Data preprocessing

Data preprocessing

- Data structure and coding
- Statistical methods
  - Definition of sample and feature
  - Sample statistics and feature distance
  - Feature normalization
  - Missing observation and outlier

Data Structure (数据结构)

- 表格数据
  - 关系记录
  - 数据矩阵
  - 向量
  - 事物数据
- 图和网络
  - 万维网
  - 社交网络
  - 分子结构

行星	周期（年）	平均距离	周期^2/距离^3
水星	0.241	0.39	0.98
金星	0.615	0.72	1.01
地球	1.00	1.00	1.00
火星	1.88	1.52	1.01
木星	11.8	5.20	0.99
土星	29.5	9.54	1.00
天王星	84.0	19.18	1.00
海王星	165	30.06	1.00

- 多媒体数据
  - 文本
  - 图像
  - 视频
  - 音频

Data Structure (数据结构)

- Computer Science
  - Float-point number, Array, Linked list, Tree, Graph
  - Function, Lambda Calculus
  - Stack, Queue, Heap
  - Object, File, Thread, Process
- Mathematics
  - Number, Vector, Matrix, Graph
  - Function, Operator
  - Random variable, Probability distribution
  - Time series, Stochastic process
  - Group, Ring, Field,
  - Set, Category

Structured Data (结构化数据)

- “表格”数据
  - 结构化数据也称作行数据，通常由二维表结构来表达
  - Array

- 严格地遵循数据格式与长度规范
  - Categorical data: 收入水平=『贫困，低收入，小康，中等收入，富有』

- How to store and retrieve these data?
  - SQL Database，Big Table

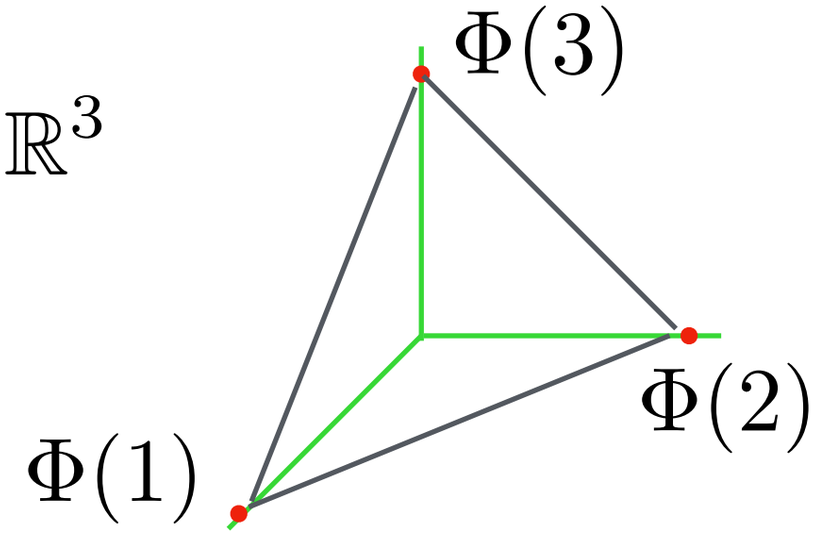
Structured Data (结构化数据)

- How to convert categorical data type for classification and regression problem?
- Integer encoding (ordering! Not good.)
  - red=1, blue=2, green=3
- Vector encoding: One-Hot Encoding
  - 对于每一个特征，如果它有m个可能值，那么经过独热编码后，就变成了m个二元特征。并且，这些特征互斥，每次只有一个激活
  - Color feature：red=(1,0,0), blue=(0,1,0), green=(0,0,1)
  - Extension: dummy encoding (哑变量编码)

Vector encoding

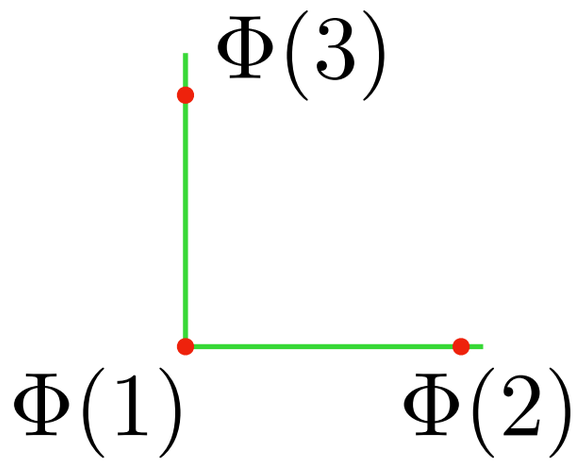
- One-Hot encoding

$$X \in \{1, 2, 3\} \qquad \Phi(X) \in \mathbb{R}^3$$



- Dummy encoding
  - Reduce one dimension

$$X \in \{1, 2, 3\} \qquad \Phi(X) \in \mathbb{R}^2$$



Statistical methods

- 研究怎样用有效的方法去收集和使用带随机性影响的数据。
- Source of randomness in data:
  - When the population is large, need to sample it.
  - When there is un-controlled or un-controllable error in the experiment.
- Effectively to collect data:
  - 抽样理论,试验设计,保证分布无偏。
- Effectively to use the data:
  - Extract information, make inference and decision as accurate and robust as possible.
  - Scientific discovery.

What is Sample?

- Population(总体)
  - 在统计中，常把研究问题中所关心的对象全体称为总体。
- Sample(样本)
  - 由于人力物力的限制，无法对总体所有的对象逐个进行观测调查。总体中若干个个体就称为样本。

What is Sample?

- Informal definition

  通过观测或试验而得到的数据，称为样本，又称样品，子样。

- *In mathematical statistics: the sample consists of finite realizations of a random variable.*

例如：在同一架天平上将一个物体秤n次，得到数据X1,…,Xn，则它们的全体，X=(X1,…,Xn)，称为样本。n称为样本大小。

What is Sample?

- Example of dice

- $X(\omega) \in \{1, \cdots, 6\}$, each ω represents one draw
- A sample consists of two independent draws

$$(X_1, X_2) = (X(\omega_1), X(\omega_2))$$

Statistics of sample

•统计量（statistics)：样本的函数

获得数据总体印象，更好的理解数据

•基本统计量：

•算术平均值（Mean）

•中值（Median）

•最大值（Max）

•最小值（Min）

•分位数（Quantiles）

•方差（Variance）

n is the sample size

- 均值（Mean）

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ 或 } \bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

- 中位数（Median）
  - 当特征值的项数n为奇数时，处于中间位置的特征值即为中位数；当n为偶数时，中位数则为处于中间位置的2个特征值的平均数。
- 众数（Mode）
  - 出现频率最高的值

Mean is much more sensitive to outlier than the median.

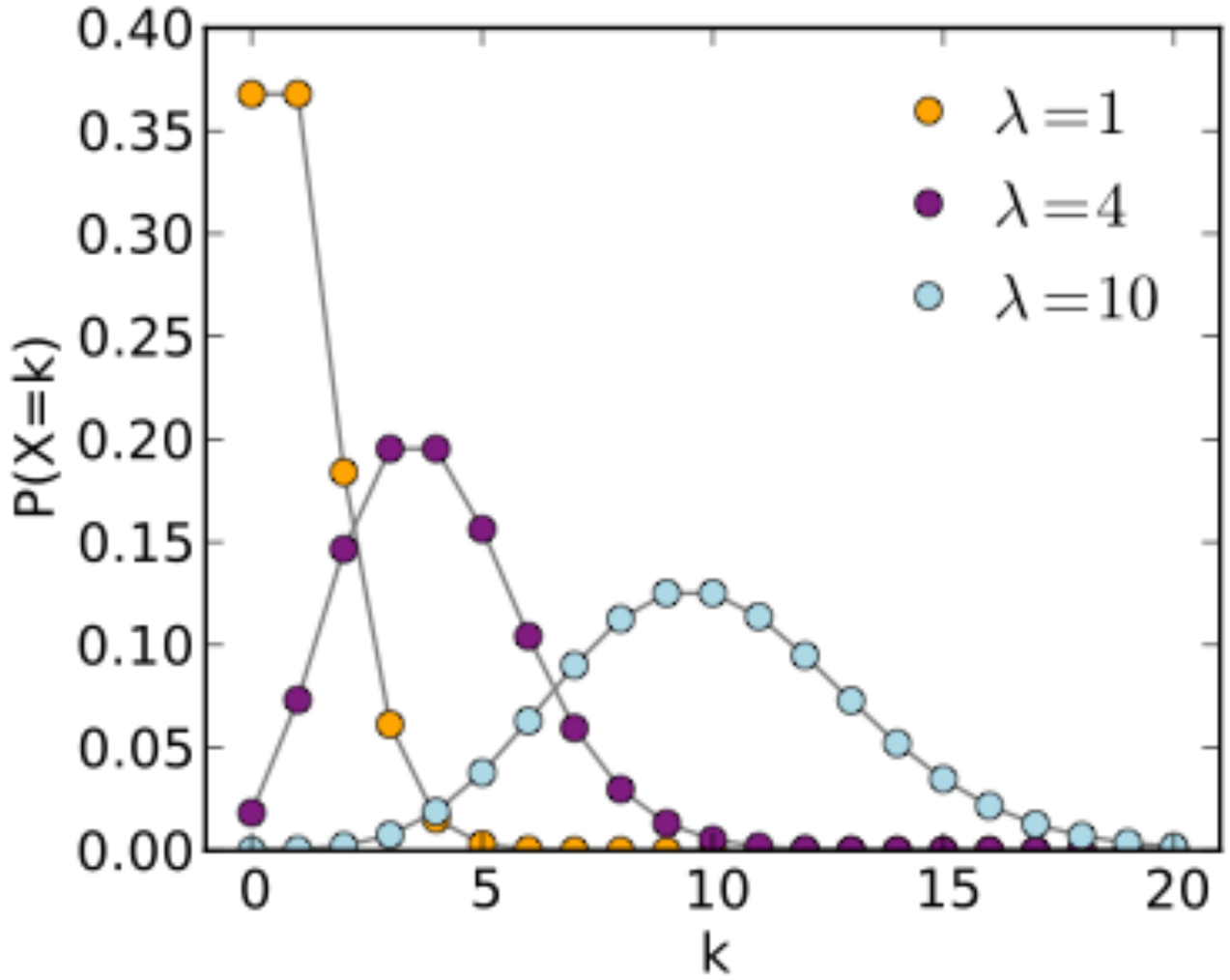
Statistics of sample

- Given x1,…x9, how to estimate their mean?
  - Use (x1 + x2 + x3 + … + x9)/9
  - Sort in increasing order: x(1) <= x(2) <= x(3) <= … <= x(9)
    - Use x(5)
    - or Use [x(1) + x(9)] / 2
    - or Use [x(2) + x(3) + … x(8)] / 7

- Which estimator is better, in what sense?

Statistics of sample

- If X follows Poisson distribution, what is the mean, mode and variance?



$$P(k \text{ events in interval}) = e^{-\lambda} \frac{\lambda^k}{k!}$$

Picture from Wikipedia

What is the sample Feature?

- 表征样本某个属性
- 类似概念：变量（variable）、属性（attribute）
- 例如在汽车性能指数：排气量、车重、油耗

连续型特征

- 特征可以为实数空间任意取值
  - 例如：温度、身高、长度、价格等
- 通常由浮点型表示

离散型特征

- 其值域为有限集或可列集
  - 如果一个集合与自然数集合之间存在一一对应关系，则这个集合称为可列集
  - 例如：汽车品牌、NBA球队等
- 布尔型、等级型、名义型

Geometry in the feature space

- 在很多场景中需要计算样本的距离或相似度
  - 样本是否重复？两个商品是否相似？客户分群？
- 假设$d(\cdot)$为某种距离函数，则通常需要满足以下条件：
  - 距离通常是非负的: $d(x,y) \geq 0$
  - 一个样本与自己的距离为零: $d(x,x) = 0$
  - 距离通常满足对称性: $d(x,y) = d(y,x)$
  - 距离的三角不等式: $d(x,z) \leq d(x,y) + d(y,z)$

Minkowski distance in L_h Space

- 闵可夫斯基距离（Minkowski distance）

$$d(i,j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \cdots + |x_{id} - x_{jd}|^h}$$

$i = (x_{i1}, x_{i2}, \cdots, x_{id})$和 $j = (x_{j1}, x_{j2}, \cdots, x_{jd})$分别代表两个 d 维数据对象， h 为序，上述距离也被称为 L_h 范式

Minkowski distance in L_h Space

- 曼哈顿距离（Manhattan distance）
  - $h = 1$，L_1 范式

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{id} - x_{jd}|$$

- 欧式距离（Euclidean distance）
  - $h = 2$，L_2 范式

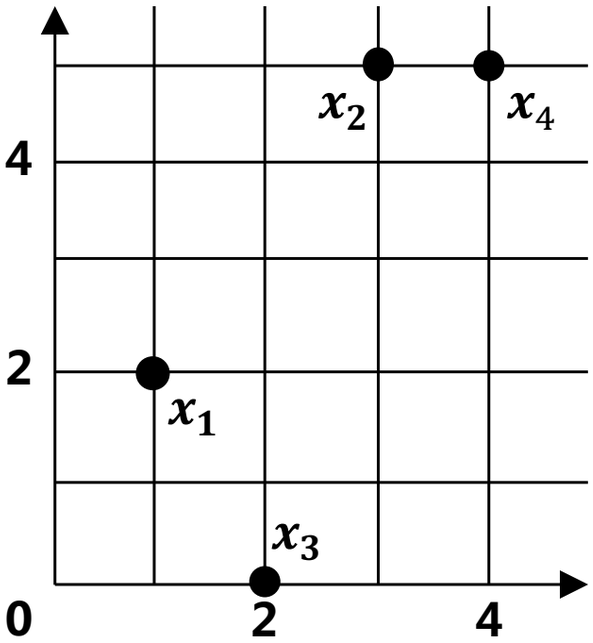
$$d(i,j) = \sqrt{(|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \cdots + |x_{id} - x_{jd}|^2)}$$

- 极大距离（supremum distance）
  - $h = \infty$，L_∞ 范式

$$d(i,j) = \lim_{h\to\infty} \left(\sum_{f=1}^{d} |x_{if} - x_{jf}|^h \right)^{\frac{1}{h}} = \max_{f}^{d} |x_{if} - x_{jf}|$$

Minkowski distance in L_h Space

点集	特征1	特征2
x_1	1	2
x_2	3	5
x_3	2	0
x_4	4	5



- 曼哈顿距离

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

- 欧式距离

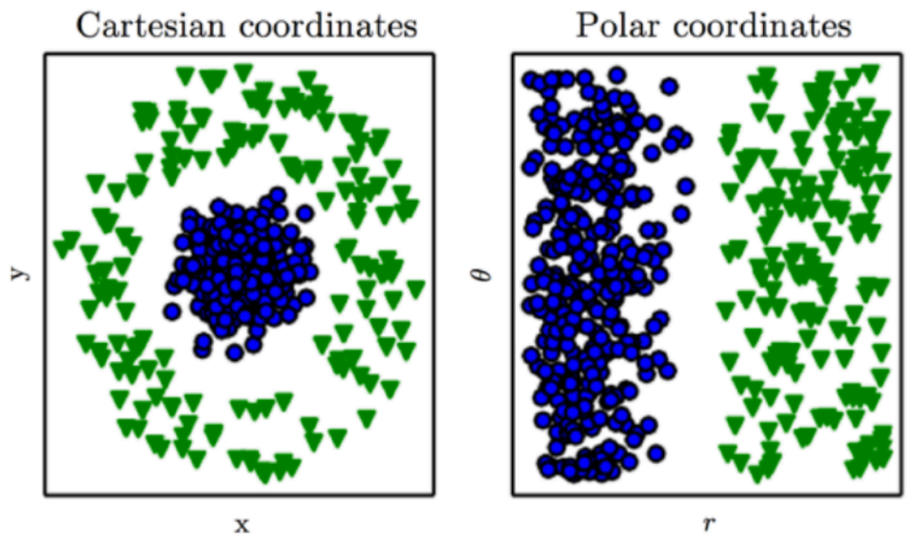
L_2	x_1	x_2	x_3	x_4
x_1	0			
x_2	3.61	0		
x_3	2.24	5.1	0	
x_4	4.24	1	5.39	0

- 极大距离

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

Distance between 2 images?

- Image Search (information retrieval)
- extension: Audio, Text, etc
  - distance between two sentences of two different languages (translation)
- Extract features, compute Euclidean distance / cosine similarity
- Mystery of representation learning (such as deep convolutional neural network)



Why standardization?

- 为什么要进行数据标准化？
  1. 数据分析及建模过程中，许多机器学习算法需要其输入特征为标准化形式。例如，SVM算法中的RBF核函数，线性模型中的L1、L2正则项，目标函数往往假设其特征均值在0附近且方差齐次；
  2. 若样本的特征之间的量纲差异太大，样本之间相似度评估结果将存在偏差
- 常见数据标准化方法：Z-score标准化、Min-Max标准化、小数定标标准化和Logistic标准化

Standardization: Z-Score

Motived by Gaussian distribution

- Assume the feature space is unbounded

  - 对特征取值中的每一个数据点作减去均值并除以标准化的操作，使得处理后的数据具有固定均值和标准差，处理函数为：

$$f_i' = \frac{f_i - \mu}{\sigma}$$

其中，f_i'为标准化后各数据点的取值，f_i为原始各数据点取值，μ为该特征取值的平均值，σ为该特征取值的标准差

  - 适用范围： Z-Score的标准化方法适用于特征的最大值或最小值未知、样本分布非常离散的情况

Standardization: Min-Max

Motived by Uniform distribution

- Assume the feature space is bounded

又称离差标准化或最大-最小值标准化，Min-Max标准化通过对特征作线性变换，使得转换后特征的取值分布在 $[0,1]$ 区间内。其处理函数为：

$$f_i' = \frac{f_i - f_{\min}}{f_{\max} - f_{\min}}$$

其中$f_{\min}$为的最小值，$f_{\max}$为的最大值

- 将特征f映射到 $[a,b]$ 区间内：$f_i' = \frac{b-a}{f_{\max}-f_{\min}}(f_i - f_{\min}) + a$
- 适用范围： 0-1标准化适用于需要将数据简单地变换映射到某一区间中。
- 缺点：当有新数据加入时，可能会导致特征的最大值或最小值发生变化，此时便需要重新定义最大值、最小值，若数据存在离群值，标准化后的效果较差

Why discretization?

- 为什么要将连续型特征进行离散化处理？
  - 算法特征类型有**要求**。如关联规则挖掘算法，ID3决策树算法
  - 为更好地提高算法的**精度**。朴素贝叶斯分类算法的正确率比没有处理的情况平均高出10%

    Dougherty, James, Ron Kohavi, and Mehran Sahami. "Supervised and unsupervised discretization of continuous features." *Machine learning: proceedings of the twelfth international conference*. Vol. 12. 1995.

  - 离散化处理本质是将连续型数据分段，因此数据中的异常值会直接划入相应的区间段中，进而增强了之后模型对于数据异常值的**鲁棒性**
  - 离散化后的特征，其取值均转化为有明确含义的区间号，相对于原始的连续型来说，含义更加明确，从而使得数据的**可解释性更强**，模型更易使用与理解
  - 将连续型特征离散化后，**特征的取值大大减少**，这样一来减少了数据集对于系统存储空间的需求，二来在算法建模中也大大减少了模型的实际运算量，从而可以提升模型的计算效率

Basic idea of discretization

1维情况

- 特征的离散化过程是将连续型特征的取值范围划分为若干**区间段(bin)**，然后使用区间段代替落在该区间段的特征取值。
  - 区间段之间的分割点称之为**切分点(cut point)**
  - 由切分点分割出来的子区间段的个数，称之为**元数 (arity)**

- 假设需要将“年龄”这个连续型特征切分成k个区间段，则需要$(k-1)$个切分点。
  “年龄”特征的取值范围在 $[0,150]$ 之间，通过4个切分点10、25、40和60，将其转化成为5个离散区间段

Basic methods of discretization

Challenge: 高维情况

- 特征离散化目标：在数据信息损失尽量少的前提下，尽可能减少元数
- 按是否参考了数据集的y值信息划分为：

无监督离散化

不参考目标特征 y ，直接根据特征本身的分布特性进行离散化处理

- 等距离散化
- 等频离散化
- 聚类离散化

有监督离散化

利用参考数据集中的目标特征 y ，将连续型特征进行离散化处理

- 信息增益离散化
- ChiMerge离散化等

Missing and outlier observation

- In probability (large deviation) theory, to observe an outlier is a rare event.
  - A rare event can have a big impact.
- A sample who has missing feature can be regarded as an outlier observation.
- But the reason why a feature is missing can be very different.

Why there is missing observation?

- 数据**采集过程**可能会造成数据缺失；
- 数据通过网络等渠道进行**传输时**也可能出现数据丢失或出错，从而造成数据缺失；
- 在数据**整合过程**中也可能引入缺失值。

学生信息数据表

入学年份	性别	年龄	足球	篮球	···	购物	化妆
2012	M	18	0	0	···	0	0
2012	F	18	0	1	···	0	0
2010	M	20	0	1	···	0	0
2012	F	18	0	0	···	0	2
2011	F	18	0	0	···	1	1
2012	F		0	0	···	1	0
2012	F	18	0	0	···	0	0
2011	M	18	2	0	···	0	0
2011	F	19	0	0	···	0	0
2012		18	0	0	···	1	0
2012	F	18	0	0	···	0	0
2011		19	0	1	···	0	0
2012	F	18	0	0	···	0	0
2012	F		0	0	···	0	2
2012	F	18	0	0	···	0	1

Why there is missing observation?

- 例如进行农作物试验，目标变量是农作物产量，控制变量有水分，肥料，温度等。试验中可能会出现意外情况，如种子没有发芽，或发芽后被鸟叼啄，造成某些产量数据缺失。
- 对于收入状况调查，低收入者普遍给出回答，高收入者普遍倾向无回答，且收入越高，无回答倾向越严重。

Deal with Missing data

- If some data is missing, what can we do?
  - Just delete it! (delete feature or samples)
  - Imputation（插补法）
    - Mean imputation
    - Nearest neighbor imputation
  - Other methods
    - Dummy variable (‘unknown’)
    - Collaborative filtering/Bayesian model (HMM, Kalman filter)

Mean Imputation

- 均值插补

计算该特征中非缺失值的平均值（数值型特征）或众数（非数值型特征），然后使用平均值或众数来代替缺失值。

统计方法

The traffic speed prediction project has used similar method to preprocess the data

学生信息数据表

入学年份	性别	年龄	足球	篮球	···	购物	化妆
2012	M	18	0	0	···	0	0
2012	F	18	0	1	···	0	0
2010	M	20	0	1	···	0	0
2012	F	18	0	0	···	0	2
2011	F	18	0	0	···	1	1
2012	F		0	0	···	1	0
2012	F	18	0	0	···	0	0
2011	M	18	2	0	···	0	0
2011	F	19	0	0	···	0	0
2012		18	0	0	···	1	0
2012	F	18	0	0	···	0	0
2011		19	0	1	···	0	0
2012	F	18	0	0	···	0	0
2012	F		0	0	···	0	2
2012	F	18	0	0	···	0	1

Nearest Neighbor Imputation

- 最近距离插补

根据研究对象在辅助变量上的接近程度来选择赋值。例如可以根据学生的兴趣爱好来估计性别，关键是如何定义兴趣爱好之间的距离。

几何方法

学生信息数据表

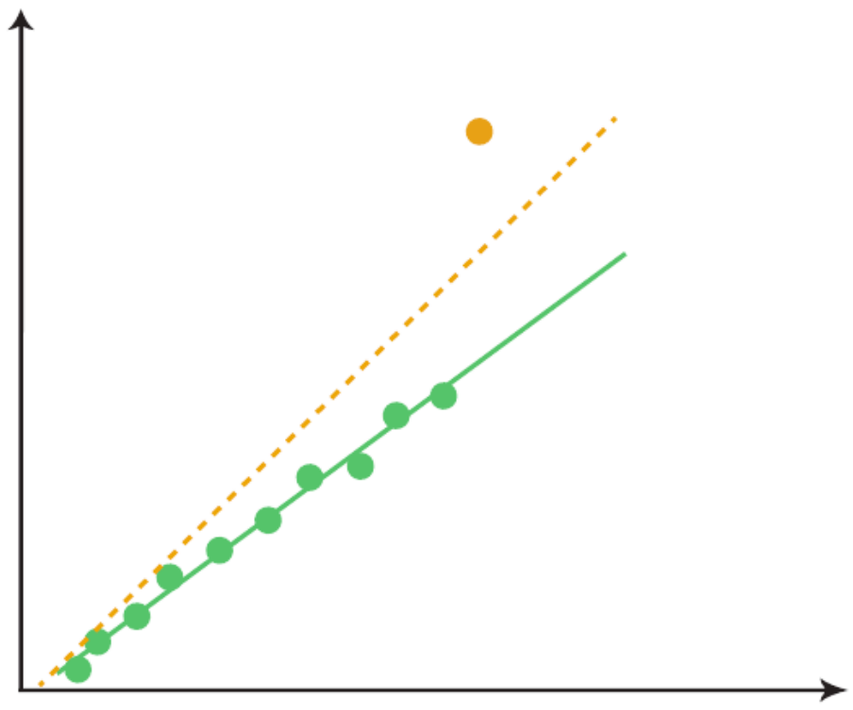
入学年份	性别	年龄	足球	篮球	···	购物	化妆
2012	M	18	0	0	···	0	0
2012	F	18	0	1	···	0	0
2010	M	20	0	1	···	0	0
2012	F	18	0	0	···	0	2
2011	F	18	0	0	···	1	1
2012	F		0	0	···	1	0
2012	F	18	0	0	···	0	0
2011	M	18	2	0	···	0	0
2011	F	19	0	0	···	0	0
2012		18	0	0	···	1	0
2012	F	18	0	0	···	0	0
2011		19	0	1	···	0	0
2012	F	18	0	0	···	0	0
2012	F		0	0	···	0	2
2012	F	18	0	0	···	0	1

What is outlier（离群值）？

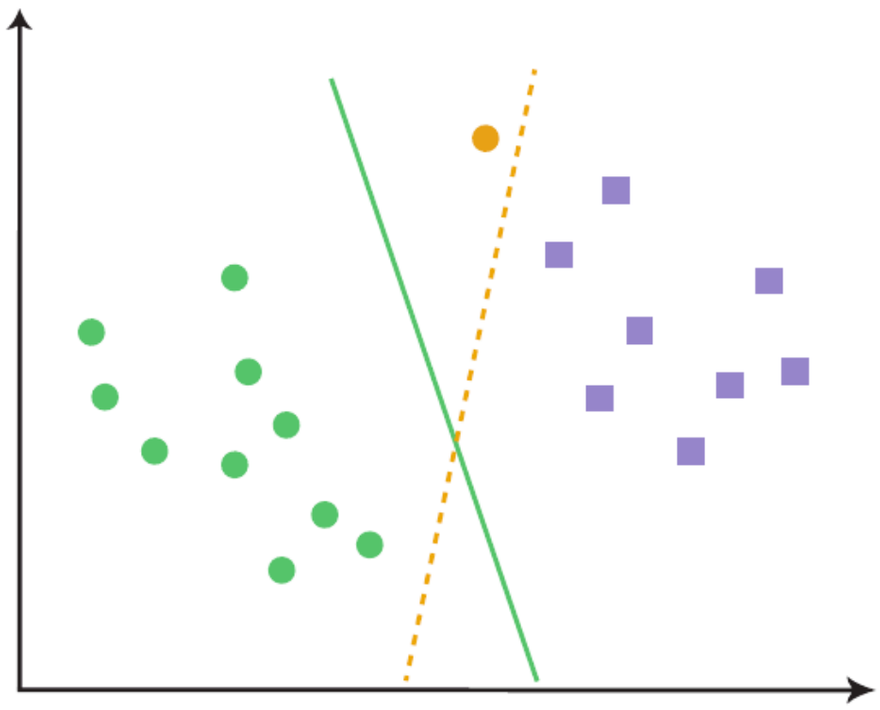
- 指一个数据集中那些明显偏离数据集中的其他样本
- 一种带有统计学味道的定义是：一个观测与其他观测偏离太多以致于值得怀疑它是由不同的机制所产生的
- 产生原因：自然变异、数据测量和收集的误差以及人工操作失误等
- 离群值检测可以作为数据预处理的一个步骤，为数据分析提供高质量的数据
- 离群值检测也可以直接用来解决很多应用问题，例如信用欺诈检测、电信欺诈检测、疾病分析和计算机安全诊断等

Example: New impact on stock market

Impact of outlier



回归

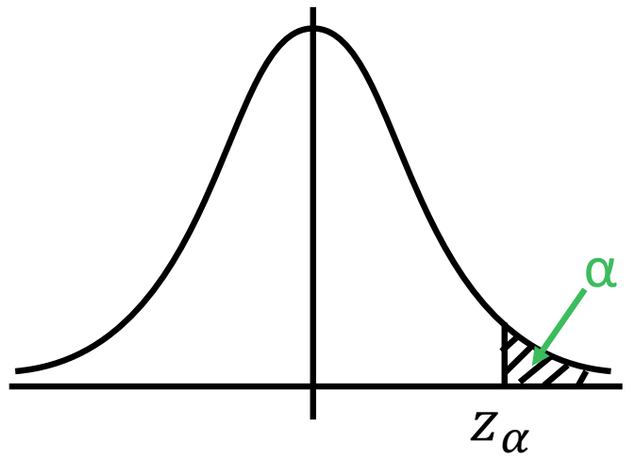


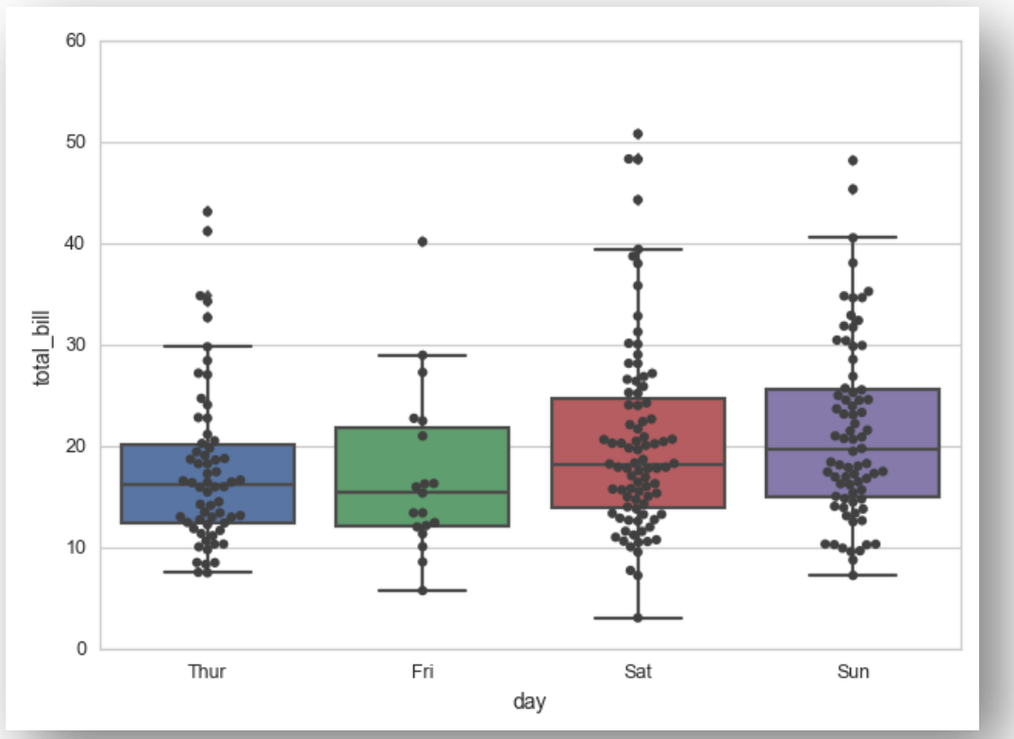
分类

Outlier detection

① 基于统计的方法

- 在上、下α分位点之外的值认为以异常值
- 盒图观察





Outlier detection

② 基于近邻的方法

- 局部异常因子算法（LOF算法，Local Outlier Factor）
- 基本想法

  通过比较每个点p和其邻域点的密度来判断该点是否为异常点，如果点p的密度越低，越可能被认定是异常点

  密度通过点之间的距离来计算，点之间距离越远，密度越低，距离越近，密度越高