数据挖掘 极限复习:

1.2 & 1.1 ① Data mining: Data $\xrightarrow{DM}$ Value

② Data is complex and interconnected, and DM is the discovery of models for data, or the study of collecting、processing、analysing and gaining useful insight from data.

③ DM思路:
- 易得info (已知) $\xrightarrow{建立关联}$ 目标info (已知成员信息)
- 易得info (新) $\xrightarrow{推测}$ 目标info (未知)

2.2. DM Pipeline Also known as: 记录、样本 → can also be: 变量、字段、特征

① What is data? 对象 objects & 属性 attributes 且 obj & attr 有 value, 且 collection of attr-value pairs describes a specific object

② Relational Data (关系型): Data Stored in a relational table with a fixed schema (模式), 其基础为 relation, 实体间有 entity-relation 图

③ Type of Attr: Numeric or Categorical

For numeric Relational Data, obj can be thought of vectors

For some numerics, they can actually be categorical

④. Binning 分箱: - Split the range of the domain of numeric attr into bins (intervals).

- Every bucket ⇒ defines a categorical value

Decision on num of bins: The desired granularity

What about bucket size? Equi-width / size (depth) / log

Or: Optimized Bin: 1D clustering | range 等长 其中等深 log(range) 等长

Equilog 法 better for skewed 分布 (如幂率分布).

⑤ Beyond relational data: Set data: Each record is a set of items from a space of possible items.

E.g. Document: Bag of Words Representation.

许多也可以转成 numeric vector data

⑥ Dependent Data: Order: 知 time order of data (时序, string)

Spatial: Data is placed on specific locations

Spatial Temporal: Data with location & time.

Graph data → Networked: Data with pairwise relations between entities

⑦ Graph Data: May have directed links (or undirected)

表示方法: ① Adjacency Matrix: sparse. wasteful 但 conceptually useful

② Adjacency List: Not easy to maintain ③ List of pairs

⑧ Pipeline: Data Collection → Data Preprocessing → Data Mining → Result Post Processing

3.1 预处理

① Sampling: 用于初步调查与最终分析

Key principle: sample should be representative.

If not: bias! ; if it has approximately the same property as the original set of data

Type: Random、with or without replacement

Stratified (分层抽样), 属于 biased sampling (bias 有用的)

② Feature Extraction: TF-IDF word weighting

Data quality problem: Noise & Outliers、Missing & Duplicate

③ Taking Text as an Example:

First cut: Remove 停顿、大小写、清除空格

break into words and keep the most popular words

But: ☆ Most frequent words are stop words

Second cut: Remove stop words (a predefined list)

④ Then: TF-IDF: The words that are best describing a document (i.e., important for this doc), but also unique.

TF(w,d): term frequency of w in documend d.

IDF(w): inverse document frequency

$\Rightarrow$ TF-IDF(w,d) = TF(w,d) × IDF(w)

More specifically, $DF(w)$: fraction of documents that contain w.

Then $IDF(w) = \log\left(\frac{1}{DF(w)}\right)$.

$TF(w,d)$: Number of times that w appears in the document d.

So, the third cut is: TF-IDF, as it takes care of stop words as well ($IDF(w)=0$ if $w \in Stop$).

② Word & Doc Representations: Eg. TF-IDF to every word, and word vector to a doc. Recent trend: embeddings.

⑥ Word2Vec., 两种: CBOW (predict missing word from context) & Skip-Gram (predict context from word)

新: BERT: 掩码, 而后重建 (Transformer-Based)

⑦ Data Normalization: a. Column Norm: Different attr take very different range of values.

做法: Divide columns by Maximum of each column.

Or: 减去 min 后, 除以 (Max-Min). Range: [0, 1] (0: min, 1: max)

b. Row Norm. 如 doc & word 场景

做法: Row 除以 Row sum (之后的 vector 变为一个分布)

★: 不同目标, 可能采取的 Norm 方式不同

Eg.

	item1	2	3
User1	1	2	3
User2	2	3	4
	-1	0	+1
	-1	0	+1

欲知: Do user1&2 rate in a similar way?

则: -Mean.

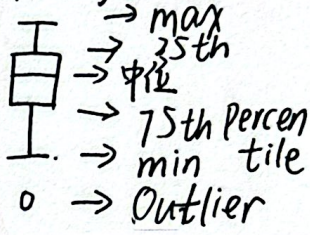
⑧补：Z-score：$z_i = \frac{x_i - mean(x)}{std(x)}$ $mean(x) = \frac{1}{N}\sum_{j=1}^{N} x_j$

$std(x) = \sqrt{\frac{\sum_{j=1}^{N}(x_j - mean(x))^2}{N}}$

3.2 Post-Processing

① Visualization: plot(折线) scatter. bar(条形图)

更进一步：hist：直方图 box：箱形：



errorbar：误差线 mean & std

pie：饼图

② 维度缩减：Principal Component Analysis

③ Statistical Significance：A result is interesting if it is not produced by randomness

☆ P-value: The prob of obtaining the observed results, assuming that the null hypothesis is true.

④ Data statistics：一些统计，如 mode(众数). pecentile等 mean. median、trimmed avg(除去min&max). Variance. Range(极差)

⑤ Distribution of Words: Power-Law 分布

$p(k) = k^{-a}$ ($\log p(x=k) = -a\log k$) ↘ 且不止words: Everywhere!

⑥ Attr Relationships: 可画折线图，也有些统计指标

$X=\{x_1,\cdots,x_n\}$ $Y=\{y_1,\cdots,y_n\}$, $corr(X,Y) = \frac{\sum_i (x_i-\mu_x)(y_i-\mu_y)}{\sqrt{\sum_i (x_i-\mu_x)^2}\sqrt{\sum_i (y_i-\mu_y)^2}}$

被称为：Pearson correlation 系数：

衡量两变量线性相关程度

Spearman rank Correlation: 两变量是否 rank-correlated

4.1 Similarity & Distance.

① Similarity metric: how close are two objs?

欲: $S(p,q) = S(q,p)$ & $S(p,q) = 1$ iff $p=q$

a. Intersection: p, q间多少元素同?

b. Jaccard: $JSim(S_1, S_2) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}$

c. Cosine Sim: 两个obj当vector看待: $Cos(d_1, d_2) = \frac{\langle d_1, d_2 \rangle}{\|d_1\| \cdot \|d_2\|}$

常用于比较 documents, 有时 vector 会用 document长度 norm, 或者文字用 tf-Idf 赋以 weight

d. Correlation Coeff: 也可以

② Distance: how different are two objs?

欲: $d(x,y) \geq 0$ $d(x,y) = d(y,x)$ $d(x,y) \leq d(x,z) + d(z,y)$

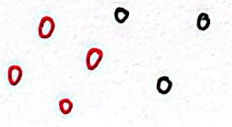
Lp-norm: $Lp(x,y) = (|x_1 - y_1|^p + \cdots + |x_n - y_n|^p)^{1/p}$

p也可取1或2; Hamming Dist: # of differ

③ Similarity → Distance: Jaccard Dist: $JDist(X,Y) = 1 - JSim(X,Y)$

Cosine Distance: $Dist(X,Y) = 1 - \cos(X,Y)$.

④ Distance Between sets of Points:

a. 

Min or Max dist over all pairs

Avg dist over all pairs

Dist between avgs

b 将点视为分布 ⇒ 两个分布间距离？

有：Variational Distance：L_1 dist between the distribution vectors

KL-Divergence：$D_{KL}(P\|Q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$

理解：Amount of additional info needed to represent P given Q

⚠ Asymmetric：可以 $\frac{1}{2}(D_{KL}(P\|Q) + D_{KL}(Q\|P))$

⑨ Ranking Distance：$R_1 < x, y, z, w > \in \mathbb{R}^n$

$R_2 < y, w, z, x >$

a. Kendall's tau：每个有 C_n^2 个 pair，则距离为 $C_n^2 - |CR_1 \cap CR_2|$

where CR_i 代表 R_i span 出个 C_n^2 个 pair

b. Spearman rank Distance：L_1 dist between ranks

⑩ 推荐系统 e.g. Def：$sim(u,v)$ Corr(u,v)：u, v 各自 vector，减去 mean（不考虑 zeros），然后算 Cosine 距离

对于 user u，找 set TopK(u)，whose ratings 与 u 最 similar

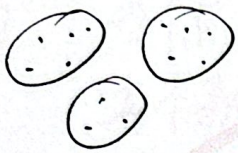
则对于 $\forall i$，$\hat{r}_{ui} = \frac{1}{Z} \sum_{v \in TopK(u)} sim(u,v) r_{vi}$，$Z = \sum_{v \in TopK(u)} sim(u,v)$

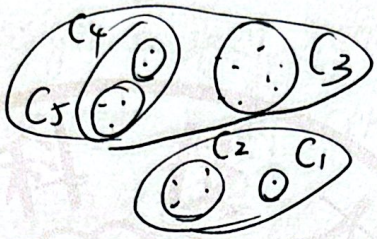
Modeling Deviations：$\hat{r}_{ui} = \bar{r}_u + \frac{1}{Z} \sum_{v \in TopK(u)} sim(u,v)(r_{vi} - \bar{r}_v)$.

Evaluation，$RMSE = \sqrt{\frac{1}{n} \sum_{i,j} (\hat{r}_{ij} - r_{ij})^2}$，Precision & Recall

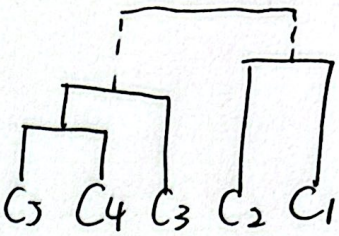
5.2&6.2 Clustering { K-means & Variants
Hierarchical clustering
DBSCAN

① Type: Partitional & Hierarchical





&



② Clustering Objective: Center of a Cluster: Centroid

- Center-Based Cluster: Obj in a cluster 到该簇 center 距离 小于到其余簇 center
- Contiguous Cluster: Cluster 中 ∀ point 的最近 neighbor 仍在该簇中
- Density-based Cluster: A cluster is a dense region of points, which is ~~represented~~ seperated by low-density regions, from other regions of high density

⇒ Used when clusters are irregular, and when noise and outliers are present.

③ Kmeans: Given a set X of n points in a d-dim space and an integer K group the points into k clusters

$C = \{C_1, C_2, \dots, C_K\}$ s.t.

$$Cost(C) = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - C_i\|^2$$

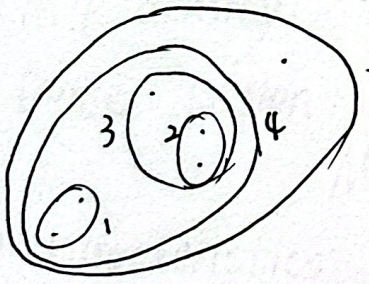
is minimized, c_i is the mean of C_i

算法：① 选K个初始 centroid　　复杂度：O(nKId)
② Repeat
③ 每个点，assign 至最近 centroid
④ 每个新 Cluster 算新 centroid
⑤ 至到 centroid 不再大幅变动

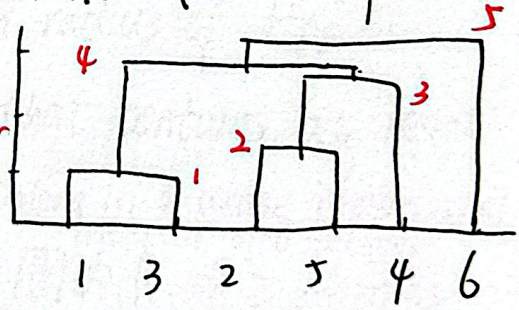
☆ 最终结果，与 K 个 centroid 初始化有关
可用 K-means++ 风格的点、初始化

Limitations：当 cluster 间：size/ densities/ non-globular shape 时，有问题；且有 Outliers 也有问题
一种解决方案：K↑，然后手动 merge

④ Hierarchical Clustering：→ one cluster at a time,
且常用 similarity/distance 矩阵
a. Agglomerative（凝聚式）：每一步，合并最近一对；
b. Divisive（分裂式）：每一步，分裂一个 cluster；



nested cluster diagram



刻度：distance.　　dendrogram

优势：不必定K，且 Dendrogram 可对应真实 taxonomies

重要: How to compute the proximity of two clusters, (Inter-Cluster Similarity) 可以是: min/max, group avg, dist between centroids ..... min → single link, max → Complete linkage

single link: 一对连上,两簇merge

Complete linkage: 两簇间全连,才merge

Stength of MIN: 可 handle non-elliptical shape ☆

但: sensitive to noise & outliers

Strength of MAX: Less susceptible to noise & outliers

但: tend to break large clusters. 且 biased towards global clusters

而 Group Avg 是一个 Compromise, Strength & Limitations 同 MAX

Complexity: $O(N^2)$ space and typically $O(N^3)$ time

⑤ DBSCAN - Density-Based Clustering

Def: a. 点 p density: # of points with radius of Eps

b. Dense Region: A radius Eps that contains at least MinPts points

c. 点的 characteristics:

- core point: Eps 圆内点 # ≥ MinPts → belong in a dense region, and interior of a cluster
- board point: # < MinPts, 但在 core point 旁边
- noise point: 非上述二者

若两个 core point 间 dist < Eps, 则有连接一条边
则最后有连通图, 连通分量#为簇#

故 algorithm: ① label points as core, border & noise
② 去掉 noise ③ For every core point p 未被分给一个簇: 创建一个 cluster, with point p, 以及所有 density-connected to p 的点
④ 把 boarder point assign 给 closest core point.

关于 MinPts, 一般认为 2×dim 比较好。随 MinPts ↑, Eps 缓慢上升, 直至一个 "knee" moment, Eps 随 MinPts 急速↑

DBSCAN resistent to noise, 可handle 不同 shape & size 的簇
但 sensitive to parameter, 且在 varying densities & 高维数据下不是很 work

直觉: 簇内紧凑 (cohesion), 簇间分离 (separation)

⑥ Clustering Evaluations: 有 internal & external Index

a. 内部: E.g. SSE, $BSS = \sum_i m_i \|c - c_i\|^2$ where m_i: size of 簇 i, c is the overall mean
(也记作 WSS)
或 $WSS + BSS = TSS = \sum_i \sum_{x \in C_i} \|x - c\|^2$

b. 轮廓系数 (Silhouette Coefficient)
For 点 i: a_i = avg distance of i to points in its own cluster
b_i = min (over clusters) of the avg dist of i to other clusters

则系数为：$S_i = 1 - a_i / b_i$，$a(i) = \frac{1}{|C_I|-1} \sum_{j \in C_I, i \neq j} d(i,j)$.

$b(i) = \min_{J \neq I} \frac{1}{|C_J|} \sum_{j \in C_j} d(i,j)$.

$S_i \in [0,1]$ (typically)，closer to 1，the better

则，可为 cluster 算平均轮廓系数

c. Correlation: Similarity & Incidence Matrix 算 corr（diagonal 不算）

=1 if associated pair is in a cluster

d. 可以考虑统计显著性：生成 random data & random clustering

看真实结果在随机基线中是否很罕见

⑦估计合适的簇数：画 SSE-K 曲线，找 elbow（只是参考）

⑧若有 GT label：

Cluster	Class 1	2	...	
1	n_{11}	n_{12}	...	m_1
2	n_{21}	n_{22}	...	m_2
⋮	⋮	⋮	⋱	
	c_1	c_2		n.

则 $Prec(i,j) = \frac{n_{ij}}{m_i} = P_{ij}$

$Rec(i,j) = \frac{n_{ij}}{c_j}$

$F(i,j) = \frac{2 \times Prec(i,j) \times Rec(i,j)}{Prec(i,j) + Rec(i,j)}$，

(F measure)

既希望簇纯，又希望覆盖完整

要同时看 precision & recall

7.1 & 7.2 有监督学习：

① 一维线性回归：$f(x_i) = w_0 + w_1 x_i$

minimize SSE：$w_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$，$w_0 = \bar{y} - w_1 \bar{x}$

多元：$w = (X^T X)^{-1} X^T y$，$f(X) = Xw$

回归对 Outlier Sensitive，但可以 remove 掉；

同时只能表达线性关系

② 为了回归，归一化很重要：推荐 z-score

$z = \frac{x - \text{mean}}{\text{std}}$，可以去均值并缩放到"按 std 计量"单位

③ Evaluation of Classification models　预测对/错 Pred.　T/F　P/N

$\text{Precision} = \frac{tp}{tp + fp}$

$\text{Recall} = \frac{tp}{tp + fn}$

$\text{Acc} = \frac{tp + tn}{tp + tn + fp + fn}$

④ Decision Tree：Hunt Algorithm：选一个 attr 分，然后重复

什么是好的分裂？让子节点、不纯度下降

常见 impurity metric：a. Entropy $= -\sum p_i \log_2(p_i)$

（纯，则熵小）

b. Gini Index：$1 - \sum p_i^2$

c. 分类 Error

⑤ 贪心分裂不一定全局最优；但决策树可解释强

却有可能过拟合（可用 early stop 缓解）。

⑥ 三类 classic classifier

a. KNN，对于x，找离它最近 k个样本，看它们类别，用 majority vote 决定 x 类别 （惰性学习器）

Good: 不train，直观 Bad: 距离敏感，高维下差*

* Causal of Dimensionality: Dim ↑，样本间距离接近，"远"和"近"差别不明显

b. SVM：两类 data point 可由一个超平面分开

⇒ 让分类边界到样本点间隔尽可能大

欲：w, b s.t. $y_i(w \cdot x_i + b) \geq 1$ for $\forall i$

且欲 $\min_w \frac{1}{2}\|w\|^2$ ↗，则是带约束的优化问题

Good: 边界维健，公式清晰，高dim下效果不错

Bad: 需要线性可分，需要解优化问题

c. Naive Bayes: $P(C|A) = \frac{P(C)P(A|C)}{P(A)}$

有 feature $A_1, \cdots, A_n$，估：$P(A_1, \cdots, A_n|C)$

$= P(A_1|C) \cdots P(A_n|C)$ （独立性假设）

则对于新样本：$(a_1, \cdots, a_n)$ 则 对于每个类别 c：

$$Score(c) = \underbrace{P(C=c)}_{\text{均从 Training Set 中获得}} \cdot \prod_{i=1}^{n} P(A_i = a_i | C = c)$$

若是连续性特征，则 $A_i|C=c \sim Normal(\mu_{ic}, \sigma_{ic}^2)$

但防止某个 $p=0$，故 Laplace 平滑：

$$P(A_i=a \mid C=c) = \frac{N(a,c)+1}{N_c+m_i}$$，m_i 为 A_i 可能取值个数

且实际计算时，使用 log 防止 float 精度问题

Naive Bayes 特别适合文本分类

Good: Simple，训练快，高维稀疏文本也不错

Bad: 条件独立假设，不太容易自然融入更复杂额外特征

9.2 Map Reduce Extract something you care about

Map → 对输入每一块做局部转换，输出 key-value

Reduce → 把有相同键数据聚在一起汇总

二者间还有 Shuffle & Sort → 还可作统计字节数&关系连接

经典：词频统计；适合大规模 data，因为并行、容错性强

MapReduce 中有：master & worker

划分/监管/监控　　↓ Map/Reduce/中间工作

适合：能拆为很多子任务的工作，大规模批处理

不适合：需要频繁交互的实时计算；迭代；低延迟

Eg. Hadoop：围绕分布式文件系统与 MapReduce 的生态 sys