集

数字型

Basic Statistics

- Mean : $EX = \min_c \mathrm{E}(X-c)^2 \approx \frac{1}{n}\sum_{i=1}^n x_i$
- Median :

$$\min_c \mathrm{E}|X-c| = \begin{cases} x_{(\frac{n+1}{2})} & \text{if } n \text{ is odd} \\ (x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)})/2 & \text{otherwise} \end{cases}$$

- Maximum : $\max_i x_i$; Minimum : $\min_i x_i$
- Quantile : a generalization of median, k-th q-quantile x_q : $P[X < x_q] \leqslant k/q$; interquartile range (IQR)=$Q_3(75\%) - Q_1(25\%)$
- Variance : $\mathrm{Var}(X) = \mathrm{E}[X - EX]^2 \approx \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2$; Standard deviation : $\sqrt{\mathrm{Var}(X)}$
- Mode : $\min_c \mathrm{E}|X-c|^0 =$ the most frequently occurring value (define $0^0 = 0$)

均值

中位数

最值

分位点

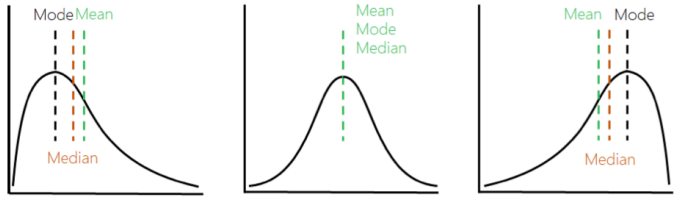
方差

标准差

众数　　零阶矩定义为：0时取0，其他取1.

Central Tendency

For one-peak skewed density distribution, empirical formula :
$Mean - Mode = 3 \times (Mean - Median)$



Positive skewed　　Symmetric　　Negative skewed

经验公式，大约满足

均值与众数之差等于3倍的均值与中位数之差

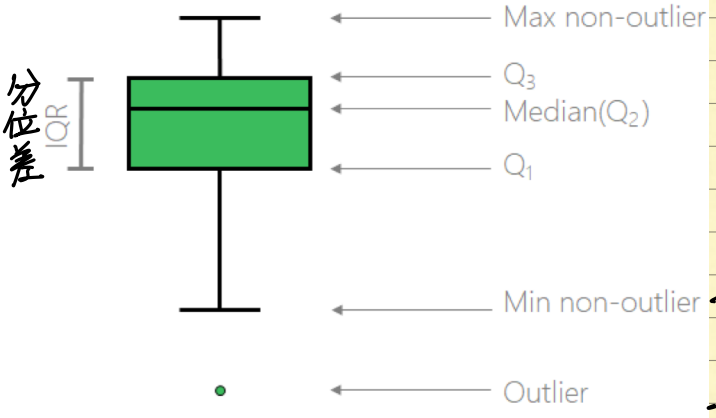
是正偏还是负偏取决于 Mean-Mode

Mean-Mode > 0　正偏 → 峰在 Median 左边

Mean-Mode < 0　负偏 → 峰在 Median 右边

Box Plot 箱形图

Measure the dispersion of data



最大非异常值
3/4分位数
中位数
1/4分位数
最小非异常值
异常值

Metrics 度量

- Proximity :
  - Similarity : range is [0, 1]
  - Dissimilarity : range is $[0, \infty]$, sometimes distance
- For nominal data, $d(\mathbf{x}_i, \mathbf{x}_j) = \frac{\sum_k I(x_{i,k} \neq x_{j,k})}{p}$; or one-hot encoding into Boolean data
- For Boolean data, symmetric distance $d(\mathbf{x}_i, \mathbf{x}_j) = \frac{r+s}{q+r+s+t}$ or Rand index $Sim_{Rand}(\mathbf{x}_i, \mathbf{x}_j) = \frac{q+t}{q+r+s+t}$; non-symmetric distance $d(\mathbf{x}_i, \mathbf{x}_j) = \frac{r+s}{q+r+s}$ or Jaccard index $Sim_{Jaccard}(\mathbf{x}_i, \mathbf{x}_j) = \frac{q}{q+r+s}$

相似度
不相似度/距离

i=1, j=1 的样本个数

Sample i \ Sample j	1	0	sum
1	q	r	$q+r$
0	s	t	$s+t$
sum	$q+s$	$r+t$	p → 所有样本个数

相关系数可看作是一种相似度

名义型data: $\frac{\text{各维度中不同的个数}}{\text{总个数}}$ 或者用独热码转成布尔型!

布尔型data：通过表格做

不对称的距离/相似度，将i=0, j=0丢掉

对布尔型data来说，只有0和1两种取值；若事实上不只两种取值而将肯定的定义为1，否则为0. 此时同时取0不能确认两者的相似但都取1时可以.

对称距离不满足三角不等式(假距离)，而不对称距离满足
但对称距离也很常用.

Distance 的性质

- Positive definiteness $d(\mathbf{x}_i, \mathbf{x}_j) \geqslant 0$ and "=" if and only if $i = j$;
- Symmetry $d(\mathbf{x}_i, \mathbf{x}_j) = d(\mathbf{x}_j, \mathbf{x}_i)$;
- Triangle inequality $d(\mathbf{x}_i, \mathbf{x}_j) \leqslant d(\mathbf{x}_i, \mathbf{x}_k) + d(\mathbf{x}_k, \mathbf{x}_j)$

正定性
对称性
三角不等式

Jaccard 相似度:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \in (0,1)$$

$$J_\delta(A,B) = 1 - J(A,B) = \frac{|A \Delta B|}{|A \cup B|} \quad (A \Delta B = (A \setminus B) \cup (B \setminus A)) \quad \text{symmetric difference}$$

Boolean Vectors: $\vec{x} = (x_1, x_2, \cdots, x_n)$, $\vec{y} = (y_1, y_2, \cdots, y_n)$

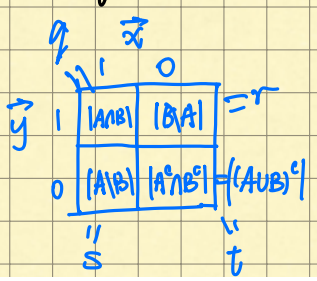
$\vec{x} = 1_A = \begin{cases} x_i = 1, & \text{if } i \in A \\ x_i = 0, & \text{if } i \notin A \end{cases}$, $\vec{y} = 1_B$ indicator function, $A, B \subset \{1, \cdots n\}$

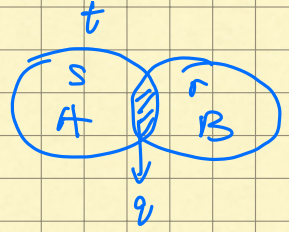
$$J_\delta(\vec{x}, \vec{y}) = J_\delta(A,B) = \frac{r+s}{r+s+q}$$

由于A,B能完全描述 $\vec{x}, \vec{y}$, 故能直接替代

$\{x_i = 1, y_i = 1\} \equiv \{i \in A \cap B\}$

$\{x_i = 0, y_i = 0\} \equiv \{i \in (A \cup B)^c\}$





- Example : Let $H = F = 1$ and $L = S = 0$, $d(LandRover, Jeep) = \frac{1+0}{4+1+0} = 0.20$, $d(LandRover, TOYOTA) = \frac{3+1}{1+3+1} = 0.80$, $d(Jeep, TOYOTA) = \frac{3+2}{1+3+2} = 0.83$

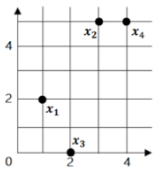
	Weight	Price	Acceleration	MPG	Quality	Sales Volume
Land Rover	H	H	F	H	H	L
Jeep	H	H	S	H	H	L
TOYOTA	L	L	F	L	H	H

Land Rover	Jeep: 1	Jeep: 0
1	$q=4$	$r=1$
0	$s=0$	$t=1$

Distance

- Minkowski distance : $d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt[h]{\sum_{k=1}^{p} |x_{ik} - x_{jk}|^h}$ is L_h-norm
  - Manhattan distance : $h = 1$, and $d(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=1}^{p} |x_{ik} - x_{jk}|$
  - Euclidean distance : $h = 2$, and $d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{k=1}^{p} |x_{ik} - x_{jk}|^2}$
  - Supremum distance : $h = \infty$, and $d(\mathbf{x}_i, \mathbf{x}_j) = \max_{k=1}^{p} |x_{ik} - x_{jk}|$

Point	Attr 1	Attr 2
x_1	1	2
x_2	3	5
x_3	2	0
x_4	4	5



L_1	x_1	x_2	x_3	x_4
x_1	0			
x_2	5	0		
x_3	3	6	0	
x_4	6	1	7	0

(a) Manhattan

L_2	x_1	x_2	x_3	x_4
x_1	0			
x_2	3.61	0		
x_3	2.24	5.1	0	
x_4	4.24	1	5.39	0

(b) Euclidean

L_∞	x_1	x_2	x_3	x_4
x_1	0			
x_2	3	0		
x_3	2	5	0	
x_4	3	1	5	0

(c) Supremum

曼哈顿：每个分量的取值差.

欧氏：直接连线

最大：比较所有分量差，取最大的.

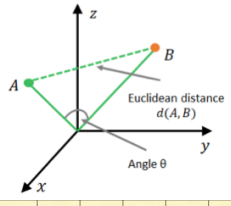
Cosine Similarity 余弦相似度（不是距离）

- Definition : $\cos(\mathbf{x}_i, \mathbf{x}_j) = \frac{\sum_{k=1}^{p} x_{ik} x_{jk}}{\sqrt{\sum_{k=1}^{p} x_{ik}^2} \sqrt{\sum_{k=1}^{p} x_{jk}^2}} = \frac{\mathbf{x}_i \cdot \mathbf{x}_j}{\|\mathbf{x}_i\| \|\mathbf{x}_j\|}$
- Example : $\cos(\mathbf{x}_1, \mathbf{x}_2) = 0.94$

Instance	Team	Coach	Hockey	Baseball	Soccer	penalty	Score	Win	Loss	Season
Instance1	5	0	3	0	2	0	0	2	0	0
Instance2	3	0	2	0	1	1	0	1	0	1

Euclidean vs. Cosine :

- Euclidean : measures the distance in absolute value, many applications
- Cosine : insensitive to absolute value, e.g., analyze users' interests based on movie ratings



欧氏距离：绝对距离

余弦相似度：相对性

若 A, B 共线，但距原点长度不一，欧氏有明显不一样了.

Other Distances

- For ordinal data, mapping the data to numerical data : $X = \{x_{(1)}, x_{(2)}, \ldots, x_{(n)}\}$, $x_{(i)} \mapsto \frac{i-1}{n-1} \in [0, 1]$
- For mixed type, use weighted distance with prescribed weights :

$$d(\mathbf{x}_i, \mathbf{x}_j) = \frac{\sum_{g=1}^{G} w_{ij}^{(g)} d_{ij}^{(g)}}{\sum_{g=1}^{G} w_{ij}^{(g)}}$$

Put the attributes of the same type into groups, for each data type g, use the corresponding distance $d_{ij}^{(g)}$

等级型data：排好序后作映射：

$x_{(i)} \mapsto \frac{i-1}{n-1} \in [0,1]$

转成数值型再处理.

混合型data：不同类型的坐标分量分别计算距离，然后加权平均即最终结果.

Reason for Data Preprocessing

- Missing values
- Noisy with outliers
- Inconsistent representations
- Redundancy
- Errors may come during data input, data gathering, and data transferring
- Errors occur in about 5% of the data

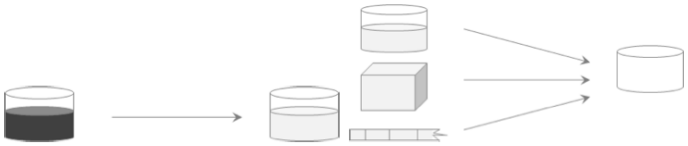
处理缺失值、离群值的噪音

处理非连续的表达

数据冗杂

5%的数据中都有错误.

4 types of Data Preprocessing



(a) Data cleaning (b) Data integration

(c) Data conversion (d) Data reduction

(a) 清洗

(b) 集成

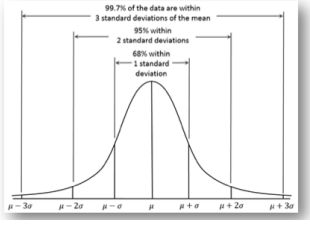
(c) 转化

(d) 约化

Data Scaling 数据标准化

- Why scaling :
  - For better performance : e.g., RBF in SVM and penalty in Lasso/ridge regression assume the zero mean and unit variance
  - Normalize different dimensions : many algorithms are sensitive to the variables with large variances, e.g., height (1.75m) and weight (70kg) in distance calculation
- Z-score scaling : $x_i^* = \frac{x_i - \hat{\mu}}{\hat{\sigma}}$, $\hat{\mu}$: sample mean, $\hat{\sigma}$: sample variance, applicable if max and min are unknown and the data distributes well

Z-score标准化



- 0-1 scaling : $x_i^* = \frac{x_i - \min_k x_k}{\max_k x_k - \min_k x_k} \in [0,1]$, applicable for bounded data sets, need to recompute the max and min when new data arrive
- Decimal scaling : $x_i^* = \frac{x_i}{10^k}$, applicable for data varying across many magnitudes
- Logistic scaling : sigmoid transform $x_i^* = \frac{1}{1+e^{-x_i}}$, applicable for data concentrating nearby origin

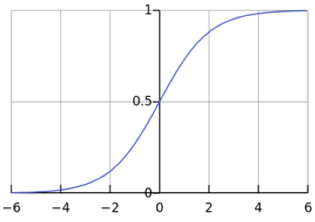
0-1标准化(min-max标准化)

$[a,b] \rightarrow f_i' = \frac{b-a}{f_{max} - f_{min}}(f_i - f_{min}) + a$

小数定标标准化：移动小数点

Logistic标准化：使数据分布在(0,1)

会消除掉原本的分布



Data Discretization 数据离散化

- Why discretization :
  - Improve the robustness : removing the outliers by putting them into certain intervals
  - For better interpretation
  - Reduce the storage and computational power
- Unsupervised discretization : equal-distance discretization, equal-frequency discretization, clustering-based discretization, 3σ-based discretization
- Supervised discretization : information gain based discretization, χ^2-based discretization

消除异常值
增加解释性
减少存储和计算

可以在一定程度上对数据进行抽象处理

节省空间，提高效率，消除异常值

Unsupervised Discretization 无监督离散化

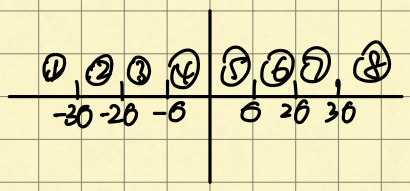
- Equal-distance discretization : split the range to n intervals (bins) with the same length, group the data into each bin, sensitive to outliers
- Equal-frenquency discretization : group the data into n subset so that each subset has the same number of points, tend to separate samples with similar values and produce uniform distribution
- Clustering-based discretization : do hierarchical clustering and form a hierarchical structure (e.g., using K-Means), and put the samples in the same branch into the same interval (a natural example is family tree)
- 3σ-based discretization : put the samples into 8 intervals, need to take logarithm first

等距离散化：均匀分成n个区间
对离群值敏感

等频离散化：每个区间都有相同的样本量、均匀分布
易使相邻区间段内的数据有相同特性。

聚类离散化：将样本分到不同的类中
每个类的平均值当作这类所有样本的值

3σ离散化：将样本划分到8个区间，需先取对数

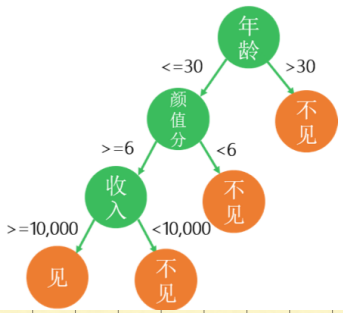


Supervised Discretization 有监督离散化

Information gain 信息增益

- Top-down splitting, similar to create a decision tree
- Do a decision tree classification using information gain, find a proper splitting point for each continuous variable such that the information gain increases the most
- The final leaf nodes summarize the discrete intervals

自上而下进行，类似于用决策树做分类。
每次分裂节点选取的是信息增益增长最快的。
叶子节点即各个离散的区间。



ChiMerge 卡方离散化

- Bottom-up : similar to hierarchical clustering
- $\hat{\chi}^2$ statistics proposed by Karl Pearson, is used to test whether the observations dramatically deviate from theoretical distribution : $\hat{\chi}^2 = \sum_{i=1}^{k} \frac{(A_i - \mathbb{E}A_i)^2}{\mathbb{E}A_i} = \sum_{i=1}^{k} \frac{(A_i - np_i)^2}{np_i}$, where n_i is the number of samples in the i-th interval $A_i = [a_{i-1}, a_i]$ (frequency of observations), $\bigcup_{i=1}^{k} A_i$ covers the range of the variable, and $\mathbb{E}A_i = np_i$ is its expectation computed from the theoretical distribution ; it can be shown that $\hat{\chi}^2 \to \chi^2_{k-1}$
- ChiMerge : Given a threshold level t,
  1. Treat each value of the continuous variable as an interval and sort them in increasing order ;
  2. For each pair of adjacent intervals, compute its $\hat{\chi}^2$ statistics, if $\hat{\chi}^2 < t$, merge them into a new interval ;
  3. Repeat the above steps until no adjacent intervals can be merged.
- Two shortcomings : t is hard to set appropriately ; too long loop for large sample set, computationally intensive

自下而上. 类似于层次聚类.

将打散的数据根据 χ^2 合并.

$\{\vec{X}\}_{i=1}^{n}$, 看 $x_{i,A}$, $x_{i,B}$ 的相关性

$A_1 = a_1$, $A_2 = a_2$, ..., $A_m = a_m$

$B_1 = b_1$ $\#\{(a_1, b_1)\} = C_{11}$ ... $\#\{(a_m, b_1)\} = C_{m1}$ $n_{\cdot 1} = \sum_{i=1}^{m} C_{i1}$

$B_2 = b_2$ ⋮ ⋮ $n_{\cdot j} = \sum_{i=1}^{m} C_{ij}$

$B_p = b_p$ C_{ij} = number of samples taking $\{(a_i, b_j)\}$ $= \#\{B = b_j\}$

$n_{1\cdot}$ ... $n_{i\cdot} = \sum_{j=1}^{p} C_{ij} = \#\{A = a_i\}$ $\#\{B=b_j\} \approx P(A=a_i)$

$\chi^2_{AB} = \sum_{i=1}^{m}\sum_{j=1}^{p} \frac{(C_{ij} - e_{ij})^2}{e_{ij}}$ 假设为独立分布: $e_{ij} = \frac{n_{\cdot j} \cdot n_{i\cdot}}{n}$ → 样本总数

取 C_{ij} 的可能性

$\approx E(\#\{(a_i, b_j)\}) = E(C_{ij})$

threshold t

若 $\chi^2_{AB} > t$, 则认为 A, B 相关;

否则无关.

$n \to \infty$ $\chi^2_{n-1} = \sum_{k=1}^{n-1} \chi_k^2$, $\chi_k^2 \sim N(0,1)$

A, B independent

若 A, B 独立, $\hat{\chi}^2 \in (E\chi^2 - 6, E\chi^2 + 6)$

是否合并区间 分类结果

A: merge interval ; B: classification

合并区间应不影响分类结果

$\chi^2_{AB} < t \Leftrightarrow$ uncorrelated $\Rightarrow$ 可以合并区间

Iris Data Example

- $\hat{\chi}^2 = \sum_{i=1}^{m}\sum_{j=1}^{k} \frac{(A_{ij} - E_{ij})^2}{E_{ij}}$, where
  $m = 2$ (two adjacent intervals)
  k is the number of classes
  A_{ij} is the number of samples in i-th interval and in class k
  $R_i = \sum_{j=1}^{k} A_{ij}$ is the total number of samples in i-th interval
  $C_j = \sum_{i=1}^{m} A_{ij}$ is the total number of samples in class j
  $N = \sum_{i=1}^{m}\sum_{j=1}^{k} A_{ij}$ is the total number of samples
  $E_{ij} = R_i \cdot \frac{C_j}{N}$
- χ^2 of 4.3 and 4.4 : $C_1 = 4$, $C_2 = 0$, $C_3 = 0$, $N = 4$, $A_{11} = 1$, $A_{12} = A_{13} = 0$, $A_{21} = 3$, $A_{22} = A_{23} = 0$, $R_1 = 1$, $R_2 = 3$, $E_{11} = 1$, $E_{12} = E_{13} = 0$, $E_{21} = 3$, $E_{22} = E_{23} = 0$, $\hat{\chi}^2 = 0$.

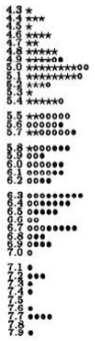


FIGURE: Sepal lengths of 3 types of iris

ChiMerge Result

Left : significance level is 0.5 and the threshold for χ^2 is 1.4 ; Right : significance level is 0.9 and the threshold for χ^2 is 4.6 ; The final results keep the intervals with χ^2 larger than the thresholds

Int	Class frequency			χ^2
4.3	16	0	0	4.1
4.9	4	1	1	2.4
5.0	25	5	0	8.6
5.5	2	5	0	2.9
5.6	0	5	1	1.7
5.7	2	5	1	1.8
5.8	1	3	3	2.2
5.9	0	12	7	4.8
6.3	0	6	15	4.1
6.6	0	2	0	3.2
6.7	0	5	10	1.5
7.0	0	1	0	3.6
7.1	0	0	12	

Int	Class frequency			χ^2
4.3	45	6	1	30.9
5.5	4	15	2	6.7
5.8	1	15	10	4.9
6.3	0	14	25	5.9
7.1	0	0	12	

Figure 2: ChiMerge discretizations for *sepal-length* at the .50 and .90 significance levels ($\chi^2 = 1.4$ and 4.6)

Data Redundancy 数据冗余

- When strong correlations exist among different attributes, then we say that the some attributes can be derived from the others (Recall linear dependency for vectors)
- E.g., two attributes "Age" and "Birthday", then "Age" can be calculated from "Birthday"
- Determine the data redundancy by correlation analysis
- For continuous variables A and B, compute the correlation coefficient $\rho_{A,B} = \frac{\sum_{i=1}^{k}(a_i-\bar{A})(b_i-\bar{B})}{k\hat{\sigma}_A\hat{\sigma}_B} \in [-1,1]$:
  1. If $r > 0$, A and B are positively correlated ;
  2. If $r < 0$, A and B are negatively correlated ;
  3. If $r = 0$, A and B are uncorrelated.

  Note that the correlation between A and B does not imply the causal inference.
- For discrete variables A and B, compute the χ^2 statistics : small $\hat{\chi}^2$ value implies small correlation

若不同属性之间强相关的话，则一类可由另一类推出.

相关系数

$$\rho_{AB} = \frac{Cov(A,B)}{\sqrt{Var(A)}\sqrt{Var(B)}}$$

对离散数据，越小的 χ^2 值有更小的相关性.

Missing Data

- Where missing data come from ?
  - Missing Completely At Random (MCAR) : the occurrence of missing data is a random event
  - Missing At Random (MAR) : depending on some control variables, e.g., the age > 20 is not acceptable in an investigation for teenager and thus is replaced by MAR
  - Missing Not At Random (MNAR) : missing data for bad performed employees after they are fired

Simple Methods

- Deleting samples : for small size of samples with missing values
- Deleting variables : for series missing values in variables

删除样本 (删除行)

删除特征 (删除列)

gradyear	gender	age	friends
2006	M	18.98	7
2006	F	18.801	0
2006	M	18.335	69
2006	F	18.875	0
2006	NA	18.995	10
2006	F		142
2006	F	18.93	72
2006	M	18.322	17
2006	F	19.055	52
2006	F	18.708	39
2006	F	18.543	8
2006	F	19.463	21
2006	F	18.097	87
2006	NA		0
2006	F	18.398	0
2006	NA		0
2006	NA		135
2006	F	18.987	26
2006	F	17.158	27
2006	F	18.497	123
2006	F	18.738	35

Filling Methods

- Filling with zero
- Filling with means for numerical type, and with modes for non-numerical type, applicable for MCAR ; drawback : concentrating in the mean and underestimating the variance ; solution : filling in different groups
- Filling with similar variables : auto-correlation is introduced
- Filling with past data
- Filling by K-Means : Compute the pairwise distances of the data using good variables (no missing values), then fill the missing values with the mean of the first K most similar good data, auto-correlation is introduced
- Filling with Expectation-Maximization (EM) : introduce hidden variables and use MLE to estimate the parameters (missing values)
- Random filling :
  - Bayesian Bootstrap : for discrete data with range $\{x_i\}_{i=1}^k$, randomly sample $k-1$ numbers from $U(0,1)$ as $\{a_{(i)}\}_{i=0}^k$ with $a_{(0)}=0$ and $a_{(k)}=1$; then randomly sample from $\{x_i\}_{i=1}^k$ with probability distribution $\{a_{(i)}-a_{(i-1)}\}_{i=1}^k$ accordingly to fill in the missing values
  - Approximate Bayesian Bootstrap: Sample with replacement from $\{x_i\}_{i=1}^k$ to form new data set $X^*=\{x_i^*\}_{i=1}^{k^*}$; then randomly sample n values from X^* to fill in the missing values, allowing for repeatedly filling missing values
- Model based methods : treat missing variable as y, other variables as $\mathbf{x}$; take the data without missing values as our training set to train a classification or regression model ; take the data with missing values as our test set to predict the miss values

填零

填均值(数型值)/众数(非数值型):
会改变分布. 聚集于均值/众数

相似变量填补(热补填补): 强相关数据的互补
过往数据填补(冷平台填补)
K-Means填补: K个最相关有效数据的均值填补

EM填补: 引入隐藏变量, 用MLE估计.

随机填补: 有样本$\{x_i\}_{i=1}^n$. 有k个非缺失值和(n-k)缺失值.
Bayesian Bootstrap: 从均匀分布$U(0,1)$取k-1随机数.
排序记为$\{a_{(i)}\}_{i=1}^{k-1}$且有$a_{(0)}=0, a_{(k)}=1$
从非缺失值中$\{x_i\}_{i=1}^k$以概率$\{a_{(i)}-a_{(i-1)}\}_{i=1}^k$采样一个值填补, 重复(n-k)次.
Approximate Bayesian Bootstrap: 从$\{x_i\}_{i=1}^k$中有放回地抽取k个值建立新集合$\{x_i^*\}_{i=1}^{k^*}$. 并分别从中随机抽取填到(n-k)个缺失值中.

模型填补: 缺失值为y, 非缺失值为x.
通过模型对x处理输出y.

Filling by Interpolation 插值填补

- For the data of numeric type, each attribute (column vector) can be viewed as the function values $z_i=f(x_i)$ at the points x_i, where x_i is a reference attribute (the reference attribute usually has no missing values, it can be chosen as the index)
- We can interpolate a function f using the existing values (x_i, z_i), and then fill in the missing values z_k with $f(x_k)$
- Linear interpolation : treat $z=f(x)$ as linear function between the neighboring points x_{k-1} and x_{k+1} of x_k
- Lagrange interpolation : interpolate the $m+1$ existing values $\{(x_{l_i}, z_{l_i})\}_{i=1}^{m+1}$ by a degree m polynomial $L_m(x)$

```
gen_data.interpolate()
```

	feature1	feature2	feature3
1	1.728534	-0.371519	1.451700
2	0.795975	-1.067026	-1.861944
3	-0.030449	-0.050409	1.299994
4	-0.856872	0.966208	0.987861

Missing value

Missing value

Special Values and Dummy Variables

对离散型特征
将缺失值当作单独取值处理

- In Python, "np.nan" means missing values (Not a Number, missing float value)
- "None" is a Python object, used to represent missing values of the object type
- Dummy variables : e.g., missing values in gender ("Male" or "Female"), then define a third value "unknown" for the missing values

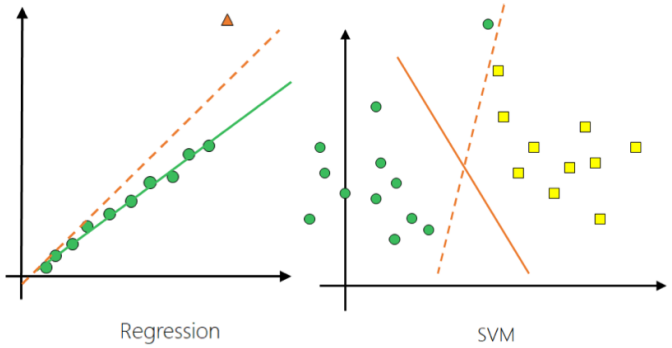
```
import pandas as pd
import numpy as np

teenager_sns = pd.read_csv('teenager_sns.csv')

print teenager_sns['gender'].value_counts()

teenager_sns['gender'] = teenager_sns['gender'].replace(np.NaN, 'unknown')

print ""
print "哑变量方法处理后: \n"
print teenager_sns['gender'].value_counts()
```

```
F    22054
M     5222
Name: gender, dtype: int64

哑变量方法处理后:

F          22054
M           5222
unknown     2724
Name: gender, dtype: int64
```

Outliers

某些点分布与大部分不同.

- Outliers : the data points seem to come from different distribution, or noisy data
- Outlier detection : unsupervised, e.g., Credit cheating detection, medical analysis, and information security, etc.

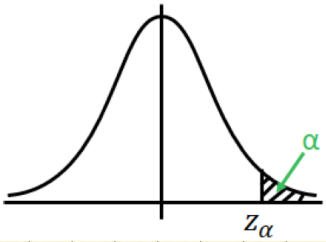


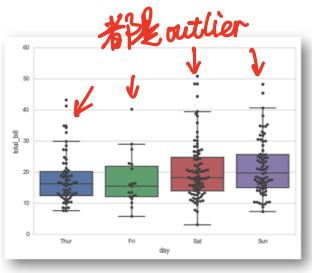
Outlier Detection

Statistics Based Methods

上、下α分位点以外的值

- The samples outside the upper and lower α-quantile for some small α (usually 1%)
- Observe from box plot

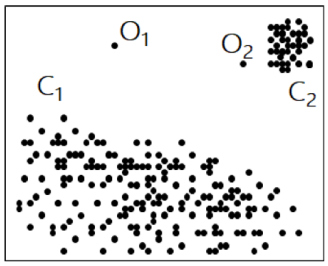




Local Outlier Factor (LOF) 局部离群因子

Local Outlier Factor (LOF) is a density based method :

1. We could compute the density at each position $\mathbf{x}$, e.g., $p(\mathbf{x})$ (how to define the density if we only have data samples) ;
2. We could compare the density of each point $\mathbf{x}$ with the density of its neighbors, i.e., compare $p(\mathbf{x})$ with $p(\mathbf{x}_k)$ where $\mathbf{x}_k$ is close to $\mathbf{x}$ (in a neighborhood of $\mathbf{x}$, but how to define the neighborhood)



周围密度较低的点，一般视作离群值.

Computing Density by Distance

Some definitions :

- $d(A, B)$: distance between A and B ;
- $d_k(A)$: k-distance of A, or the distance between A and the k-th nearest point from A
- $N_k(A)$: k-distance neighborhood of A, or the points within $d_k(A)$ from A ;
- $rd_k(B, A)$: k-reach-distance from A to B, the repulsive distance from A to B as if A has a hard-core with radius $d_k(A)$, $rd_k(B, A) = \max\{d_k(A), d(A, B)\}$; note that $rd_k(A, B) \neq rd_k(B, A)$, which implies that k-reach-distance is not symmetric.

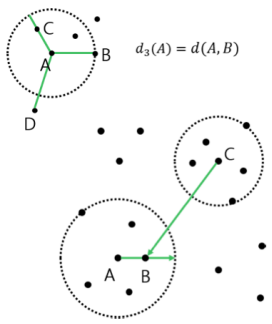


FIGURE: $rd_5(B, A) = d_5(A)$ and $rd_5(B, C) = d(B, C)$

$d_k(A)$：距点A第k近的点的距离

$N_k(A)$：以$d_k(A)$为半径的圆内的点

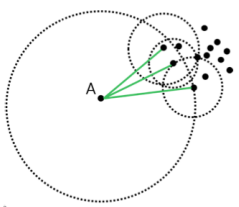
$rd_k(B,A) = \max\{d_k(A), d(A,B)\}$

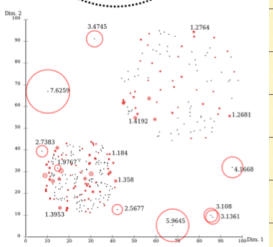
不对称：$rd_k(A,B) \neq rd_k(B,A)$

Local Outlier Factor

Some definitions :

- $lrd_k(A)$: local reachability density is inversely proportional to the average distancep, $lrd_k(A) = 1/\left(\frac{\sum_{O \in N_k(A)} rd_k(A,O)}{|N_k(A)|}\right)$; intuitively, if for most $O \in N_k(A)$, more than k points are closer to O than A is, then the denominator is much larger than $d_k(A)$ and $lrd_k(A)$ is small ; e.g., $k = 3$ in the figure
- $LOF_k(A)$: local outlier factor, $LOF_k(A) = \frac{\sum_{O \in N_k(A)} \frac{lrd_k(O)}{lrd_k(A)}}{|N_k(A)|}$;
- $LOF_k(A) \ll 1$, the density of A is locally higher, dense point ; $LOF_k(A) \gg 1$, the density of A is locally lower, probably outlier





局部可达密度

$$lrd_k(A) = \left(\frac{\sum_{O \in N_k(A)} rd_k(A,O)}{|N_k(A)|}\right)^{-1}$$

局部离群因子

$$LOF_k(A) = \frac{\sum_{O \in N_k(A)} \frac{lrd_k(O)}{lrd_k(A)}}{|N_k(A)|}$$

当$LOF_k(A) \gg 1$时，为离群值

- Other methods for outlier detection :
  - K-Means
  - K Nearest Neighbors
  - Isolation Forest
  - One-class support vector machine
  - Robust covariance
- Outlier processing :
  - Delete outliers (treat them as missing values)
  - Robust regression
  - Theil-Sen regression