# 注意事项

We will specify whether to follow the Relationship Model or RDBMS. If not specified, you should explain your answer considering both perspectives.

# 考点

Week2

- 根据 properties of data，判断 data types (i.e. nominal, ordinal, etc.)
- statistical 的计算
    - Central Tendency
    - Dispersion
        - percentile 的计算很关键

Week3

- Box plot 的特点
- Correlation
- 搞清楚 Bar, Pie, Histogram, Box, scatterplot 的用法

Week4

- RDBMS
- SQL
    - Creation
    - Insertion
    - Updating
    - Deleting
- KEY Properties
    - PK
    - FK
    - CANDIDATE KEY
- IC
    - Domain
    - Key
    - Referential
- 识别 star 和 snow flake Fact table

Week5

- NULL
- 怎么用 SQL Aggregation 实现要求的问题 — TUT
    - mode
    - percentile

Week6

- 区分 Observational Study 和 Experimental Study
- 区分 variable type 的类型; dep, indep
- 判断 Q, H0, H1, P-VALUE, ALPHA, P-HACKING, one-side, two-side
- 求 P-value 的方法
- Holdout & cross validation methods; LOOCV
- Accuracy, precision, recall, f1

Week7

- 计算 frequent itemset
- Apriori, FP Tree
    - frequent itemset
    - rule generation

Week8

- cluster distance
    - 对噪音敏感
- K-means 区分 cluster 的步骤
- Agglomerative 区分 cluster 的步骤
- SSE, Silhouette coefficient
- 根据 eignvalue 和所占比重删除对应的 PCA Component

Week9

- 知道如何求 Simple Linear Regression 的$\alpha$和$\beta$
- multitple linear regression 的 weight 迭代

- logistic regression
    - 对噪音敏感
    - 使用公式判断 class 是什么
- R^2
- Gradient Descent

Week10

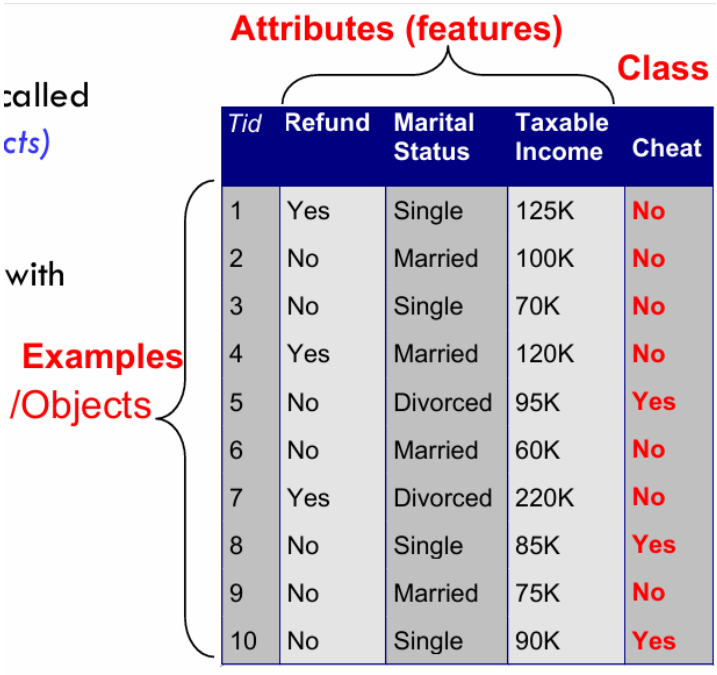
- Decision Tree
    - 对噪音敏感
- Entropy
- 使用 IG 判断 attribute 是否好

Week11

- TFIDF
- Naive Bayes
    - 对噪音不敏感
    - Laplace
    - Text classification

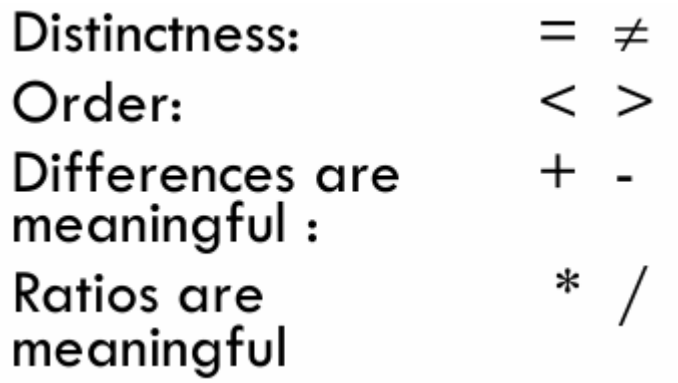
# Week2

## Attributes, Class, Examples/Objects




| Tid | Refund | Marital Status | Taxable Income | Cheat |
|---|---|---|---|---|
| 1 | Yes | Single | 125K | No |
| 2 | No | Married | 100K | No |
| 3 | No | Single | 70K | No |
| 4 | Yes | Married | 120K | No |
| 5 | No | Divorced | 95K | Yes |
| 6 | No | Married | 60K | No |
| 7 | Yes | Divorced | 220K | No |
| 8 | No | Single | 85K | Yes |
| 9 | No | Married | 75K | No |
| 10 | No | Single | 90K | Yes |

## Properties of Data




## Types of Data

- **Nominal** (distinctness)
  - 一般用 Bar Chart 表示数据之间是离散的
  - Examples: ID, eye color, zip codes

- **Ordinal** (distinctness, order)
  - 一般用 Bar Chart 表示数据之间是连续的
  - **Central tendency** can be measured by mode or median. 不能用 mean 因为不支持计算 mean；可以求 percentile
  - Examples: ranking, grades, height, educational levels
- **Interval** (distinctness, order & differences are meaningful)
  - **Central Tendency**
    - Mode, median, mean
  - **Dispersion**
    - IQR, stdev, variance
  - No true zero, 这意味着区间尺度下的 0 并不表示“无”某个特征的状态，而是数字 0 的意思，比如 0 度表示不是没有温度。这是 Interval 和 Ratio 的区别
  - Equal intervals between values
  - Examples: calendar dates, temperatures in Celsius or Fahrenheit
- **Ratio** (all 4 properties)
  - 一般用 scatterplot
  - Has true zero (Zero defined), 这意味着区间尺度下的 0 并表示“无”某个特征的状态。比如 0kg 表示无重量。
  - Examples: temperature in Kelvin (Kelvin 的 **0 K** 是**绝对零度**：也就是 **分子完全没有运动**，这是物理上真正的“无温度”状态。), length, counts, elapsed time

| | Nominal | Ordinal | Interval | Ratio |
|---|---|---|---|---|
| Countable | ✓ | ✓ | ✓ | ✓ |
| Order defined | | ✓ | ✓ | ✓ |
| Difference defined (addition, subtraction) | | | ✓ | ✓ |
| Zero defined (multiplication, division) | | | | ✓ |

可以看到只有 ratio 是 zero defined

| | Nominal | Ordinal | Interval | Ratio |
|---|---|---|---|---|
| Mode | ✓ | ✓ | ✓ | ✓ |
| Median | | ✓ | ✓ | ✓ |
| Mean | | | ✓ | ✓ |

**Central Tendency** 可以看到只有 interval 和 ratio 支持 mean 的计算

| | Nominal | Ordinal | Interval | Ratio |
|---|---|---|---|---|
| Counts / Distribution | ✓ | ✓ | ✓ | ✓ |
| Minimum, Maximum | | ✓ | ✓ | ✓ |
| Range | | ✓ | ✓ | ✓ |
| Percentiles | | ✓ | ✓ | ✓ |
| Standard deviation, Variance | | | ✓ | ✓ |

**Dispersion** 只有 Interval 和 Ratio 支持 Stdev

## Central Tendency 的计算

- Mode：数据集中**出现次数最多的值**
- Median：等同于 50 percentile
  - 需要先 sort，再计算
  - 假设总数是 N，索引从 0 开始
    - N 是 odd，则 median 是在[(N-1)/2] idx 上的值.减一是

因 idx 从 0 开始

- N 是 Even，则 median 是在(N/2)-1 和(N/2) idx 上 2 个值的 mean

- Mean: 所有值的和除以 COUNT 即是 Mean

**Dispersion 的计算**

- **Variance**: 这节课应该都是默认除以 N-1 的（即样本方差）

$$\frac{\sum(X_i - mean)^2}{N-1}$$

- **Stdev**: Square root of variance，和 variance 一样都是表示数据是否集中，即 stdev 越小，则数据越集中于 mean
- **Percentiles -- interpolating**
  - 需要先 sort，再计算
    - 这里默认 index begin from 0
    - N 表示的是数据点的数量，比如有 10 个 row，则 N=10
    - N 和 index 不同
  - 10th percentile is item at index 0.1 * (N - 1).
  - 90th percentile is item at index 0.9 * (N - 1).
  - 对于 interval，ratio 来说，如果 index 不是整数怎么办？这里以 5.4 为例子
    - Percentile_Value = value at index 5 + 0.4 * (value at index 6 - value at index 5)
  - 对于 ordinal 来说，如果 index 不是整数怎么办？这里以 5.4

为例子

- 我们选择 idx6 的 value 作为 Percentile_Value

- **IQR**:
  - 主要用来衡量数据的 **离散程度**，也就是数据的“分散”或者“变异”有多大。数据的 IQR 越小，说明中间 50%的数据越集中，波动较小
  - **IQR**=Q3-Q1；其中 Q1 表示 25 percentile，Q3 表示 75 percentile
- Counts, min, max, range 不需要计算，其中 Range 值 Max-Min

**Data Cleaning**

- Type and name conversion.
- Filtering of missing or inconsistent data. NA
- Unifying semantic data representations. 避免情绪用词
- Matching entries from different sources.
- Rescaling and optional dimensionality reduction.比如 PCA,在 Week8

**Terminology**

- **Pivot table**: Summarize data by calculating statistics over sub-populations，不是 bar chart
  - e.g., count of industry by name.

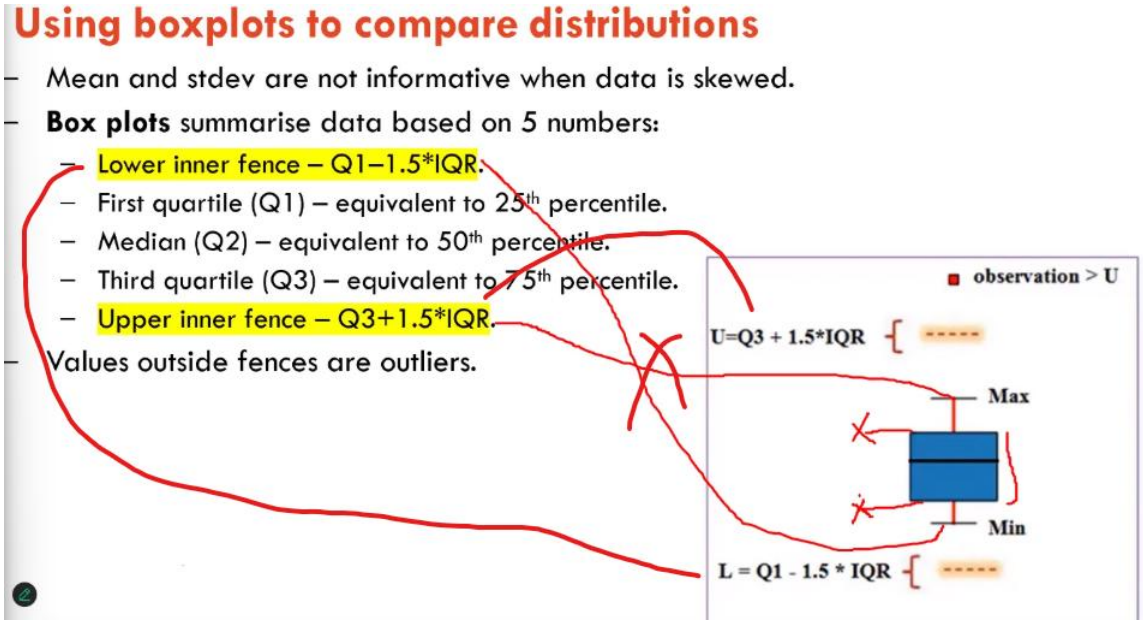
| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | 商品名 | 月日 | 数量（辆） | 价（辆） | 金额（辆） |
| 2 | 甲 | 2013-9-24 | 11 | 15 | 165 |
| 3 | 乙 | 2013-9-25 | 12 | 16 | 192 |
| 4 | 丙 | 2013-9-26 | 13 | 17 | 221 |
| 5 | 丁 | 2013-9-27 | 14 | 18 | 252 |
| 6 | 戊 | 2013-9-28 | 15 | 19 | 285 |
| 7 | 己 | 2013-9-29 | 16 | 20 | 320 |
| 8 | 庚 | 2013-9-30 | 17 | 21 | 357 |
| 9 | 辛 | 2013-10-1 | 18 | 22 | 396 |
| 10 | 壬 | 2013-10-2 | 19 | 23 | 437 |
| 11 | 癸 | 2013-10-3 | 20 | 24 | 480 |
| 12 | 甲 | 2013-10-4 | 21 | 25 | 525 |
| 13 | 乙 | 2013-10-5 | 22 | 26 | 572 |

# Week3

## NaNs VS NaTs

- NaNs: not a number, float 类型
- NaTs: not a time, datetime 类型

## Box Plot




- Min: 非离群值中的最小值（≥ Lower Inner Fence）
- Max: 非离群值中的最大值（≤ Upper Inner Fence）
- **Outer fence**: Q1/3 +/- 3*IQR

## Correlation

### Pearson's

- Capture linear relationships
  - 1 indicates a perfect positive linear relationship
  - –1 indicates a perfect negative linear relationship,

- 0 indicates no linear relationship.

**Spearman's**

- Capture both linear and nonlinear monotonic relationships
  - Monotonic: 当一个变量增加时，另一个变量要么**总是增加**，要么**总是减少**，这种关系就叫做**单调关系**
- Particularly suitable for ordinal, ratio data.
- measures the strength and direction of the relationship between two variables when they are monotonically related.
- This means that the relationship is consistent in direction (either always increasing or always decreasing), regardless of whether it is linear or nonlinear. 只要方向一致（总是增加或总是减少），不论这种关系是线性的还是非线性的，都可以使用 Spearman。
  - 1: 一个变量上升时另一个也总是上升（单调递增）
  - 0: 没有单调关系, no monotonic
  - -1: 一个变量上升时另一个总是下降（单调递减）

Kendall's tau

- for ordinal data
- 和 Spearman 一致，但是适合更小的 dataset

Plot Properties

- Bar: Count frequency of each category (Nominal, Ordinal)
- Pie: 显示部分与整体的比例关系。
- Histogram: Count frequency of numerical (Interval, Ratio)
- Box: 展示 numeric 的**分布特征**，包括中位数，四分位数，异常值等

- Scatterplot: The most suitable to visualize and analyze the correlation between two variables

# Week4

**Relation schema**: specifies name of relation, and name and data type (domain) of each attribute.

**Relation instance** is a set of tuples (a table) for a schema.

**Ideal Relation**

- unique table name
- column with unique names
- All rows have the same structure
- Every column is atomic — known as 1NF
- Every row is unique
- The order of the rows is immaterial.

**Real word Relation — RDBMS (在 ideal 上的扩展)**

- **allow duplicated rows**
- support ordering tuples and attributes
- **allows "null"**

**DDL**

- CREATE
- ALTER
- DROP

**DML**

- INSERT, DELETE, UPDATE
- SELECT … FROM … WHERE

**Syntax**

CREATE TABLE NAME(…);

DROP TABLE NAME CASCADE;

INSERT INTO table (list-of-columns) VALUES (list-of-expression);

UPDATE table SET column = expression WHERE search_condition;

DELETE FROM table WHERE search_condition;

**PRIMARY KEY（主键）**

- **唯一性**：每个表只能有一个主键（At most one per table），并且主键字段的值必须是唯一的。
- **不允许 NULL 值**：主键字段不能包含空值（Automatically disallows NULL values）
- **Composite PK 可以自己设定**，但必须满足唯一性约束。

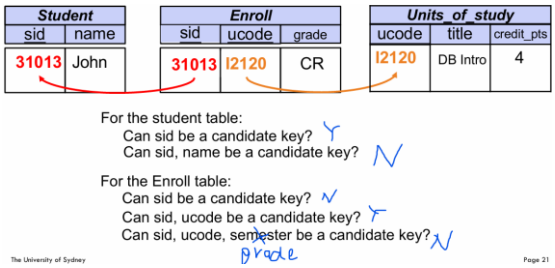
**CANDIDATE KEY**

- **唯一性**：候选键中的字段也必须是唯一的（all must be declared as UNIQUE）。
- **最小性（Minimality）**：没有多余的属性，可以删除任何一个属性而不再满足唯一性的条件。这里的最小是不是指 size，而是说不能去掉任何属性。
- **可以包含 NULL 值**

**SUPER KEY**

- 如果只满足唯一性，那么就叫做 Super Key

- 所有 Candidate key 和 PK 都是 Super key




## FOREIGN KEY

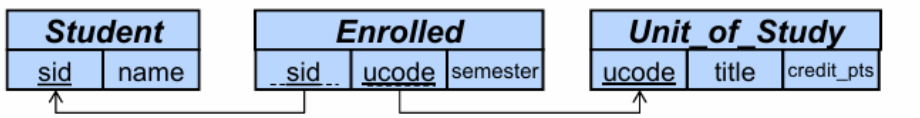
- 如默认允许 NULL 值
- If there must be a parent tuple, then must combine with NOT NULL constraint

**Domain Constraints**: restrict attributes to valid domains

- NOT NULL / NULL
- DEFAULT
- CHECK -- User defined domain
    - 确保每个属性的值符合预定义的数据类型和取值规则。比如，成绩”只能是 'A', 'B', 'C', 'D', 'F' 之一

## Key Constraints & Referential Integrity

- Primary key
- Unique
- FOREIGN key
    - **Referential Integrity**




```
CREATE TABLE Student ( sid INTEGER, ... ,
    CONSTRAINT Student_PK PRIMARY KEY (sid)
);
CREATE TABLE UoS  ( ucode CHAR(8), ... ,
    CONSTRAINT UoS_PK PRIMARY KEY (ucode)
);
CREATE TABLE Enrolled ( sid INTEGER, ucode CHAR(8), semester VARCHAR,
    CONSTRAINT Enrolled_FK1 FOREIGN KEY (sid) REFERENCES Student,
    CONSTRAINT Enrolled_FK2 FOREIGN KEY (ucode) REFERENCES UoS,
    CONSTRAINT Enrolled_PK  PRIMARY KEY (sid,ucode)
);
```

**对于 NF**

只用知道 5NF 包含 4NF，并且更加严格；主要是用来解决 Redundant； A relation needs to be decomposed if it does not satisfy the restrictions。

**OLTP** (Online Transactional Processing)

Designed to handle day-to-day business operations.
Focuses on **maintaining dynamic relationships** between business entities.

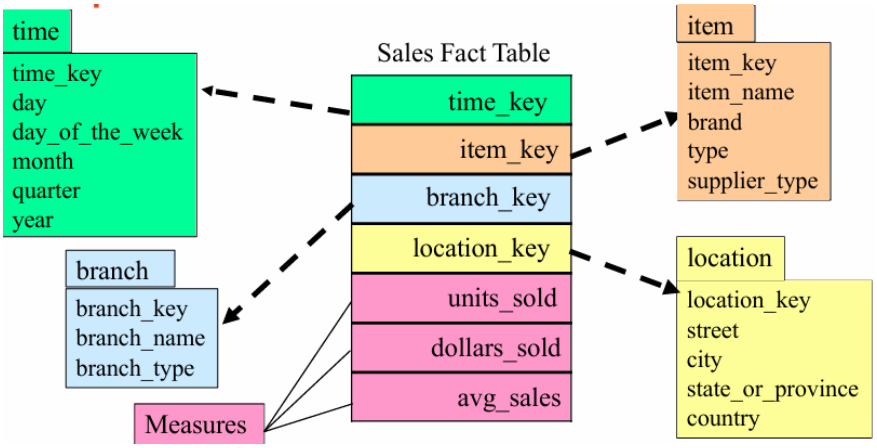
**OLAP** (Online Analytical Processing)

Designed for analyzing **large volumes of historical data**.

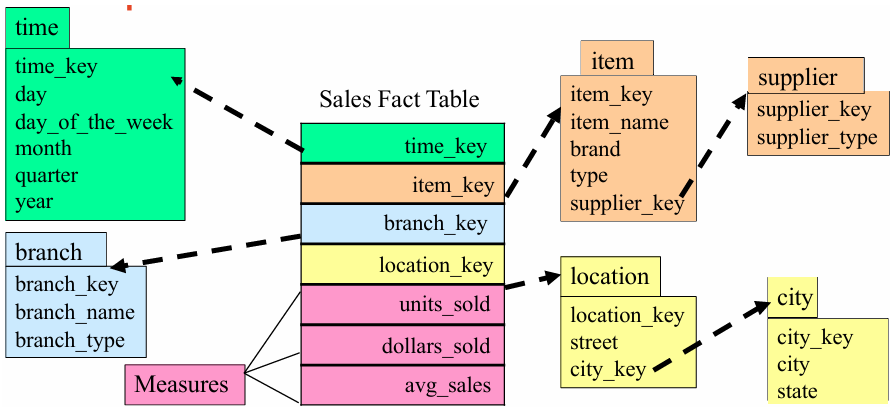
**Data Warehouse**

- Properties
  - Subject-oriented: Organized by subject, not by application
  - Integrated: integrating multiple, heterogeneous data sources
  - Time variant: Large volume of historical data
  - Non-volatile: Updates infrequent or does not occur; maybe only append
- It contains a large central table – Fact Table
  - 目地是 Contains the data without redundancy
  - can be viewed as multidimensional
  - Star schema
    - 结构
      - 1 Fact table

        - N dimension tables with foreign key relationships from the fact table
        - denormalization
    - 特点
        - fast query (因为结构简单，不用去访问多个表)
        - Higher redundancy due to denormalization (存储会重复)
        - 更加简单，没有多个层次的子表，扩展和修改时比较直接




- Snowflake schema
    - 结构
        - The schema consists of multiple dimension tables that are **normalized** into a set of smaller dimension tables
        - 可能不止一个 Fact table
    - 特点
        - slow query (结构复杂，需要访问多个表)
        - lower redundancy due to normalization (存储会避免重复)
        - want to store data more efficiently and perform detailed analyses on the dataset




**事实表**

- 存储维度表主键（作为外键）
- storing numerical measures

**维度表**

- 存储维度属性

# Week5

SQL is case insensitive
这一章节主要是针对 SQL 语法的

### NULL and Three-valued Logic

- 任何和 NULL 进行的运算都是 NULL，比如 5+NULL=NULL
- Any comparison with NULL returns unknown，比如 5<NULL 是 unkown
    - Result of WHERE clause predicate is treated as false if it evaluates to unknown
    - SELECT sid FROM enrolled WHERE marks < 50;
        - ignores all students without a mark, that is if a student with null mark, then we will not out put it in above query
- Three-Valued Logic
    - UNKOWN OR TRUE = True
    - UNKOWN AND FALSE = False

别的都是 UNKOWN

## Strings (CHAR, VARCHAR)

- SQL string literals must be enclosed in single quotes ('like this').
- CHAR: fixed length; VARCHAR: variable length strings up-to max length.
- String comparison is case-sensitive.
- Pattern matching with LIKE operator and % placeholders.
- String concatenation: || (eg. 'hello' || 'there').

```
SELECT *
FROM TelescopeConfig
WHERE tele_array LIKE 'H%';
```

## Date

| SQL Type | Example | Description |
|---|---|---|
| DATE | '2012-03-26' | A date (some systems incl. time) |
| TIME | '16:12:05' | A time, often down to nanoseconds |
| TIMESTAMP | '2012-03-26 16:12:05' | Time at a certain date: SQL Server: DATETIME |
| INTERVAL | '5 DAY' | A time duration |

**CURRENT_DATE**: db system's current date.

**CURRENT_TIME**: db system's current timestamp.

**EXTRACT**(component **FROM** date).

- e.g., **EXTRACT**(year **FROM** startDate)

**DATE** string (Oracle syntax: **TO_DATE**(string,template))

- e.g., **DATE** '2012-03-01'
- Some systems allow templates on how to interpret string.
- Oracle syntax: **TO_DATE**('01-03-2012', 'DD-Mon-YYYY')

**+/- INTERVAL**

- e.g. '2012-04-01' **+ INTERVAL** '36 HOUR'

```
SELECT
  EXTRACT(DAY FROM event_duration) AS days,
  EXTRACT(HOUR FROM event_duration) AS hours,
  EXTRACT(MINUTE FROM event_curation) AS minutes
FROM
EventSchedule;
```

```
SELECT *
  FROM Measurement
 WHERE date = '2005-04-29';

SELECT *
  FROM Measurement
 WHERE date = '29/04/2005';
```

JOIN

- Implicit join is Cartesian join/CROSS join
  - FROM A, B
  - 如果 A 或 B 为空，则 CROSS 也为空
- Inner join，特殊写法 using 替代 on；省略 Inner
  - Theta Join（θ -Join）
  - Equi-Join 等值连接
  - Natural Join
    - Natural Join Caveat

Natural Join Caveat：如果两个表中没有相同的列名，则 `NATURAL JOIN` 会返回笛卡尔积（不推荐）。

- $R \bowtie S = \pi_{unique\_attributes}(\sigma_{equality\_of\_common\_attributes}(R \times S))$

- Outer join
  - LEFF JOIN
    - NATURAL LEFT OUTER JOIN: 不需要 ON 语句，它会自动匹配两个表中相同名称的列，并以左表（Employee）为主表. 结果可能和 LEFT JOIN 不同，因为 NATURAL JOIN 可能匹配多个同名列，而 LEFT JOIN 只会匹配 ON 指定的列。
  - RIGHT JOIN

    - **FULL JOIN:** 不会返回重复的行。在右表没有孤儿行的情况下等同于 LEFT JOIN
- **Natural Join vs. Explicit Equi-Join:** 如果用 explicit equi-join，会导致连接字段出现 2 次,除非用 USING 替代 ON

## Aggregate function

- COUNT(*)计算 NULL，别的都不算，包括 COUNT(x)
- 不能 MIN(AVG())
- 都对 DUPLICATE 起效，除非 COUNT(DISTINCT x)
- If you use GROUP BY function, then in SELECT or HAVING line, you can only include aggregate function or the attribute you use for GROUP BY
- Predicates in the HAVING clause are applied after the formation of groups, whereas predicates in the WHERE clause are applied before forming groups

| SQL Aggregate Function | Meaning |
|---|---|
| COUNT(*attr*) ; COUNT(*) | Number of *Not-null-attr* ; or of all values |
| MIN(*attr*) | Minimum value of *attr* |
| MAX(*attr*) | Maximum value of *attr* |
| AVG(*attr*) | Average value of *attr*(arithmetic mean) |
| MODE() WITHIN GROUP (ORDER BY *attr*) | Mode function over *attr* |
| PERCENTILE_DISC(0.5) WITHIN GROUP (ORDER BY *attr*) | Median of the *attr* values |

```
SELECT COUNT(value),
       MIN(value),
       Max(value),
       MAX(value) - MIN(value)                                  AS Range,
       AVG(value)                                               AS Mean,
       MODE()  WITHIN GROUP (ORDER BY value)                    AS Mode,
       PERCENTILE_DISC(0.5) WITHIN GROUP (ORDER BY value)       AS Median,
       PERCENTILE_DISC(0.25) WITHIN GROUP (ORDER BY value)      AS Percentile25,
       PERCENTILE_DISC(0.75) WITHIN GROUP (ORDER BY value)      AS Percentile75,
       PERCENTILE_DISC(0.75) WITHIN GROUP (ORDER BY value)
       - PERCENTILE_DISC(0.25) WITHIN GROUP (ORDER BY value) AS IQR
  FROM Measurement WHERE sensor='temp';
```

# Week6

**Observational Study**:

- Simply observing what happens. **No intervention**
- Records information about subjects without applying any treatments to subjects (passive participation of researcher).
- ex. analyzes customer by using their transaction records

**Experimental Study**

- 核心判断依据：**有 intervention，即对 independent variable 有操作**
- Records information about subjects while applying treatments to subjects and controlling study conditions to some degree (active participation of researcher)
- **不能只用 Control group 和 Treatment group 作为依据**
- ex. A researcher gives one group of patients a new medication and another group a placebo, then compares the outcomes to determine the drug's effectiveness.

**Main difference between observational studies and experiments**

- Most experiments use random assignment while observational studies do not.
- 确定因果
  - **Observational** studies typically **only establish correlation** but not causality
  - **Experimental** studies establish **causality**

**Independent Variable**

The variable that is **manipulated** to observe its effect on the dependent variable. The one thing you change in experiment.
*Example: Type of treatment, amount of exercise, temperature.*

**Dependent Variable**

The **measure of interest**; the outcome you observe or measure.
*Example: Blood pressure, test scores, reaction time.*

**Controlled Variables (Constants)**

Conditions or factors **kept the same** throughout the experiment to ensure a fair test.
*Example: Equipment used, environment, measurement method*

## Hypothesis Testing

- use hypothesis testing to specify whether to accept or reject a claim
- Related Definitions
  - **Q**: Asks whether the **independent variable** has an effect on the **dependent variable**.
  - **H0**: Assumes that there is **no effect**. **initial assumption. Means** "no effect" or "no difference."
  - **H1**: Assumes that there **is an effect** or **a difference**.
- **p-value** and **Significance Level (α)**
  - **p-value** 在原假设 H0 为真的前提下，观察到当前样本数据(一般为极端结果)的概率。比如公平硬币连续 10 次都是 back 的概率是 0.5^10 = 0.00098.
  - **α** is the probability of (wrongly) rejecting H0 given that it is true (**Type I error rate**, i.e., false positive) 如果原假设 H0 为真，对于 **α**=0.05 我们有 5% 的概率会错误地拒绝它。

| P-value Compared to α | Indicates | Reject $H_0$? |
|---|---|---|
| *p-value* < α | Strong evidence against the null hypothesis ($H_0$). **代表极端值几乎不可能发生** | ✅ Yes |
| *p-value* > α | Weak evidence against the null hypothesis ($H_0$) | ❌ No |
| *p-value* = α | Marginal (on the boundary) | ⚠️ NA |

  - Example

假设：

- $H_0$：硬币是公平的。
- 你扔 10 次，结果 10 次都是正面。
- 你计算出：在公平前提下，出现这种情况的概率 $p = 0.00098$。

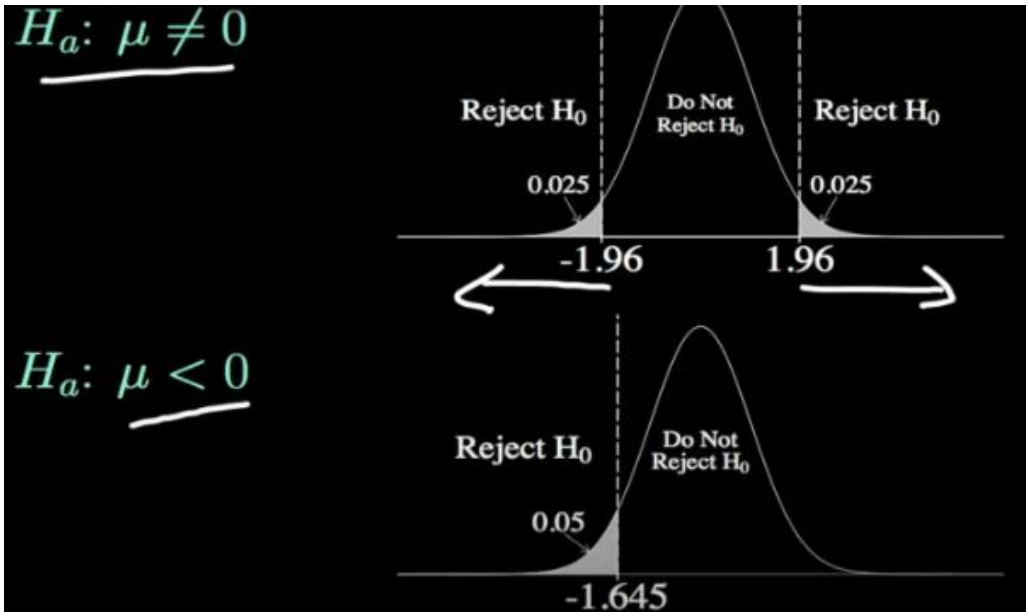
如果你设定 α = 0.05，那么：

> $p = 0.00098 < 0.05$，我们就说："这么极端的结果如果硬币是公平的，几乎不会发生，我们怀疑硬币不公平"，因此**拒绝** $H_0$。

- p-hacking
  - If we perform many tests, we are likely to get false positives
- **Type of Error**

| | Accept $H_0$ | Reject $H_0$ |
|---|---|---|
| $H_0$ is True | ✅ Correct Decision *(No difference)* | ❌ Type I Error *(False positive)* |
| $H_0$ is False ($H_1$ is true) | ❌ Type II Error *(False negative)* | ✅ Correct Decision *(Difference exists)* |

**One-side test VS Two-side**




- Two-side : H1 只能是不等于
- One-side : H1 可以是大于或者小于

**求 P-Value 的方法**

parametric 方法假设数据服从某个特定的分布模型

- 2 组
  - **Pearson's Correlation — parametric**
    - **$H_0$**：两个变量之间没有线性相关关系。
    - Assumption:
      - 变量是连续的且近似服从正态分布。
      - 变量间的关系是线性的。
      - 数据是成对的独立观测。
  - **Paired Student's t-test (pairwise t-test) — parametric**
    - H0: two population means are equal
    - Assumption:

        - The samples are paired (e.g. before and after treatment).
        - Populations are normally distributed.
        - Standard deviations are equal.
    - **Wilcoxon Signed-Rank Test** — **nonparametric**
        - H0: two related paired samples come from the same **distribution**.
        - Assumption:
            - The samples are paired
    - **Unpaired Student's t-test** — **parametric**
        - H0: two population **means** are equal
        - Assumption:
            - The samples are independent.
            - Populations are normally distributed.
            - Standard deviations are equal (by default).
    - **Mann-Whitney U test** — **nonparametric**
        - H0: the **distribution** underlying sample x is the same as that of sample y
        - **A**ssumption:
            - The two samples are independent.
            - Does not require normal distribution.
- 2 组以上
    - Analysis of Variance (**ANOVA**) — **parametric**
        - H0: two or more groups have the same population **mean**.
        - **A**ssumption:
            - the samples are independent.
            - Populations are normally distributed.
            - Standard deviations are equal.
    - **Kruskall-Wallis H-test** — **nonparametric**
        - H0: the population **medians** of all groups are equal
        - **A**ssumption:
            - the samples are independent.

**Holdout & cross validation methods**

- **Holdout:** Splits the data randomly into two independent sets
    - Training set (e.g., 2/3) for model construction.
    - Test set (e.g., 1/3) for accuracy estimation.
    - Repeat holdout k times, accuracy = avg. of the accuracies obtained
- **Cross Validation Methods** (k-fold, where k = 10 is most popular)

- Split data into 10 sets set1,.., set10 of approximately equal size
- A classifier is built 10 times. Each time the testing is on 1 set and the training is on the remaining 9 sets together
- 计算 Avg(acc1,…acc10)

- **LOOCV** (Leave-one-out cross validation)
  - A special form of n-fold cross-validation；其中 n 等于样本的总数

**对比 model 的性能 -- Confusion matrix**

Accuracy

what proportion of cases got the right label

$$\frac{Correctly\ labeled\ cases}{All\ test\ cases} = \frac{TP+TN}{TP+TN+FP+FN}$$

Precision:

what proportion of cases guessed as positive were actually positive

**不希望错判，即 不希望把负类错判成正类**（如正常邮件被错判成垃圾邮件）

不要乱猜错

$$\frac{Cases\ gussed\ as\ positive\ that\ are\ positive}{Cases\ guessed\ as\ positive} = \frac{TP}{TP+FP}$$

Recall

proportion of cases that should be positive were guessed as positive

**不希望漏判，即 不希望把正类错判成负类**

不要漏掉重要的

$$\frac{Cases\ gussed\ as\ positive\ that\ are\ positive}{Cases\ that\ are\ real\ positive} = \frac{TP}{TP+FN}$$

F-Score

harmonic mean of Precision and Recall; bigger is better

$$F = \frac{2}{\frac{1}{P}+\frac{1}{R}} = \frac{2PR}{P+R}$$

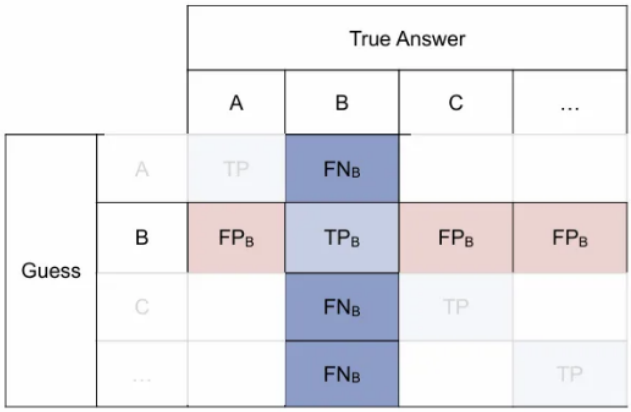
重点关注大于 2-dimensional 的计算

第一个表示是否正确

第二个表示预测的结果，

FP 表示错误的预测为 Positive

FN 表示错误的预测为 Negative




$$\text{Recall for B} = \frac{\text{Cases gussed as B that are B}}{\text{Cases that are B}}$$

$$= \frac{TP_B}{TP_B + FN_B}$$

# Week7

**Supervised learning**

The training data are accompanied by labels indicating the class of the observations

目标是学习一个从输入到输出的映射关系，用于预测未知数据的标签

**Unsupervised learning**

The class labels of training data is unknown

目标

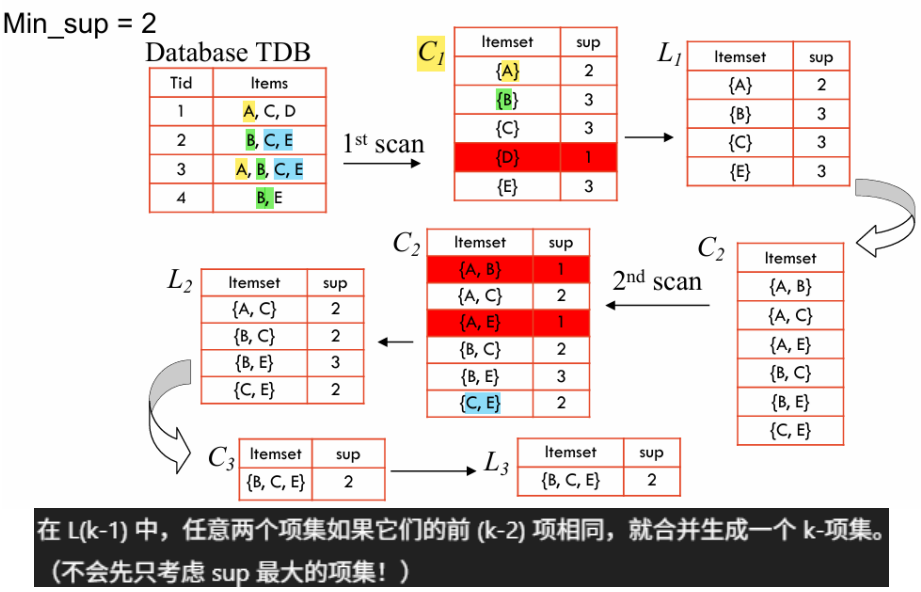
- Establishing the existence of classes or clusters in the data
- Discovering hidden patterns or rules

**Association Rule Mining -- Unsupervised learning**

- **k-itemset**: an itemset containing k items.
- Frequency — support
  - Support Count (σ) 是项集的出现次数（频率）
  - Support (s) 是归一化后的项集频率（也就是比例）
  - Frequent Itemset: 如果一个 itemset 的 s 大于等于 threshold min_support，则 s 就可以被认为是 frequent itemset
- Association Rule — Confidence
  - Confidence（c）衡量的是：在买了 X 的交易中，有多大概率也买了 Y

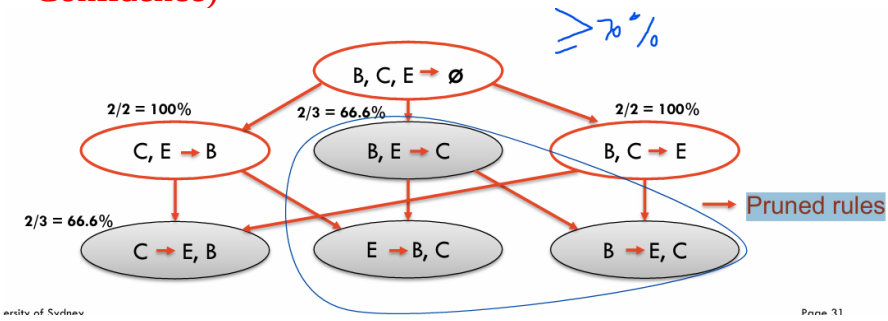
$$c(x \rightarrow y) = \frac{support(X \cup Y)}{support(X)}$$

  - Must satisfy: 如果 c 大于等于 min_conf 我们说 X 和 Y 有强关联
- **Apriori Principle**
  - Bottlenecks:
    - Generate huge candidate frequent sets (2^N)-1
    - Multiple scans of database
  - Def
    - If an itemset is infrequent, then its supersets are also infrequent
    - If an itemset is frequent, then all of its subsets must also be frequent
  - Apriori Algorithm Step (这里只用了 Frequency — support)




所以在生成 C3 的时候，只有{B,C}和{B,E}满足前缀相同

- **Rule Generation**
  - 除去 2 个无效情况（空集和整个集合本身），总共有(2^N)-2 个可能的 rule，因此要根据下面的 Pruned Tree
  - **Pruned rules**:
    - $c(ABC \rightarrow D) \geq c(AB \rightarrow CD) \geq c(A \rightarrow BCD)$
  - 假如我们根据上面的 Apriori Algorithm Step 得到了 frequent itemset {B,C,E}，则按照下面的步骤得到最后的 Rule，这里假设 70%的 Confidence threshold，并且用到了(Association Rule — Confidence)




**FP Tree (Frequent-Pattern Tree) — unsupervised learning**

这个建议还是看看 example 更加直接

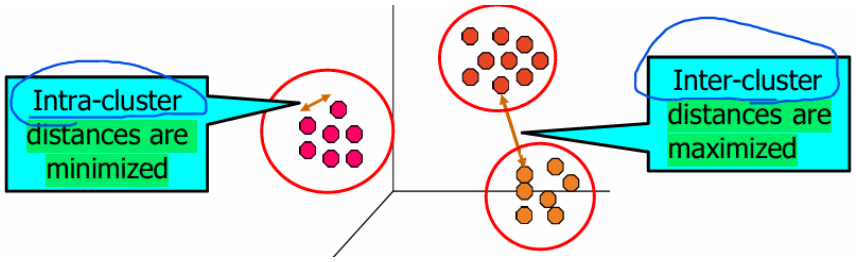
也是用来求 Frequent Itemset，之后可以再求 rule generation

- Avoid costly repeated database scans
- divide-and-conquer methodology

# Week8

## Clustering — unsupervised

- Intra and Inter Cluster




- Similarity and Dissimilarity
  - Distances are normally used to measure the similarity dissimilarity between two data objects.

Minkowski Distance

$$d(i,j) = \sqrt[q]{(x_{i1} - x_{j1})^q + (x_{i2} - x_{j2})^q + \ldots + (x_{ip} - x_{jp})^q}$$

Manhattan Distance — q=1

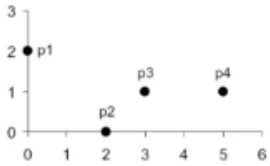
$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \ldots + |x_{ip} - x_{jp}|$$

Euclidean Distance (L2) — q=2

$$d(i,j) = \sqrt[2]{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ldots + (x_{ip} - x_{jp})^2}$$

  - Dissimilarity matrix: expresses the pairwise dissimilarities (distances) between points

| point | x | y |
|---|---|---|
| p1 | 0 | 2 |
| p2 | 2 | 0 |
| p3 | 3 | 1 |
| p4 | 5 | 1 |




Pairwise Euclidean Distances:

1. p1-p2
   $\sqrt{(2-0)^2 + (0-2)^2} = \sqrt{4+4} = \sqrt{8} \approx 2.83$
2. p1-p3
   $\sqrt{(3-0)^2 + (1-2)^2} = \sqrt{9+1} = \sqrt{10} \approx 3.16$
3. p1-p4
   $\sqrt{(5-0)^2 + (1-2)^2} = \sqrt{25+1} = \sqrt{26} \approx 5.10$
4. p2-p3
   $\sqrt{(3-2)^2 + (1-0)^2} = \sqrt{1+1} = \sqrt{2} \approx 1.41$
5. p2-p4
   $\sqrt{(5-2)^2 + (1-0)^2} = \sqrt{9+1} = \sqrt{10} \approx 3.16$
6. p3-p4
   $\sqrt{(5-3)^2 + (1-1)^2} = \sqrt{4+0} = \sqrt{4} = 2.00$

| L2 | p1 | p2 | p3 | p4 |
|---|---|---|---|---|
| p1 | 0 | 2.828 | 3.162 | 5.099 |
| p2 | 2.828 | 0 | 1.414 | 3.162 |
| p3 | 3.162 | 1.414 | 0 | 2 |
| p4 | 5.099 | 3.162 | 2 | 0 |

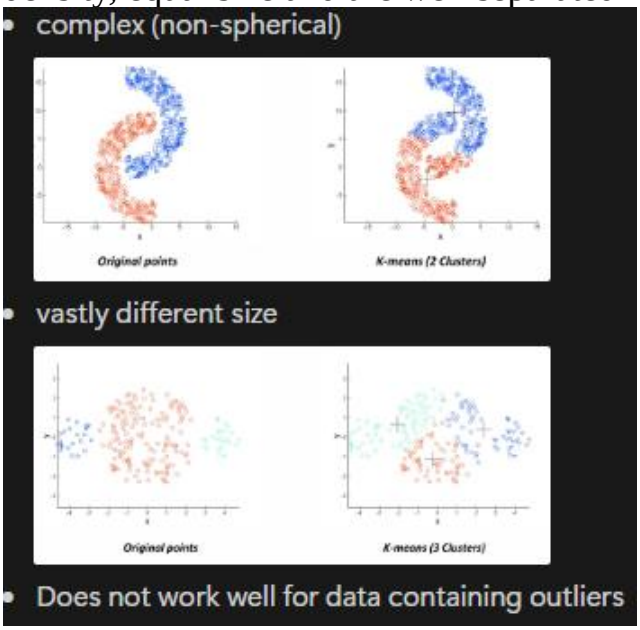
- Partitional clustering: A division of data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset
  - K-Means $O(n \times k \times i \times d)$

$n$ = number of points, $k$ = number of clusters,
$i$ = number of iterations, $d$ = number of attributes (or dimensions)

**3 steps**

1. choose $k$ examples as the initial centroids of the clusters; typically chosen randomly
2. Form $k$ clusters by assigning each example to the ***closest*** centroid
3. At the end of each epoch:
   a. re-computer the centroid of the clusters; 一般使用mean来计算
   b. check if the stopping condition is satisfied (i.e. centroids do not change)
      i. if yes stop
      ii. if no, repeat above step 2 and 3

- K-means works well if the clusters are spherical, of equal density, equal size and are well separated




- Most of the convergence happens in the first few iterations
- 个 **centroid** 是一个**向量的均值**，也就是说，它是属于**特征空间中的一个“点”**，但不需要刚好是数据集中某个具体的点。
- Outlier need be removed

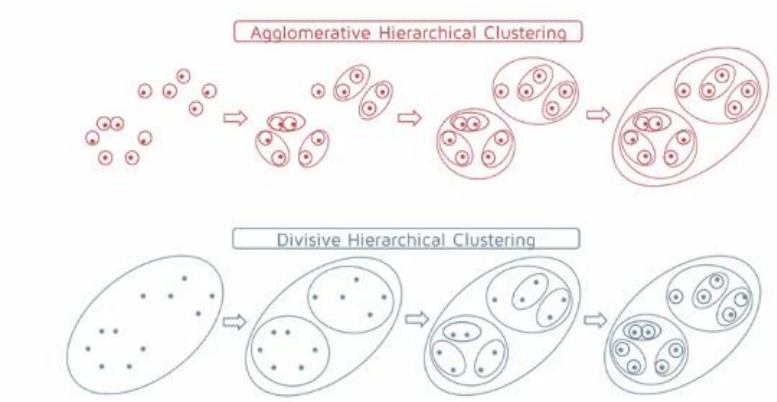
- **K-means++**

- 避免 centroids 选择不好（k-means 的 random 选择有可能不好）；如果 centroids 不移动，则可以停止了。

**K-means++ Initialization Steps:**

1. **Randomly pick the first centroid**
   随机从数据中选出第一个中心（centroid）。
2. **Compute distance for each point**
   对于数据集中每个点 $x$，计算它到最近已选中心的距离平方（通常记为 $D(x)^2$）。
3. **Select next centroid probabilistically**
   根据距离平方值 $D(x)^2$ 确定概率，选择下一个中心：
   - 概率 $p(x) = \frac{D(x)^2}{\sum_x D(x)^2}$。
   - 即，离现有中心越远的点，更有可能被选为下一个中心。
4. **Repeat until k centroids are chosen**
   重复步骤 2-3，直到选出 $k$ 个初始中心。
5. **Run standard K-means**
   使用这 $k$ 个中心作为 K-means 算法的初始化，然后按普通 K-means 进行迭代聚类。

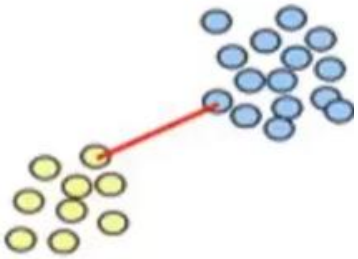
- **Hierarchical clustering**: A method of cluster analysis which seeks to build a hierarchy of clusters. It produces a set of nested clusters organized as a hierarchical tree
  - Time complexity: O(n^3)
  - Space complexity: O(n^2)
  - No need to specify the number of clusters



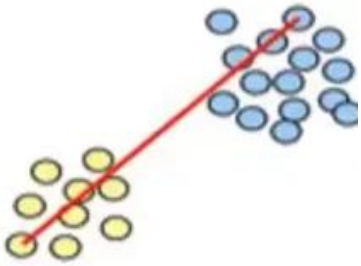

  - Agglomerative (bottom-up) – merges clusters iteratively

1. Compute the proximity matrix
2. Let each data point be a cluster
3. **Repeat**
4. Merge the two closest clusters
5. Update the proximity matrix
6. **Until** only a single cluster remains

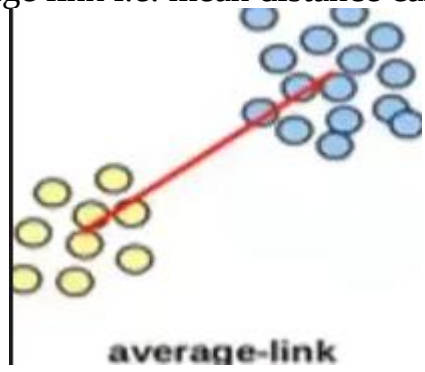
- Single link (MIN)




- Complete link (MAX)




- Average link i.e. mean distance calculation




Hierarchical clustering does not require predefining cluster number and reveals data hierarchy via dendrograms.

**Measures of Cluster Validity**

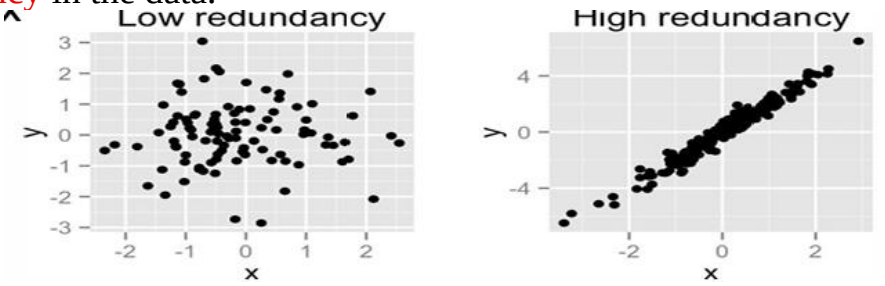
- **External** (值越高越好)
  - **Homogeneity**: ranges from 0 to 1，每个聚类中的数据点是否都属于同一个真实类。preferring each cluster contains only members of a single class, analogous to precision, P = TP / (TP+FP)
  - **Completeness**: ranges from 0 to 1，同一类的所有数据点是否都被分到同一个聚类中。preferring all members of a given class are assigned to the same cluster, analogous to recall, R = TP / (TP+FN)
  - **V-measure**: 是 Homogeneity 和 Completeness 的**调和平均数**，综合这两个指标
- **Internal**
  - SSE

    $$SSE = \sum_{i=1}^{k} \sum_{x \in K_i} dist(x, c_i)^2$$
  - Silhouette coefficient
    - The silhouette coefficient measures how similar an object is to its own cluster compared to other clusters. It combines **cohesion** (how close the object is to other points in its own cluster) and **separation** (how far it is from points in other clusters).
    - The closer to 1, the better; -1 表示非常差
    - s = 1–a/b if a<b
    - s = b/a-1 if a>=b
    - a 是 average distance of x to points in its cluster
    - b 是 average distance of x to points in the next nearest cluster
    - 我们要分别对某一个 cluster x 求里面每个点的 a 和 bs；最后计算均值，得到 cluster x 的质量，如果接近 1 则表明分类的非常好
- **可以自己设置一个 threshold，如果所有 cluster 的 Silhouette Coefficient 达标，则说明 k 选的很好。**

**Dimensionality reduction (PCA) — unsupervised**

- we used Euclidean Metric so far (L2-Norm)
- but others possible too, eg. Manhatten Distance (L1-Norm)

- PCA method is particularly useful when the variables within the data set are highly correlated. Because Correlation indicates that there is redundancy in the data.




- **Covariance Matrix**

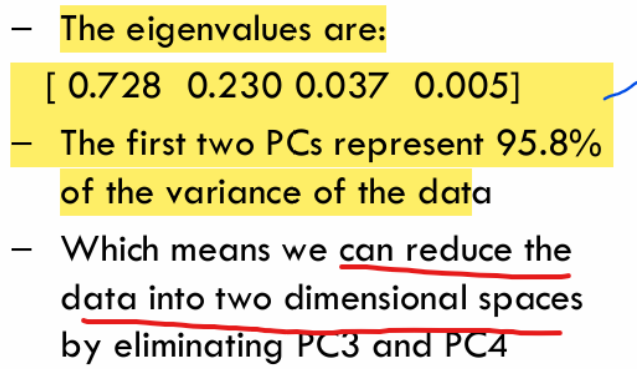
| | | |
|---|---|---|
| 1.343730 | -.1601522 | .1864702 |
| -.1601522 | .61920562 | -.1266842 |
| .1864702 | -.1266842 | 1.485549 |

  - off-diagonal values are **different from zero**. This indicates the presence of redundancy in the data. 因为 redundant，所以要进行 PCA
    - COV 为正：当 X 增大时，*Y*也倾向于增大。
    - 负：当 *X*增大时，Y 倾向于减小。
    - 接近零：两者之间几乎没有线性关系（不一定表示独立!

- PCA Matrix
  - PCA creates uncorrelated PC variables (called eigenvectors) having zero covariations and variances (called eigenvalues, 1.65 1.22. 0.57) sorted in decreasing order.

| | | |
|---|---|---|
| 1.65135 | .000000 | .000000 |
| .000000 | 1.220288 | .000000 |
| .000000 | .0000000 | .576843 |

  - 经过 PCA 后，可以发现 PCA component1 占到了 1.651/3.448 = 47.9% of the overall variability
  - 我们可以通过删除占比低的 PCA components，达成 reduce attributes dimension 的目地




    - 比如比较经典的，保留占 95%重要程度 principal components (PC) 的方法
- 可以发现 Covariance Matrix 和 PCA Matrix 的 diagonal 之和是相同的

# Week9

## Simple Linear Regression

$$Y = \alpha + \beta X + \varepsilon$$

- It is simple because it only has one independent variable X
- Method for finding the line of best fit between the dependent variable Y and the independent variable X
- β: slope; Regression coefficient
- α: intercept
- ε: error, SSE

$$\epsilon = SSE = \sum (y_i - \hat{y}_i)^2;\ \text{这里}y_i\text{是target，}\hat{y}_i\text{是prediction}$$

- lower SSE means a better fit
- Ordinary Least Squares (**OLS**)

通过最小化误差项的平方和来确定回归系数（$\alpha$ 和 $\beta$）。即发现Loss function

$$\frac{1}{n}\sum_{i=1}^{n}\left((\hat{\alpha}+\hat{\beta}x_i)-y_i\right)^2 \text{的minimum}$$

其中$\hat{\beta}$，$\hat{\alpha}$的计算公式如下

$$\hat{\beta} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2} = \frac{\operatorname{cov}(x,y)}{\operatorname{Var}(x)} = \frac{r(x,y)\cdot \operatorname{sd}(y)}{\operatorname{sd}(x)}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

这里$r$是指correlation

- **SSE** (Sum of Squared Errors)

$$SSE = \sum_i (y_i - \hat{y}_i)^2$$

  - Measures the difference between the real value and the model's predicted value
  - **Smaller** values indicate **better** predictive accuracy.
- **SSR** (Sum of Squared Regression)

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

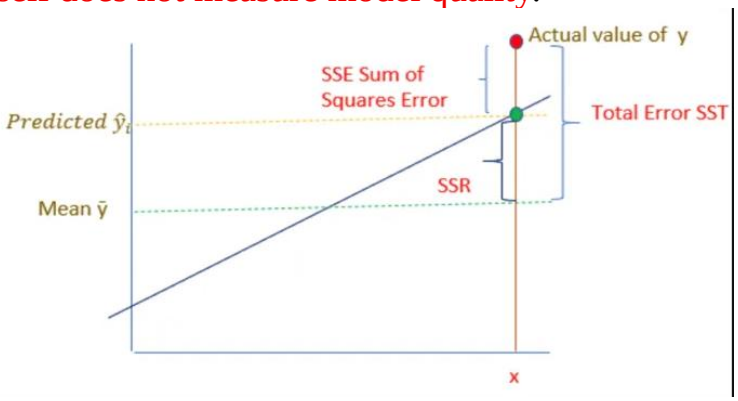
  - Measures the difference between the model's predicted value and the mean of the real value.
  - **Higher** values indicate a **better** model fit, as more variance is explained.
- **SST** (Sum of Squared Total)

- Represents the total dispersion of the dependent variable around its mean (**SST=SSR+SSE**).
- Smaller values mean the data points are closer to the mean of y, but SST itself does not measure model quality.

$$SSE = \sum_i (y_i - \hat{y}_i)^2$$

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$




- **Coefficient Determination (R^2)**

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{SSE}{SST}$$

=SSR/SST

  - 是一个衡量回归模型拟合优度的指标。它表示的是**自变量 x 对因变量 y 的解释能力**。
  - Ranges from 0 to 1, with higher values indicating better fit for the linear model
  - 当 SSE>SST 的时候会有 negative R^2 表示 model is very poor, perform worse than the mean of the data
  - Conveys goodness of fit but not precision

**Multiple linear regression**

$$Y = h_\theta(X) = \alpha + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_d X_d + \varepsilon = \sum_{j=0}^{d} \theta_j X_j = \theta^T X$$

- It is NOT simple because it has more independent variable X_i
- **Cost Function**; we can fit model by minimize it

$$MSE = J(\theta) = \frac{1}{2n} \sum_{i=1}^{n} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

  - $h_\theta(x^{(i)})$是第i个样本的预测值
  - $y^{(i)}$是第i个样本的真实值

- **How to minimize MSE? Always converge to Global min**
  - Calculate Gradient descent for each theta respectively.

$$\theta_j \leftarrow \theta_j - \alpha \cdot \frac{\partial}{\partial \theta_j} J(\boldsymbol{\theta})$$

这里α代表learning rate，而我们用导数进行梯度下降。下面是梯度的计算公式

$$\frac{\partial}{\partial \theta_j} J(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \left( h_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}) - y^{(i)} \right) x_j^{(i)}$$

- 如果 derivative of J 是正数
  - 如果导数是正数，说明坡度朝上，你站的位置右侧是上坡。你想走下坡找到最低点，所以不能往上坡方向走（不能往导数正的方向走），而是得往反方向（左边）走。
- Add non-linearity

$$Y = h_{\boldsymbol{\theta}}(\mathrm{x}) = \theta_0 + \theta_1 \mathrm{x}^1 + \theta_2 \mathrm{x}^2 + \ldots + \theta_d \mathrm{x}^{\mathrm{d}} = \sum_{j=0}^{d} \theta_j \mathrm{x}^{\mathrm{j}}$$

- **Overfitting**: 只在 training data work well
  - **Regularization**: aims to penalize for large values of coefficients ($\theta\_j$) to reduce overfitting 是对代价函数（cost function）添加一个**惩罚项（penalty term）**; No regularization on $\theta$_0

$$J(\boldsymbol{\theta}) = \underbrace{\frac{1}{2n} \sum_{i=1}^{n} \left( \mathrm{h}_{\boldsymbol{\theta}}(\mathrm{x}^{(\mathrm{i})}) - \mathrm{y}^{(\mathrm{i})} \right)^2}_{\text{Model fit to data}} + \underbrace{\lambda \sum_{j=1}^{d} \theta_j^2}_{\text{regularization}}$$

$\lambda$ is the regularization parameter ($\lambda \geq 0$)

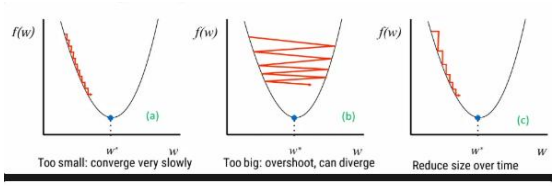
  - 当 λ 较小时，正则化项的影响较小，模型可能更复杂，但容易过拟合；当 λ 较大时，正则化项会抑制权重，使模型更简单，但可能欠拟合。
  - Can also address overfitting by eliminating features (either manually or via model selection)，比如去除噪音等。

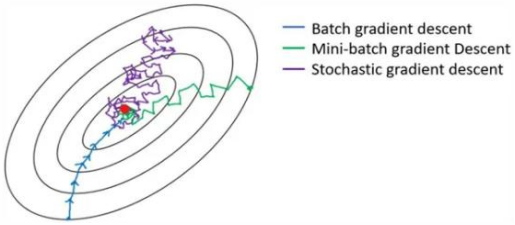
**Gradient Descent**

- If $\alpha$ is small, gradient descent can be slow
- If $\alpha$ is too large, gradient descent might overshoot the minimum

Steps:

- Choose an initial value for $\theta$
- Iteratively choose a new value for $\theta$ to reduce $J(\theta)$ (e.g., through gradient descent)
- Repeat Until we reach a minimum $J(\theta)$




Batch:

- Repeat until converge $\left\{\theta_j \leftarrow \theta_j - \alpha\frac{1}{n}\sum_{i=1}^{n}\left(h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right) - y^{(i)}\right)x_j^{(i)}\right\}$, for every $j$
- Accurate but slow:
  - has to scan through the entire training set before taking a single step
  - costly operation if $n$ is large

Stochastic:

For $i = 1$ to $n$

$$\left\{\theta_j \leftarrow \theta_j - \alpha\left(h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right) - y^{(i)}\right)x_j^{(i)}\right\}, \text{ for every } j$$

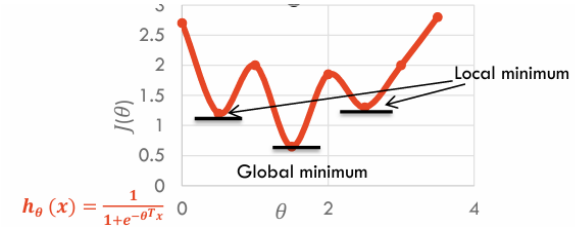
- Fast, start making progress right away.
- It may not converge to the minimum

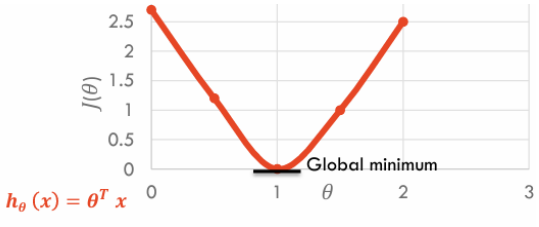
可以看到，Batch GD 会很慢，但是比较精确，因为用了整个 dataset。而 Stochastic GD 会很快，但是不一定能 converge，因为只用了 random row from dataset。

**Logistic Regression**

$$g(Z) = h_\theta(x) = \frac{1}{1 + e^{-(W^T X^{(i)} + b)}}$$

- 不一定 converge to global min，因为有很多 local min（看左图）




- 虽然名字带有 regression，但其实是一种 classification 方法
  - **Regression** assigns a numerical value, Output is a continuous variable

    - **Classification** assigns a class to each example, Output is a discrete / categorical variable
- g(Z)返回的结果位于 0-1 之间，而 **Binary classification** 的分类依据如下

这里$g(Z) = h_\theta(x) = \frac{1}{1+e^{-(W^T X^{(i)}+b)}}$

  - $Z = W^T X^{(i)} + b$ can be a large positive or negative value while $y$ is Yes or No (0 or 1) in the case of binary classification
  - $Z = W^T X^{(i)} + b \geq 0 \rightarrow g(Z) \geq 0.5 \rightarrow \hat{y} = 1$
  - $Z = W^T X^{(i)} + b < 0 \rightarrow g(Z) < 0.5 \rightarrow \hat{y} = 0$

- **Multi-class classification**
    - One-vs-rest strategy
        - 训练多个（K 个）逻辑回归模型。
        - 每个模型判断是否是某一个类。
        - 最后选出输出概率最大的那个类。
    - SoftMax function
        - 一个模型，对于每个 label 都有 probability，从中选择概率最大的那一个作为输出 label
- **Cross-Entropy** (即 cost function)

  - $\hat{y}_i = \sigma(W^T X^{(i)} + b)$
  - 偏置$b$是模型的一部分

$$L = f(w) = -\frac{1}{n}\sum_{i=1}^{n}[y_i log\sigma(W^T X^{(i)} + b) + (1 - y_i) log[1 - \sigma(W^T X_i + b)]]$$

    - 这里可以看到当 yi=0，则只有第 2 个 term 有效；如果 yi=1，则只有第 1 个 term 有效

# Week10

## Decision Tree

- non-leaf node: each non-leaf node corresponds to a test for the value of an attribute (i.e. outlook)
- branch: Attributes value (i.e. windy = false)
- leaf node: each leaf node represent a class (labeled attribute)(i.e. play)
- a top-down recursive divide-and-conquer manner
- Entropy H(S)= $-\sum_i P_i log_2 P_i$
  - higher entropy => higher uncertainty.
  - Lower entropy => lower uncertainty.
  - The smaller the entropy, the greater the purity of the data set.
    - 对于 2个类别，最大熵是 $\log_2 2 = 1$。
    - 对于 3个类别，最大熵是 $\log_2 3 \approx 1.585$。
    - 对于 4个类别，最大熵是 $\log_2 4 = 2$。
- 如何用 Information Gain 来选择 node？(判断分的是否好)
  - Gain = T1-T2
  - Higher Gain is better, so we need to choose attribute with higher Gain as next node

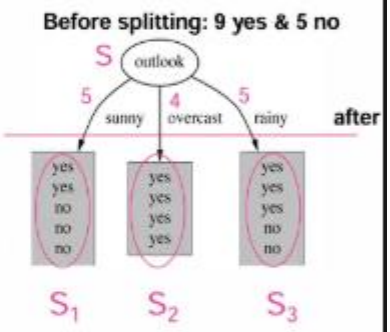
Example

下面计算量对于attribute `outlook` 的 Information Gain；如果要和别的attribute对比，则对于别的attribute重复下面的计算步骤，最后比较information gain，最后选择最小Gain的attribute。其中$i$是不同的class

$$H(S) = I(S) = -\sum_i P_i log_2 P_i$$

- Let's calculate the information gain of the attribute *outlook*
- *T1* is the entropy of the set of examples S before the split:

$$T1 = H(S) = I(\frac{9}{14}, \frac{5}{14}) = 0.940\ bits$$




- T2 is the remaining entropy in S, after S is split by the attribute
- It takes into consideration the entropies of the child nodes and the distribution of the examples along each child node
  - e.g. for a split on outlook, it will consider the entropies of S1, S2 and S3 and the proportion of examples following each branch (5/14, 4/14, 5/15):

$$T2 = H(S \mid outlook) = \frac{5}{14} \cdot H(S_1) + \frac{4}{14} \cdot H(S_2) + \frac{5}{14} \cdot H(S_3)$$

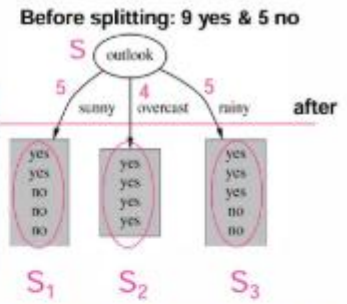
现在来计算$T_2$，这里sunny的weight是5/14，因为从14个data中分了5个到该branch

$$T2 = H(S \mid outlook) = \frac{5}{14} \cdot H(S_1) + \frac{4}{14} \cdot H(S_2) + \frac{5}{14} \cdot H(S_3)$$

$$H(S \mid outlook = sunny) = I(\frac{2}{5}, \frac{3}{5}) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.971\,bits$$

$$H(S \mid outlook = overcast) = I(\frac{4}{4}, \frac{0}{4}) = -\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4} = 0\,bits$$

$$H(S \mid outlook = rainy) = I(\frac{3}{5}, \frac{2}{5}) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} = 0.971\,bits$$




$$H(S \mid outlook) = \frac{5}{14} \cdot 0.971 + \frac{4}{14} \cdot 0 + \frac{5}{14} \cdot 0.971 = 0.693\,bits$$

$$Gain(S|outlook) = H(S) - H(S|outlook) = 0.940 - 0.693 = 0.247 \text{ bits}$$

| outlook | temp. | humidity | windy | play |
|---|---|---|---|---|
| sunny | hot | high | false | no |
| sunny | hot | high | true | no |
| overcast | hot | high | false | yes |
| rainy | mild | high | false | yes |
| rainy | cool | normal | false | yes |
| rainy | cool | normal | true | no |
| overcast | cool | normal | true | yes |
| sunny | mild | high | false | no |
| sunny | cool | normal | false | yes |
| rainy | mild | normal | false | yes |
| sunny | mild | normal | true | yes |
| overcast | mild | high | true | yes |
| overcast | hot | normal | false | yes |
| rainy | mild | high | true | no |

16

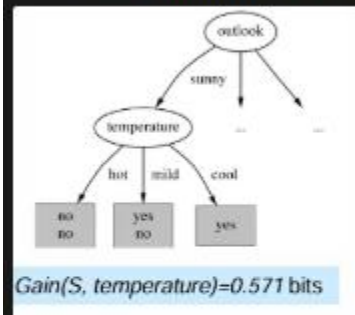
我们省略其它3个attributes 的计算步骤，直接进行比较部分，发现outlook的information gain最大

- Similarly, the information gain for the other three attributes is:

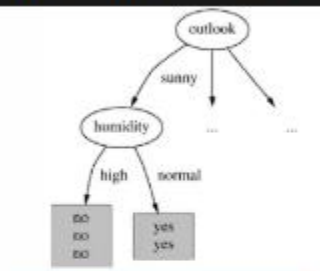
  *Gain(S|temperature)=0.029 bits*
  *Gain(S|humidity)=0.152 bits*
  *Gain(S|windy)=0.048 bits*

- => we select outlook as it has the highest information gain

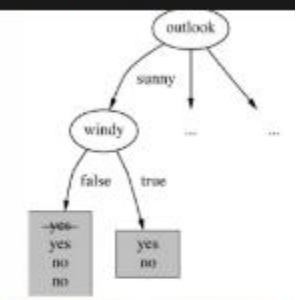
选择完第一个node后，我们还需要对于branches（i.e. sunny，rainy）继续选择node（不考虑overcast是因为这个branch里面全是一个class），也是通过对比Information Gain，其中用到的data是split到branch的data。



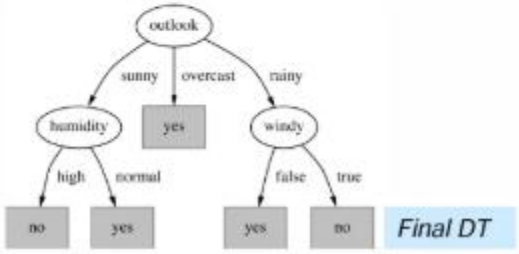

Gain(S, temperature)=0.571 bits




Gain(S, humidity)=0.971 bits



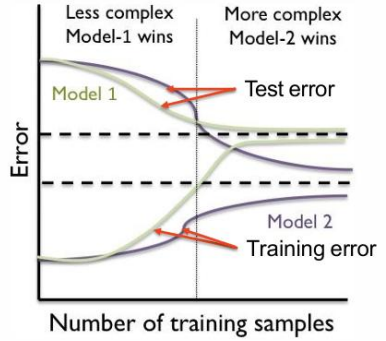

Gain(S, windy)=0.020 bits



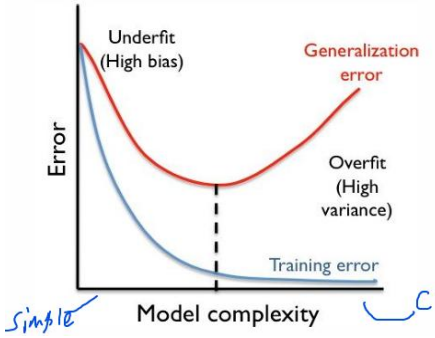

Final DT

| outlook | temp. | humidity | windy | play |
|---|---|---|---|---|
| sunny | hot | high | false | no |
| sunny | hot | high | true | no |
| overcast | hot | high | false | yes |
| rainy | mild | high | false | yes |
| rainy | cool | normal | false | yes |
| rainy | cool | normal | true | no |
| overcast | cool | normal | true | yes |
| sunny | mild | high | false | no |
| sunny | cool | normal | false | yes |
| rainy | mild | normal | false | yes |
| sunny | mild | normal | true | yes |
| overcast | mild | high | true | yes |
| overcast | hot | normal | false | yes |
| rainy | mild | high | true | no |

# Complex model VS simple model




- Limited data → simple model is better
- enough data → complex model is better




- The dashed line on right shows point where we switch from under-fitting to overfitting.
- **Goal**: Find this dotted line.
- Generalization error should model application as closely and reliably as possible.
  - Sample must be representative.
  - Larger sample better.

- **Training Error**：这是模型在**训练数据集**上的误差，即模型对它“见过的数据”的预测有多准确
- **Generalization Error**：这是模型在**未见过的数据（测试集）上的误差，也就是模型对新样本的预测能力**

**Data drift (non-stationary data)**

模型在生产环境中，因输入数据分布变化而性能下降的现象。比如客户的爱好变化。

- 训练集和测试集假设数据分布稳定（stationary），但在生产环境中这种假设只短时间成立。Typical train/test setups assume stationarity
- Only near true in production for a little while

解决办法有：

- Monitor offline metric on live data
- May require monitoring/annotation
- If there are large changes, then retrain on new data
- Online/incremental learning：通过**在线学习**（模型持续实时更新）或**增量学习**（分批更新模型参数）方式，让模型能够**动态适应数据变化**，减少手动干预。

**Feature Engineering（特征工程）**

简单来说，就是从原始数据中提取、转换和构造新的特征，以便让机器学习模型更好地理解和利用数据。

具体来说，特征工程包括：

- **选择**：从已有数据中挑选对预测目标有用的变量（特征）。
- **转换**：对原始数据进行处理，比如归一化、标准化、离散化、编码（如类别变量转数字）等。
- **构造**：根据业务知识或者数据规律，创建新的特征，比如用两个变量相乘、计算时间差、提取文本关键词等。
- **清洗**：处理缺失值、异常值，确保特征质量。
- **降维**：减少特征数量，去除冗余，提高模型效率。

**Data Preprocessing（数据预处理）**

目标：　把原始数据变得干净、规范，适合模型训练。

常见内容:

- 处理缺失值（填补、删除）
- 处理异常值（剔除、修正）
- 数据清洗（去重、纠正错误）
- 数据格式转换（比如字符串转数字）
- 数据标准化、归一化（使不同特征处于同一量纲）
- 编码分类变量（one-hot 编码、标签编码）
- 分割数据集（训练集、测试集）

## Ensemble of predictors（集成预测器）

- **集成学习**是指将多个模型（预测器）组合起来共同做决策，通常能得到比单个模型更好的效果。
- 例如：**投票法（Voting）**，多个分类器对样本进行投票，最终选票数最多的类别作为预测结果。
- **Random Forest（随机森林）**
  - 原理：对训练数据进行 bootstrap 采样（有放回抽样）生成多个训练子集，针对每个子集训练一棵决策树。和 boosting 不同的是“每个决策树的每个节点**随机选取部分特征”**
  - 最后将所有树的预测结果综合（通常是投票）得到最终预测。
  - **优点**
    - 更加稳定，即 lower variance
  - **缺点**
    - More biased
    - Lose explainability of trees

**Subtractive feature analysis**

- Assess impact of each feature by removing it.
- If performance goes up, it’s not a good feature.
- BIGGER negative means have higher influence

# Week11

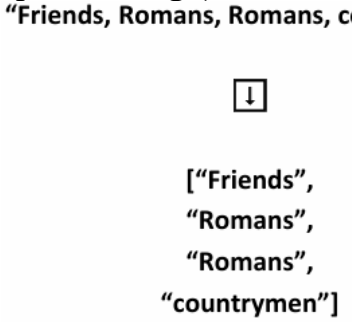
## Structured data VS Unstructured Data

| 项目 | Structured Data | Unstructured Data |
| --- | --- | --- |
| 格式 | 表格、模式清晰 | 无固定结构 |
| 存储 | 关系型数据库 | 文件、云存储 |
| 分析工具 | SQL, Excel 等 | NLP, AI, 图像识别等 |
| 可扩展性 | 处理较简单 | 处理更灵活但复杂 |
| 占比 | 数据总量中较少 | 占 80% 以上的企业数据 |

There are more unstructured data than structured data

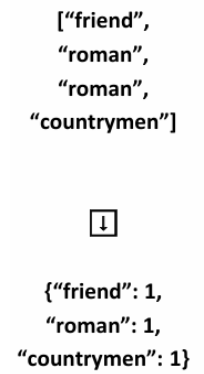
## Bag of words

- **Tokenization**: Split a string (document) into pieces called tokens




- **Normalization**: Map similar words to the same token
  - Stemming/lemmatisation
    - E.g., "was" => "be"
  - Lower casing, encoding
    - E.g., "Naïve" => "naive"


- **Indicator features**: Binary indicator feature for each word in a document；如果 token 在文档中出现，赋值 1，否则赋值 0




- **Term frequency weighting**: Give more weight to terms that are common in documents
  - TF = |occurrences of term in doc|
- **Damping**: Sometimes want to reduce impact of high counts
  - TF = log(|occurrences of term in doc|)

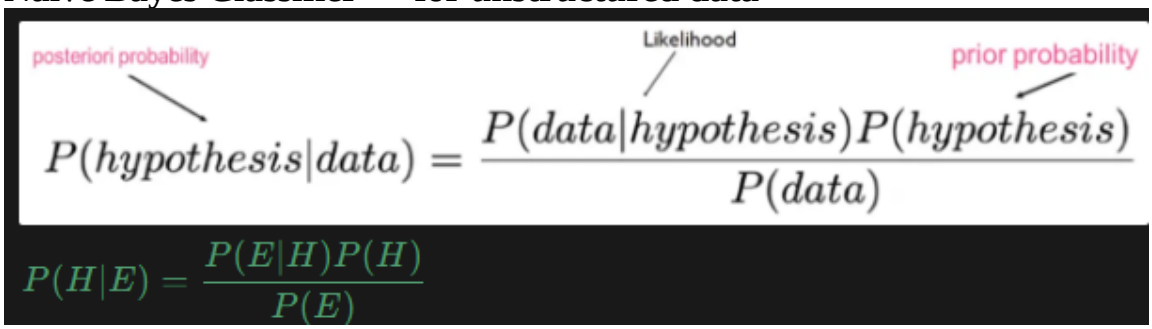

- **TFIDF**: Reduces the weight of common words across multiple documents while increasing the weight of rare words. If a term appears in fewer documents, it is more likely to be meaningful and specific.

TFIDF = TF * IDF

$$IDF(t, D) = \log \frac{\text{Total number of documents in corpus D}}{\text{Number of documents containing term t}}$$

“Total number of documents in the corpus (D)”意思是指在整个语料库中包含的文档总数。

**Naïve Bayes Classifier — for unstructured data**




- Assumption: no relation between attributes — why it is called naïve
  - Correlated attributes reduce the power of Naïve Bayes,
- **ZERO-Frequency** (i.e. 如果某个 P(E_i|H) = 0 怎么办)
  - 用 Laplace correction

$$p(w|c) = \frac{\text{count(w,c)}+1}{\text{count(word,c)}+\text{count(word)}}$$

- count(w,c) = Count of word w in class c
- count(word,c) = Count of all words in class c
- count(word) = Count of all distinct words in the dataset

**Text-driven forecasting**

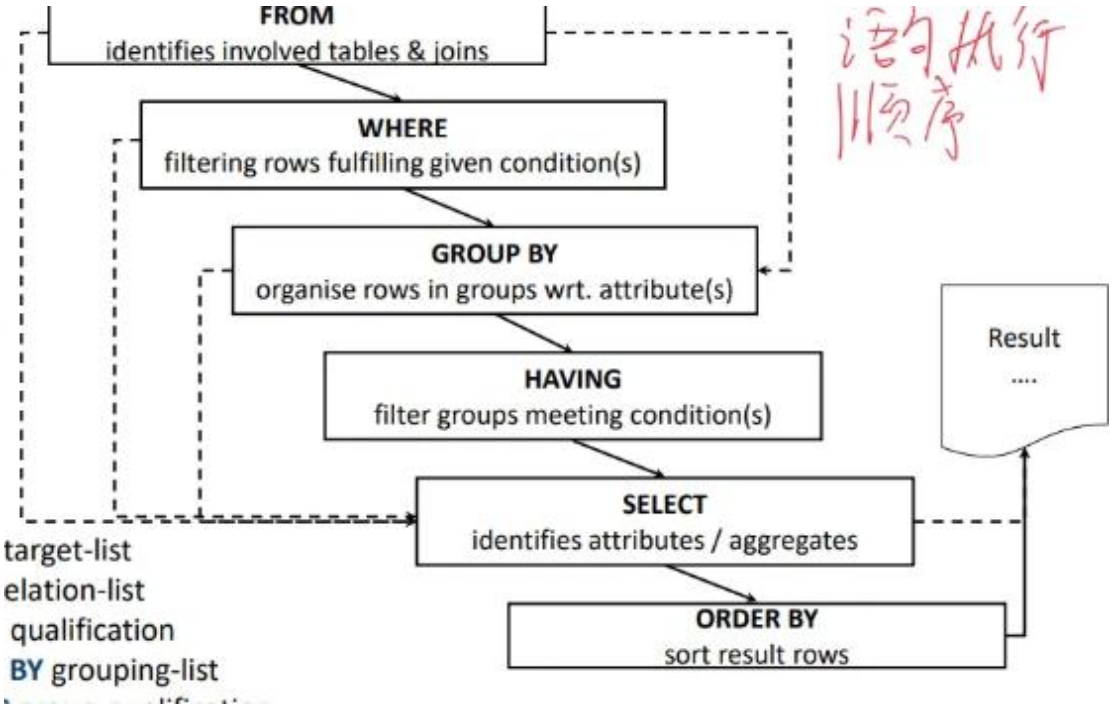
Given a body of text T pertinent to a social phenomenon, make a concrete prediction about a measurement M of that phenomenon, obtainable only in the future.

**Some text-driven forecasting tasks**

- Predict box office gross for films
  - T: description, script, reviews, etc
  - M: how much the film earns at the box office
- Predict volatility of a stock
  - T: annual report, etc
  - M: volatility over the following year
- Predict blog reader behaviour
  - T: political blog posts, etc
  - M: number of reader comments

# Syntax

SQL 语句执行顺序




## Aggregate function

- COUNT(*)计算 NULL，别的都不算，包括 COUNT(x)
- 不能 MIN(AVG())
- 都对 DUPLICATE 起效，除非 COUNT(DISTINCT x)
- If you use GROUP BY function, then in SELECT or HAVING line, you can only include aggregate function or the attribute you use for GROUP BY
- Predicates in the HAVING clause are applied after the formation of groups, whereas predicates in the WHERE clause are applied before forming groups

## Syntax

CREATE TABLE NAME(…);

DROP TABLE NAME CASCADE;

INSERT INTO table (list-of-columns) VALUES (list-of-expression);

UPDATE table SET column = expression WHERE search_condition;

DELETE FROM table WHERE search_condition;

JOIN

- Implicit join is Cartesian join/CROSS join
  - 类似于 FROM A, B
- Inner join，特殊写法 using 替代 on; 省略 Inner
  - Theta Join（θ -Join）
  - Equi-Join 等值连接
  - Natural Join
    - Natural Join Caveat

Natural Join Caveat：如果两个表中没有相同的列名，则 NATURAL JOIN 会返回笛卡尔积（不推荐）。

- $R \bowtie S = \pi_{unique\_attributes}(\sigma_{equality\_of\_common\_attributes}(R \times S))$

- Outer join
  - LEFF JOIN
    - NATURAL LEFT OUTER JOIN: 不需要 ON 语句，它会自动匹配两个表中相同名称的列，并以左表（Employee）为主表. 结果可能和 LEFT JOIN 不同，因为 NATURAL JOIN 可能匹配多个同名列，而 LEFT JOIN 只会匹配 ON 指定的列。
  - RIGHT JOIN
  - FULL JOIN: 不会返回重复的行。在右表没有孤儿行的情况下等同于 LEFT JOIN
- Natural Join vs. Explicit Equi-Join: 如果用 explicit equi-join，会导致连接字段出现 2 次,除非用 USING 替代 ON

# Exercise

# Week2

## 求 interval 或者 ratio data 的 Percentile

Suppose you have the following data which is a simple series of 10 random numbers [41,18,95,62,33,25,77,89,12,50]

Step 1:- First you need to sort the data in ascending order:

[12,18,25,33,41,50,62,77,89,95]

Step 2:

note the number of data points (N) : 10

Step 3:- Find the index for that percentile:

Index = Percentile_Value_In_Decimal * (N - 1).
Index = 0.10 * (10 - 1) = 0.1 * 9 = 0.9.

Step 4:- Calculate percentile value:-

Value at Index 0 + 0.9*(value at index 1 - value at index 0)

12 + 0.9 * (18 - 12) = 12 + 5.4 = 17.4.

Therefore, the 10th percentile of the dataset is 17.4

## 分析 data type 和应该使用什么类型的 plot

1. The table below contains students' grades and the number of students who achieved that grade for a specific unit of study. What is the data type of the "Grades" column and what type of chart would you use to analyse the distribution of the dataset?

| Grades | Frequency |
|---|---|
| HD | 11 |
| DI | 24 |
| CR | 28 |
| P | 16 |
| F | 6 |

In the table above, the data type of the "Grades" column is "Ordinal

Data" because the grades mentioned have a specific ordering /ranking to them where the ranks of each grade is defined as follows: HD > DI > CR > P > F. The best type of graph that can be used to analyse the distribution of the dataset is to create a bar chart of the table above because a bar chart can help visualise some key aspects of the data, such as which grade was the most common and which was the least. It can also help in comparing data points in different grade categories which will provide a holistic view of how the students performed.

这里注意的是，最好结合每个 datatype 的特点进行描述，比如上面说的 ordinal 有 rank 而 nominal 是没有的；或者 ratio 的 zero 是 defined，但是 interval 的不是，等。

**分析数据类型以及合适的分析方法（这里强调的是 central tendency）**

4. Identify the data type (nominal, ordinal, interval, or ratio) for each of the following variables and state which measures of central tendency would be appropriate.

    a) Education level (High school, Bachelor's, Master's, PhD)

    b) Annual income in dollars

    c) Customer satisfaction rating (1-5 stars)

    d) Zip codes

    e) Temperature in degrees Celsius

4. a) Ordinal data - appropriate measures: median, mode
   b) Ratio data - appropriate measures: mean, median, mode
   c) Ordinal data - appropriate measures: median, mode
   d) Nominal data - appropriate measure: mode only
   e) Interval data - appropriate measures: mean, median, mode

# Week3

## 计算 Boxplot 中的 outlier

Question Using the IQR method, identify the outliers in the dataset:

-20, -3, -2, -1, 0, 1, 2, 3, 4, 5, 8, 9, 10, 11, 12, 13, 15, 16, 19, 40, 50

**Data Points:** n = 21
**Complete Dataset:** -20, -3, -2, -1, 0, 1, 2, 3, 4, 5, 8, 9, 10, 11, 12, 13, 15, 16, 19, 40, 50
**Quartiles:** $Q_1$ = 1, $Q_2$ (median) = 8, $Q_3$ = 13
**Interquartile Range:** IQR = $Q_3$ - $Q_1$ = 13 - 1 = 12

**Lower Fence:** L = $Q_1$ - 1.5×IQR = 1 - 1.5×12 = -17
**Upper Fence:** U = $Q_3$ + 1.5×IQR = 13 + 1.5×12 = 31
**Outliers:** Values < -17 or > 31 (-20, 40, 50)
**Non-outlier Range:** -17 to 31

这里 Q1 = 0.25*(21-1) = 5,而 idx5 对应的是 1

Q1. What type of plot is most suitable to visualise and analyse the correlation between two variables? Why?

A scatterplot is the most suitable visualisation to observe and analyse the correlation between two variables because it shows the strength and direction of the correlation between two observed variables.

Q2. Given the boxplot below, calculate Q1, Q2, Q3, IQR, Upper Inner, Lower Inner Fence. Also give examples of outlier data points.

a. Q1 = 7

b. Q2 = Median = 20

c. Q3 = 32

d. Max = 47

e. Min = 1

f. IQR = Q3 – Q1 = 32 – 7 = 25

g. Upper Inner Fence = Q3 + (1.5 * IQR) = 32 + (25 * 1.5) = 69.5

h. Lower Inner Fence = Q1 – (1.5 * IQR) = 7 – (25 * 1.5) = -30.5.

i. Outlier data points would be any data points that are beyond either the lower inner fence or upper inner fence.

Q3. A researcher analyzes the relationship between age and income, where income increases rapidly with age up to a certain point and then plateaus. What would you expect if the researcher uses Pearson correlation instead of Spearman correlation.

The relationship between **age and income** is **non-linear**: it increases rapidly and then **levels off** (plateaus). Pearson correlation only captures linear relationships, so once the income plateaus (i.e., stops increasing with age), the linear association weakens. As a result, Pearson correlation may return a relatively low correlation coefficient, even though there is a strong underlying monotonic relationship (i.e., income never decreases with age).

Since income increases monotonically with age (even if the rate of increase slows), Spearman would capture this more accurately, resulting in a higher correlation coefficient than Pearson in this case.

# Week4

Q1. Explain why the following table does not qualify as a relation and describe a potential solution to fix the issue.

| SID | Student Name | Student Contact | City |
|---|---|---|---|
| 87292 | Mark Anthony | +61788292342 ; +61282662834 | Sydney |
| 91623 | Tim Berrel | +61639715239 ; +61427882639 | Sydney |
| 64872 | Violet Aura | +61912213572 ; +61732215543 | Sydney |
| 42881 | Karl Brown | +61528193732 ; +61633037253 | Sydney |

对于这种题，我们应该使用 RDBMs 判断准则?

这里因为 atomic 不符合, 即 Student Contact 有 2 种; one potential solution is to create another table that maps SID to Student Contact number.

Q2. Consider the following tables below. Write the SQL query to:

a. Create both the tables.

Movie Details Table:

| Movie_ID | Movie_Title | Year_of_Release |
|---|---|---|
| 1001 | The Avengers: Age of Ultron | 2015 |
| 1002 | Harry Potter and the Sorcerer's Stone | 2001 |
| 1003 | Percy Jackson & the Olympians: The lightning Thief | 2010 |

Sales Order Table:

| Order_ID | Date_Of_Purchase | Store_ID | Movie_ID |
|---|---|---|---|
| 0001 | 11/11/2018 | 9403AF | 1001 |
| 0002 | 10/02/2019 | 9403AF | 1003 |
| 0003 | 25/06/2016 | 4572ZH | 1003 |
| 0004 | 04/12/2023 | 9403AF | 1002 |
| 0005 | 29/07/2020 | 4572ZH | 1002 |
| 0006 | 13/01/2024 | 4572ZH | 1003 |

CREATE TABLE Movie(

```
    Movie_ID INT PRIMARY KEY,
    Movie_Title VARCHAR(255),
    Year_of_Release INT
);
CREATE TABLE Sales(
    Order_ID CHAR(4) PRIMARY KEY,
    Date_Of_Purchased DATE,
    Store_ID VARCHAR(6),
    Movie_ID INT,
    FOREIGN KEY (Movie_ID) REFERENCES Movie (Movie_ID)
);
```

Q3. You are hired as a data modelling expert by Star Truck Suppliers. Given the immense amount of data, how would you decide whether to use a star schema or snowflake schema to use for data modelling.

The decision to either implement a snowflake schema or a star schema really depends on the business's needs.

A **snowflake** schema is used when you want to **store data more efficiently** and **perform detailed analyses** on the dataset. This type of schema typically requires a lot of joins which may **impact the query processing speed**. Snowflake Schema has **lower redundancy** due to normalization, which avoids repeated data.

On the other hand, if the business requires **fast query** processing speeds and does not have a very complex dataset, then the star schema is the optimal choice. **Star Schema** has **Higher redundancy** in dimension tables because of denormalization.

可以看到数据量对于选择 star 和 snowflake 并无明显的影响

Q4.

Consider the following Orders and Customers tables:

**Customers Table:**

| customer_id | customer_name | email |
|---|---|---|
| **1** | John Doe | john@example.com |
| **2** | Jane Smith | jane@example.com |
| **3** | Bob Brown | bob@example.com |

**Orders Table:**

| order_id | customer_id | product_name | order_date |
|---|---|---|---|
| **101** | 1 | Laptop | 2024-07-12 |
| **102** | 2 | Smartphone | 2024-07-15 |
| **103** | 1 | Tablet | 2024-07-18 |
| **104** | 3 | Headphones | 2024-07-20 |

1. Identify the primary key in both tables. Explain why you selected those keys.
2. What is the difference between a primary key and a foreign key?
3. Write an SQL command to create these tables

The “customer_id” is chosen as the primary key because it uniquely identifies each customer in the Customers table. No two customers will have the same “customer_id”. The “order_id” is chosen as the primary key because it **uniquely identifies** each order in the Orders table. **Each order has a distinct** “order_id”, ensuring that **no two orders are the same**.

PK is not null and unique, while FK can be null and have duplicates

```
CREATE TABLE Customers (
    customer_id INT PRIMARY KEY,
    customer_name VARCHAR(100) NOT NULL,
```

```
    email VARCHAR(100) NOT NULL UNIQUE
);
CREATE TABLE Orders (
    order_id INT PRIMARY KEY,
    customer_id INT,
    product_name VARCHAR(100) NOT NULL,
    order_date DATE NOT NULL,
    FOREIGN KEY (customer_id) REFERENCES Customers(customer_id)
);
```

Q5.
You have two tables, employees and departments, in your database. The employees table contains the following columns: employee_id, first_name, last_name, department_id, and salary. The departments table contains the columns: department_id and department_name. Write a SQL query to retrieve the first_name of each employee along with the department_name they belong to.

```
SELECT
    first_name, department_name
FROM
    employees
NATURAL JOIN
    departments
;
```

# Week5

Q1. Write a SQL query to find the first_name, last_name, and the number of days each student has been enrolled. Hint: don't forget the constants that exist in the database. The database schema consists of a single table, students. This table stores each students first_name, last_name and enrolment_date.

```
SELECT
    first_name,
    last_name,
    CURRENT_DATE – enrolment_date
From
    students
;
```

Q2. Write a SQL query to select the first_name and hire_date of all employees who were hired in the year 2023. The database schema consists of a table, employees. It includes the first_name, last_name and hire_date of all employees.

```
SELECT
    First_name,
    Hire_date
FROM
    employees
WHERE
    EXTRACT(YEAR FROM hire_date) = 2023;
```

Q3. We have Employee table and Department table as shown below.

| | employeeid [PK] integer | firstname character varying (50) | lastname character varying (50) | departmentid integer | businessid integer |
|---|---|---|---|---|---|
| 1 | 1 | Alice | Smith | 100 | 1 |
| 2 | 2 | Bob | Johnson | 101 | 1 |
| 3 | 3 | Charlie | Williams | 102 | 2 |
| 4 | 4 | David | Jones | 105 | 2 |
| 5 | 5 | Emily | Li | 104 | 4 |

| | departmentid [PK] integer | departmentname character varying (50) | managerid integer | businessid integer |
|---|---|---|---|---|
| 1 | 100 | Sales | 1 | 1 |
| 2 | 101 | Marketing | 2 | 3 |
| 3 | 102 | HR | 5 | 2 |
| 4 | 104 | Finance | 4 | 4 |
| 5 | 105 | Design | 4 | 5 |

**Part 1:** Write a query using an INNER JOIN to retrieve a list of employees and their corresponding department names.

**Part 2:** Write a query using a NATURAL JOIN to retrieve a list of employees and their corresponding department names. Discuss how this result differs from the INNER JOIN.

```
SELECT *
FROM Employee E
INNER JOIN Department D
ON E.departmentid = D.departmentid;
```

```
SELECT *
FROM Employee
NATURAL JOIN Department;
```

这里的不同在于，natural join 会同时使用 departmentid 和 bussinessid，因为两个表都有这些。

**Example 1:** How many measurements were done *per each sensor*?

```
  SELECT sensor, COUNT(*)
    FROM Measurement
   GROUP BY sensor;
```

**Example 2:** How many measurements of *distinct* stations were done *per each sensor*?

```
SELECT sensor, COUNT(DISTINCT station)
  FROM Measurement
 GROUP BY sensor
 ORDER BY count DESC;

SELECT station, sensor, COUNT(*)
  FROM Measurement
 GROUP BY station, sensor
 ORDER BY count DESC;
```

Determine some basic statistics about the measured temperature values **per each station**, including minimum temperature, maximum temperature, range of temperature values, mean, mode, 25th and 75th percentile:

```
SELECT siteName,
       MIN(value),
       Max(value),
       MAX(value) - MIN(value)                                  AS Range,
       AVG(value)                                               AS Mean,
       MODE()  WITHIN GROUP (ORDER BY value)                    AS Mode,
       PERCENTILE_DISC(0.5) WITHIN GROUP (ORDER BY value)       AS Median,
       PERCENTILE_DISC(0.25) WITHIN GROUP (ORDER BY value)      AS Percentile25,
       PERCENTILE_DISC(0.75) WITHIN GROUP (ORDER BY value)      AS Percentile75,
       PERCENTILE_DISC(0.75) WITHIN GROUP (ORDER BY value)
       - PERCENTILE_DISC(0.25) WITHIN GROUP (ORDER BY value) AS IQR
  FROM Measurement NATURAL JOIN Station
  WHERE sensor ='temp'
 GROUP BY siteName;
```

```
#1 In which time period were all the measurement done?
query_stmt = "SELECT min(date), max(date) FROM Measurement;"
query_result = pgquery (conn, query_stmt, None)
print("Q1:", query_result)

#2 At how many distinct Stations were used? Which Stations?
query_stmt2 = "SELECT COUNT(DISTINCT station) FROM Measurement;"
query_result2 = pgquery (conn, query_stmt2, None)
print("Q2_1:", query_result2)

query_stmt2 = """SELECT DISTINCT station, sitename
FROM Measurement NATURAL JOIN Station;"""
query_result2 = pgquery (conn, query_stmt2, None)
print("Q2_2:", query_result2)

#3 Do the same statistical analysis for temprature measurements as above, but for just measurements from the year 2005
query_stmt3 = """SELECT DISTINCT station, sitename
FROM Measurement NATURAL JOIN Station
WHERE extract(year FROM date) = 2005 and
sensor = 'temp';"""
```

# Week6

1. You are a data scientist working for an e-commerce company. The marketing team wants to know whether offering a discount on products increases the average time users spend browsing the website. To investigate this, you randomly select 200 users and split them into two groups: one group sees discounted prices, and the other sees regular prices. After one week, you collect data on the average time spent on the website by users in both groups.

   What type of statistical study is this, and why?
   What is the research question being investigated?

Experimental Study, because the researcher is actively **manipulating the independent variable discount.**

The research question is: "whether offering a discount on products affect the average time users spend browsing the website".

2. Assume you have a simple binary classification model that classifies whether a given house classifies as a mansion or a regular house (not a mansion). Given below is the confusion matrix of the model's results. Based on the matrix below, calculate the accuracy, precision, recall, and F1 measures.

| | | MODEL PREDICTIONS | |
|---|---|---|---|
| | | Mansion | Not Mansion |
| ACTUAL RESULTS | Mansion | 40 | 30 |
| | Not Mansion | 25 | 35 |

这里是比较简单的，为 2-dimensional；还需要知道 multi-dimensional 怎么做. 还有就是注意这里列是 actual，row 是 predict

TP FN

FP TN

整体的

Accuracy = (40+35)/(40+30+25+35)=

默认求正类（Mansion）的

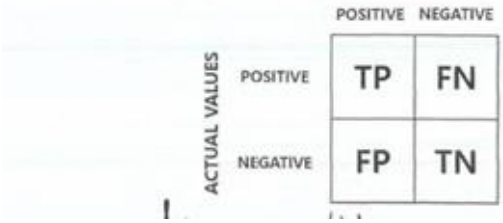
Precision = 40/(40+25) = 0.6154

Recall = 40/(40+30) = 0.5714

F1 = (2*0.6154*0.5714)/( 0.6154+0.5714)=0.6088

**Question 3. Performance Metrics. (5 points)**

Given a dataset consisting of 360 images, with 220 images labeled as chickens and 140 images labeled as dogs. A k-Nearest Neighbors (kNN) algorithm is applied to classify these images. Out of the total images, the model predicts 100 images as dogs. Of these, 80 images are correctly labeled as dogs, while 20 images are misclassified as chickens. Calculate Precision, F1 Score, Recall, and Accuracy of the model.




$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Precision=80/100=0.8
Recall=80/140=0.57
F1=2.0.45/1.37=0.66
Acc=280/360=0.77

| | Act dog | Act chicken | |
|---|---|---|---|
| Pred dog | 80 | 20 | 100 |
| chicken | 60 | 200 | NA |
| | 140 | 220 | |

1 Which of the following tests is used to compare the means of two independent groups under the assumption of normality and equal variances?

a) Paired Student's t-test

b) Wilcoxon Signed-Rank Test

c) Unpaired Student's t-test

d) Mann-Whitney U test

C

2 The Mann-Whitney U test is a nonparametric equivalent of which parametric test?

a) Paired Student's t-test

b) Analysis of Variance (ANOVA)

c) Unpaired Student's t-test

d) Kruskal-Wallis H-test

C

3 Which test would you use to compare the medians of two related samples when the data does not meet the assumptions of normality?

a) Pearson's Correlation

b) Paired Student's t-test

c) Wilcoxon Signed-Rank Test

d) Kruskal-Wallis H-test

Ð C

4 Pearson's correlation coefficient measures the strength of the

a) Non-linear relationship between two variables

b) Linear relationship between two variables

c) Difference in variances between two groups

d) Difference in means between two groups

B

5 Which test compares the medians of more than two independent samples in a nonparametric way?

a) Kruskal-Wallis H-test

b) Analysis of Variance (ANOVA)

c) Wilcoxon Signed-Rank Test

d) Mann-Whitney U test

A

**6** **The Wilcoxon Signed-Rank test requires that the samples are:**

a) Independent and normally distributed

b) Related/paired and ordinal data

c) Related/paired and continuous data

d) More than two groups

C

**7** **Analysis of Variance (ANOVA) is used to test for differences in means:**

a) Between two independent samples

b) Between paired samples

c) Between more than two independent samples

d) Between medians of independent samples

C

# Week7

| Transaction Records | |
|---|---|
| **Transaction ID** | **Items** |
| #1 | apple, banana, coca-cola, doughnut |
| #2 | banana, coco-cola |
| #3 | banana, doughnut |
| #4 | apple, coca-cola |
| #5 | apple, banana, doughnut |
| #6 | apple, banana, coca-cola |

1. Build the FP-tree using a minimum support ***min_sup*** = 2. Show how the tree evolves for each transaction.

这里要删除 min_sup 小于 2 的 item，另外在 transaction 里面也要删除对应的 item。但是因为这里没有，所以这一步没有显示出来。

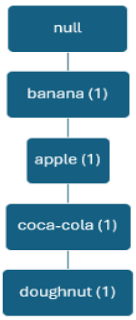
1. Calculate the frequency of each item across all transactions:

banana: 5
apple: 4
coca-cola: 4
doughnut: 3

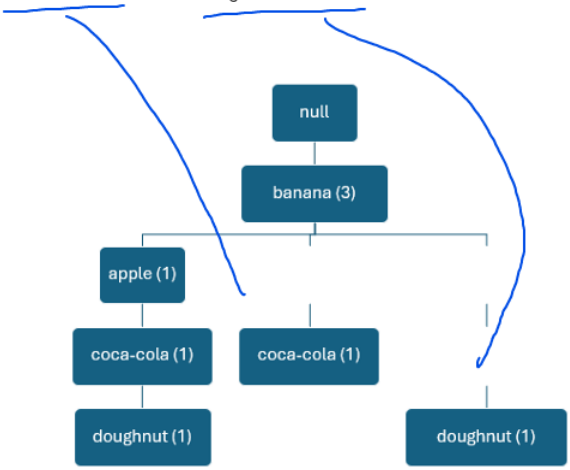
2. Sort the items in each transaction based on frequency:

   Transaction 1: banana, apple, coca-cola, doughnut
   Transaction 2: banana, coca-cola
   Transaction 3: banana, doughnut
   Transaction 4: apple, coca-cola
   Transaction 5: banana, apple, doughnut
   Transaction 6: banana, apple, coca-cola

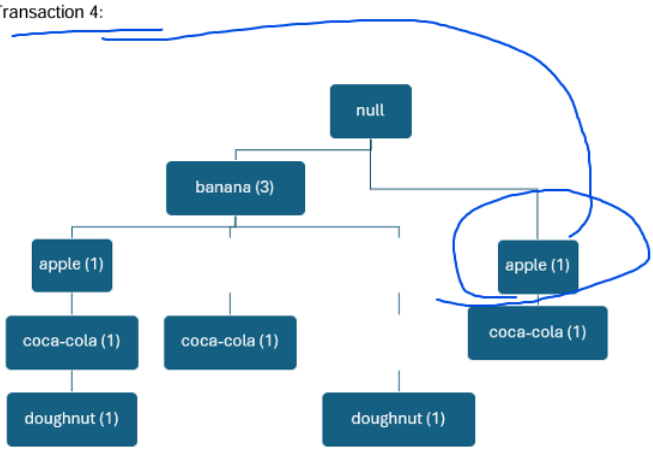
3. Adding Transaction 1:



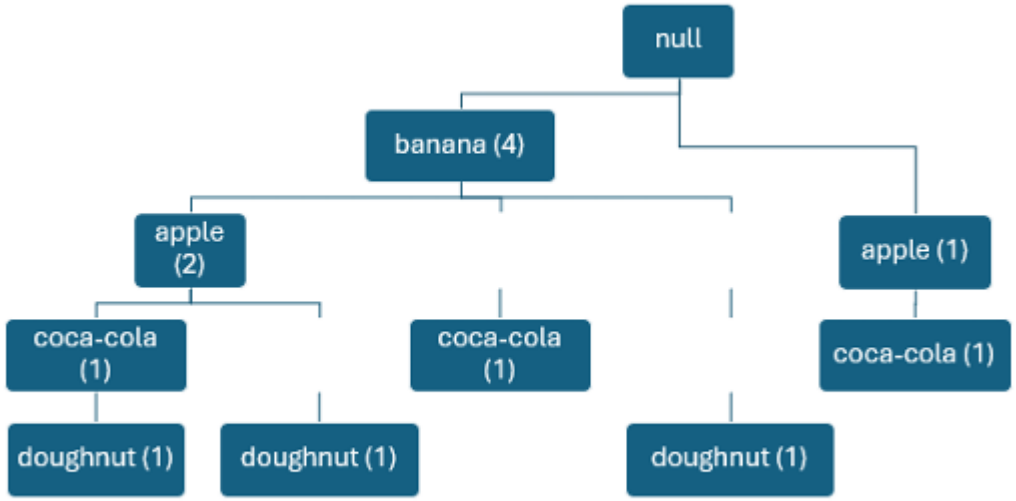

Adding Transaction 2 and then Adding Transaction 3:



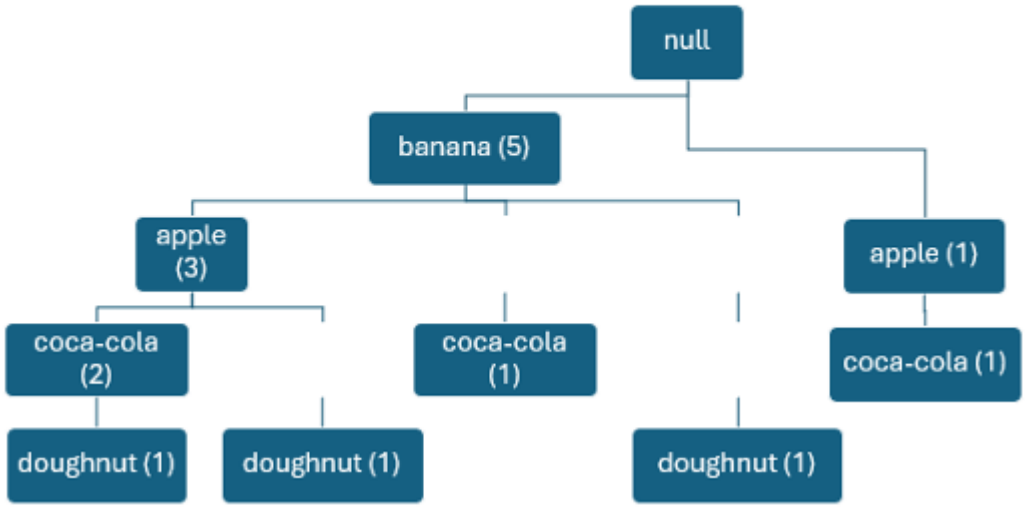

Adding Transaction 4:



Adding Transaction 5:




Adding Transaction 6:



2. With the previous transaction record, Use the Apriori algorithm on this dataset and verify that it will generate the same set of frequent itemsets with **min_sup** = 2.

We'll apply the Apriori algorithm with a minimum support (min_sup = 2) to find the frequent itemsets.

1. Generate Frequent 1-Itemsets:

{banana}: 5
{apple}: 4
{coca-cola}: 4
{doughnut}: 3

2. Generate Frequent 2-Itemsets:

{banana, apple}: 3
{banana, coca-cola}: 2
{banana, doughnut}: 2
{apple, coca-cola}: 3
{apple, doughnut}: 2

3. Generate Frequent 3-Itemsets:

{banana, apple, coca-cola}: 2
{banana, apple, doughnut}: 2

4. Generate Frequent 4-Itemsets:

There are no 4-itemsets that meet the minimum support threshold.

5. Frequent Itemsets using Apriori:

{banana}: 5
{apple}: 4
{coca-cola}: 4
{doughnut}: 3
{banana, apple}: 3
{banana, coca-cola}: 2
{banana, doughnut}: 2
{apple, coca-cola}: 3
{apple, doughnut}: 2
{banana, apple, coca-cola}: 2
{banana, apple, doughnut}: 2

3. Suppose that { **Apple, Banana, Doughnut** } is a frequent item set, derive all its association rules with

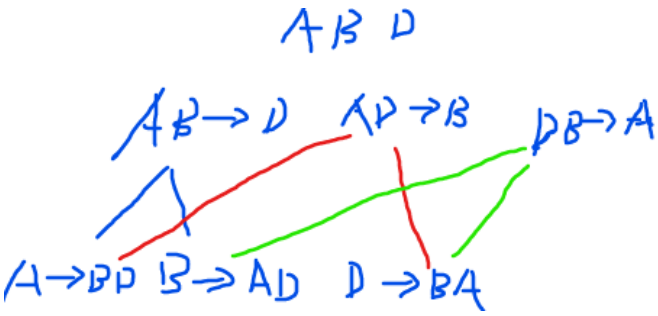
***min_confidence = 70%***

需要通过 FP tree 或者在原始数据中看，不能通过 Apriori

For the frequent itemset {apple, banana, doughnut}, the possible association rules are:

{apple} → {banana, doughnut}
{banana} → {apple, doughnut}
{doughnut} → {apple, banana}
{apple, banana} → {doughnut}

{apple, doughnut} → {banana}
{banana, doughnut} → {apple}



AB->D 和 DB->A 要 prune，但是 D->BA 还需 check

Now, we calculate the confidence for each rule:

$$confidence(\{apple\} \rightarrow \{banana, doughnut\}) = \frac{\{apple, banana, doughnut\}}{\{apple\}} = \frac{2}{4} = 0.50$$

$$confidence(\{banana\} \rightarrow \{apple, doughnut\}) = \frac{\{apple, banana, doughnut\}}{\{banana\}} = \frac{2}{5} = 0.40$$

$$confidence(\{doughnut\} \rightarrow \{apple, banana\}) = \frac{\{apple, banana, doughnut\}}{\{doughnut\}} = \frac{2}{3} \approx 0.67$$

$$confidence(\{apple, banana\} \rightarrow \{doughnut\}) = \frac{\{apple, banana, doughnut\}}{\{apple, banana\}} = \frac{2}{3} \approx 0.67$$

$$confidence(\{apple, doughnut\} \rightarrow \{banana\}) = \frac{\{apple, banana, doughnut\}}{\{apple, doughnut\}} = \frac{2}{2} = 1$$

$$confidence(\{banana, doughnut\} \rightarrow \{apple\}) = \frac{\{apple, banana, doughnut\}}{\{banana, doughnut\}} = \frac{2}{2} = 1$$

Only the following rules meet the minimum confidence threshold of 70%:

{apple, doughnut} → {banana} with 100% confidence.
{banana, doughnut} → {apple} with 100% confidence.

这里 confidence {banaba, doughnut}-> apple 应该是 2/3?是的

# Week8

## K-means

Suppose a dataset consists of the following 1D points:
[1, 2, 3, 10, 11, 12]
Perform K-means clustering with k = 2, using initial centroids as:
•Cluster 1 (C1): 2
•Cluster 2 (C2): 11

**Iteration 1**

**1.Assign points to nearest centroid :**

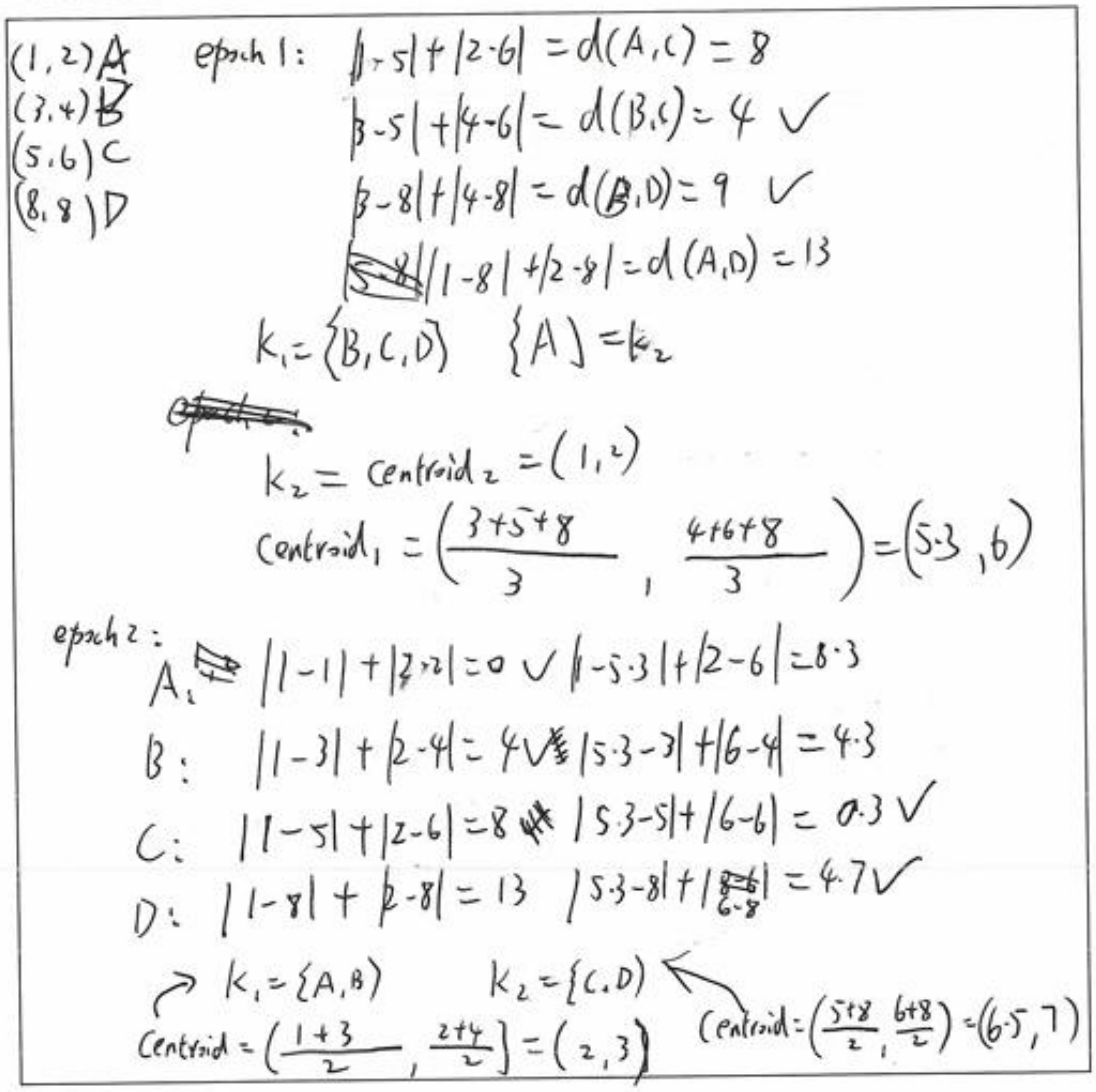
1. For point 1: |1-2| = 1, |1-11| = 10 → Assign to C1
2. For point 2: |2-2| = 0, |2-11| = 9 → Assign to C1
3. For point 3: |3-2| = 1, |3-11| = 8 → Assign to C1
4. For point 10: |10-2| = 8, |10-11| = 1 → Assign to C2
5. For point 11: |11-2| = 9, |11-11| = 0 → Assign to C2
6. For point 12: |12-2| = 10, |12-11| = 1 → Assign to C2

Resulting clusters:
Cluster 1: [1, 2, 3]
Cluster 2: [10, 11, 12]

Step 2: Compute new centroids (mean of points in each cluster)
•New C1 = mean(1, 2, 3) = (1+2+3)/3 = 2.0
•New C2 = mean(10, 11, 12) = (10+11+12)/3 = 11.0
So, updated centroids:
•C1 = 2.0
•C2 = 11.0

Explain the K-means clustering algorithm and perform clustering using K-means for the following dataset of 2-dimensional four data points: (1, 2), (3, 4), (5, 6), (8, 8). Follow these steps: Choose k = 2 and initialize the centroids with (1, 2), (3, 4). Assign each point to the nearest centroid. Update the centroids based on the new clusters. Repeat the process until the centroids stabilize. Provide the resulting clusters and the positions of the centroids after 2 iterations. Use the Manhattan distance formula to calculate the distances between points.

Manhattan distance is defined by: $d((u, v), (p, q)) = |u - p| + |v - q|$.




## Agglomerative

1. Get distance matrix

| | a | b | c | d | e | f |
|---|---|---|---|---|---|---|
| a | 0 | 184 | 222 | 177 | 216 | 231 |
| b | 184 | 0 | 45 | 123 | 128 | 200 |
| c | 222 | 45 | 0 | 129 | 121 | 203 |
| d | 177 | 123 | 129 | 0 | 46 | 83 |
| e | 216 | 128 | 121 | 46 | 0 | 83 |
| f | 231 | 200 | 203 | 83 | 83 | 0 |

2. merge the closest data items

- Next step is to merge the closest data items.
  - In this case: {b , c} are merged.
  - Therefore, the first clustering process generates: {a}, {b, c}, {d},{e},{f}.

| | a | b | c | d | e | f |
|---|---|---|---|---|---|---|
| a | 0 | 184 | 222 | 177 | 216 | 231 |
| b | 184 | 0 | 45 | 123 | 128 | 200 |
| c | 222 | 45 | 0 | 129 | 121 | 203 |
| d | 177 | 123 | 129 | 0 | 46 | 83 |
| e | 216 | 128 | 121 | 46 | 0 | 83 |
| f | 231 | 200 | 203 | 83 | 83 | 0 |

→

| | a | b,c | d | e | f |
|---|---|---|---|---|---|
| a | 0 | ? | 177 | 216 | 231 |
| b,c | ? | 0 | ? | ? | ? |
| d | 177 | ? | 0 | 46 | 83 |
| e | 216 | ? | 46 | 0 | 83 |
| f | 231 | ? | 83 | 83 | 0 |

3. according to either single, complete, or average to update matrix，可以看到新的cluster (b,c) 里面离a最近的是b，而这里我们选择用single，所以新的matrix里面用184进行填充a到(b,c)的距离

| | a | b | c | d | e | f |
|---|---|---|---|---|---|---|
| a | 0 | 184 | 222 | 177 | 216 | 231 |
| b | 184 | 0 | 45 | 123 | 128 | 200 |
| c | 222 | 45 | 0 | 129 | 121 | 203 |
| d | 177 | 123 | 129 | 0 | 46 | 83 |
| e | 216 | 128 | 121 | 46 | 0 | 83 |
| f | 231 | 200 | 203 | 83 | 83 | 0 |

→

| | a | b, c | d | e | f |
|---|---|---|---|---|---|
| a | 0 | 184 | 177 | 216 | 231 |
| b, c | 184 | 0 | 123 | 121 | 200 |
| d | 177 | 123 | 0 | 46 | 83 |
| e | 216 | 121 | 46 | 0 | 83 |
| f | 231 | 200 | 83 | 83 | 0 |

4. repeat until when there is only one cluster

**SSE**

Suppose we have 3 clusters:

- Cluster 1: [2, 4] with centroid at 3
- Cluster 2: [5, 6, 7] with centroid at 6
- Cluster 3: [8, 10, 12] with centroid at 10

Squared error for each cluster:

- $SE1 = (2-3)^2 + (4-3)^2 = 1 + 1 = 2$
- $SE2 = (5-6)^2 + (7-6)^2 = 1 + 1 = 2$
- $SE3 = (8-10)^2 + (12-10)^2 = 4 + 4 = 8$

$SSE = SE1 + SE2 + SE3 = 12$

**Silhouette Coefficient**

**这里要求 Cluster 1 的质量**

- Suppose we have 3 clusters:
  - Cluster 1 = [ [1,0], [1,1] ]
  - Cluster 2 = [ [1,2], [2,3], [2,2], [1,2] ],
  - Cluster 3 = [ [3,1], [3,3], [2,1] ]
- Take a point [1,0] in cluster 1
- Calculate its average distance to all other points in its cluster, i.e. cluster 1
- So a1 = $\sqrt{(1-1)^2 + (0-1)^2} = \sqrt{0+1} = 1$

先求 a1 (对于 Cluster 1 的第一个点)

Now for the point [1,0] in cluster 1 calculate its average distance from all the points in cluster 2 and cluster 3.

Of these take the minimum average distance.

So for cluster 2:

- [1,0] → [1,2], distance = $\sqrt{(1-1)^2 + (0-2)^2} = \sqrt{0+4} = 2$
- [1,0] → [2,3], distance = $\sqrt{(1-2)^2 + (0-3)^2} = \sqrt{1+9} = 3.16$
- [1,0] → [2,2], distance = $\sqrt{(1-2)^2 + (0-2)^2} = \sqrt{1+4} = 2.24$
- [1,0] → [1,2], distance = $\sqrt{(1-1)^2 + (0-2)^2} = \sqrt{0+4} = 2$

Therefore, the average distance of point [1,0] in cluster 1 to all the points in cluster 2 = (2+3.16+2.24+2)/4 = 2.35

Similarly, for cluster 3.

- [1,0] → [3,1], distance = $\sqrt{(1-3)^2 + (0-1)^2} = \sqrt{4+1} = 2.24$
- [1,0] → [3,3], distance = $\sqrt{(1-3)^2 + (0-3)^2} = \sqrt{4+9} = 3.61$
- [1,0] → [2,1], distance = $\sqrt{(1-2)^2 + (0-1)^2} = \sqrt{1+1} = 1.41$

Therefore, the average distance of point [1,0] in cluster 1 to all the points in cluster 3 = (2.24+3.61+1.41)/3 = 2.42

Now, the minimum average distance of the point [1,0] in cluster 1 to the other clusters 2 and 3 is,

b1 =2.35 (2.35 < 2.42)

现在得到 b1，这里会从 c1 点 1 到 c2 或 c3 的距离中选择最小的作为代表；再计算 s1

So the silhouette coefficient of point [1,0] in cluster 1

s1= 1-(a1/b1) = 1- (1/2.35) = 1-0.43 = 0.57

后续还要求 c1 中其它的点的 s。最后 average them to calculate the overall silhouette coefficient to evaluate the resultant cluster

The closer to 1 the better

# Week9

## Simple linear regression

Q: If you have a multiple regression model for house prices with coefficients:

- $\beta_0 = 50{,}000$
- $\beta_1 = 100$ (for square footage)
- $\beta_2 = 5{,}000$ (for number of bedrooms)

Write out the full equation of the model.

House Price = 50,000 + 100(square footage) + 5,000(number of bedrooms)

## Multiple linear regression Gradient descent

$$\hat{y}^{(i)} = w_0 + w_1 x_1^{(i)} + w_2 x_2^{(i)}$$

有三个点(x1,x2,y): (1,1,4),(2,1,5),(1,2,6)

初始 w 为：(0,0,0)

Aplha 为 0.1

---

这里只看第一个 epoch

先获取 predictions: 0, 0, 0 for all 3 points

Update w0 for x_0

$$\frac{\partial J}{\partial w_0} = \frac{1}{3}(-4 + (-5) + (-6)) = \frac{-15}{3} = -5$$

这里-4=第一个样本的预测值-第一个样本的实际值=0-4

而 x_0 = 1 的，是一个 constant

$$w_0 := 0 - 0.1 \cdot (-5) = 0 + 0.5 = 0.5$$

Update w1 for x_1

$$\frac{\partial J}{\partial w_1} = \frac{1}{3}(-4 \cdot 1 + (-5) \cdot 2 + (-6) \cdot 1) = \frac{1}{3}(-4 - 10 - 6) = \frac{-20}{3} \approx -6.6667$$

$$w_1 := 0 - 0.1 \cdot (-6.6667) = 0 + 0.66667 \approx 0.667$$

Update w2 for x_2

$$\frac{\partial J}{\partial w_2} = \frac{1}{3}(-4 \cdot 1 + (-5) \cdot 1 + (-6) \cdot 2) = \frac{1}{3}(-4 - 5 - 12) = \frac{-21}{3} = -7$$

$$w_2 := 0 - 0.1 \cdot (-7) = 0 + 0.7 = 0.7$$

## 📈 Simple Linear Regression

### ✅ 单选题

In simple linear regression, which of the following is assumed? ~~C~~ D

- a) The errors are independent and normally distributed.
- b) The errors have constant variance (homoscedasticity).
- c) There is a linear relationship between X and Y.
- d) All of the above.

In simple linear regression, which term measures the strength and direction of the linear relationship between X and Y? ~~B~~ C

- a) Intercept
- b) Slope
- c) Correlation coefficient
- d) Mean squared error

In simple linear regression, what does the slope $\beta_1$ represent?
a) The expected value of $y$ when $x = 0$
b) The average change in $y$ when $x$ increases by 1 unit
c) The square of the residuals
d) The log odds of $y$

B

In simple linear regression, which of the following is minimized to find the best-fitting line?
a) The residual sum of squares (RSS)
b) The correlation coefficient
c) The log-likelihood
d) The log odds

A

🔟 What does it mean if the residuals show a clear pattern in a residual plot?
a) The model is perfect
b) A linear model may not be appropriate
c) The data has no noise
d) The residuals are normally distributed
✅ 答案：b)

### ✅ 判断题

The coefficient of determination R^2 can be negative in simple linear regression. T
(True / False)

The slope coefficient in simple linear regression is sensitive to outliers.
(True / False) F T

### ✅ 简答题

解释什么是残差（residual），以及它们在模型拟合中的作用。

---

## 📊 Multiple Linear Regression

## ✅ 单选题

In multiple linear regression, multicollinearity refers to: A

- a) Having highly correlated independent variables
- b) A linear relationship between dependent and independent variables
- c) The presence of outliers
- d) A large residual variance

In multiple linear regression, which of the following indicates how well the model explains the variability in the dependent variable? D

- a) Adjusted R^2
- b) Residual standard error
- c) F-statistic
- d) All of the above

In multiple linear regression, what does the adjusted $R^2$ account for?
a) Adjusts for the number of predictors
b) Adjusts for outliers
c) Adjusts for nonlinearity
d) Adjusts for variance in residuals

2. A

In multiple linear regression, the **adjusted $R^2$** modifies the regular $R^2$ by penalizing the addition of unnecessary predictors. It accounts for the number of predictors to avoid overfitting — so it only increases if the new predictor improves the model more than would be expected by chance.

8 The coefficient of determination $R^2$ in simple linear regression can be interpreted as:
a) The correlation between residuals and predictors
b) The proportion of variance in $y$ explained by $x$
c) The intercept's contribution
d) The slope's value squared

2. B R^2 是一个衡量模型拟合优度的指标，表示**因变量 y 中有多少方差被自变量 x 所解释**。

2 Which of the following is NOT an assumption of multiple linear regression?
a) Linearity
b) Homoscedasticity
c) Independence of residuals
d) All independent variables are binary
✅ 答案：d)

5 What is the main purpose of adjusted $R^2$?
a) Make $R^2$ larger
b) Estimate the residuals
c) Penalizes unnecessary predictors
d) Provide residual variance
✅ 答案：c)

### ✅ 判断题

In multiple linear regression, adding more predictors always improves the adjusted R^2. F
(True / False)

In multiple linear regression, the p-value for a coefficient tests whether that predictor is significantly associated with the dependent variable, controlling for other variables.~~F~~ T
(True / False)

### ✅ 简答题

10 为什么 adjusted R^2 比普通 R^2 更能反映多元回归模型的拟合好坏？考虑了模型的复杂度，通过惩罚无关变量的加入，避免过拟合，更准确反映模型的拟合优度。

---

## ⚖️ Logistic Regression

### ✅ 单选题

In logistic regression, what is the range of predicted probabilities? B

- a) $[-\infty,\infty]$
- b) [0, 1]
- c) $[0,\infty]$
- d) $[-\infty,0]$

The log-odds (logit) function in logistic regression is: B A

a) $\log(\frac{p}{1-p})$

b) $\frac{1}{1+e^{-x}}$

c) $\log(p \cdot (1-p))$

d) $p \cdot (1-p)$

In logistic regression, what is the dependent variable?

a) Continuous

b) Binary (0 or 1)

c) Count data

d) Categorical with more than 2 levels

B

The output of the logistic regression model is:

a) A linear equation

b) A probability between 0 and 1

c) A count of occurrences

d) A mean value

B

When using logistic regression, the cost is very high if:
a) The predicted probability is close to the true label
b) The predicted probability is far from the true label
c) The residual sum of squares is small
d) The adjusted $R^2$ is large

B

If the correlation coefficient between two variables is 0, what does this imply?
a) The variables are independent
b) The variables have no linear relationship
c) The variables have no relationship at all
d) The variables are strongly related

B

In logistic regression, which method is typically used to estimate the model coefficients?
a) Least squares
b) Maximum likelihood estimation (MLE)
c) Bootstrapping
d) Gradient boosting

B；这一题是 estimate 所以是 MLE,如果问计算就是 Gradient

5 In logistic regression, the output of the logit function can be:
a) Any real number
b) Only between 0 and 1
c) Only positive numbers
d) Only integers
✅ 答案：a)
$\text{logit}(p) \in (-\infty, +\infty)$

## ✅ 判断题

Logistic regression is suitable for predicting continuous numerical outcomes.

(True / False)F

The logistic regression model uses maximum likelihood estimation (MLE) to find the best-fit parameters.F T
(True / False)

## 简答题

简述 logistic 回归的 Cost function，并解释当预测概率接近真实标签时，Cost 趋于什么数值。

# Week10

IG

Given is the following training data where ***location, weather*** and ***expensive*** are the features and ***holiday*** is the class.

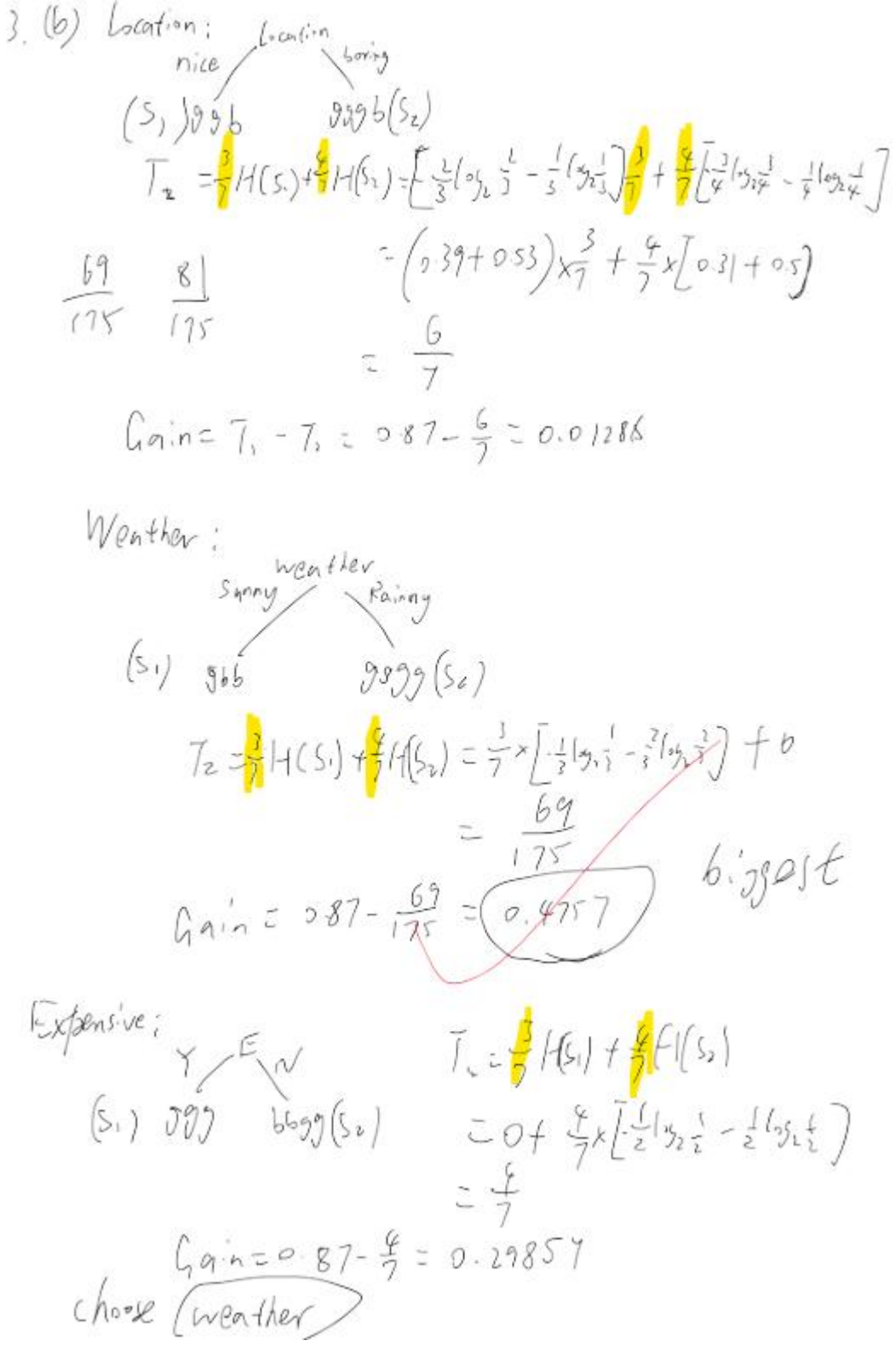
| location | weather | expensive | holiday | |
|---|---|---|---|---|
| nice | sunny | Y | good | + |
| nice | sunny | N | bad | − |
| boring | rainy | Y | good | + |
| boring | sunny | N | bad | − |
| nice | rainy | Y | good | + |
| boring | rainy | N | good | + |
| boring | rainy | N | good | + |

7 datas in total
2− 5+

a) What is the entropy of this set of training examples with respect to the class?
b) We would like to build a decision tree using information gain. Which attribute will be selected as a root of the tree? Show your calculations.

(a) $E = -\frac{2}{7}\log_2\frac{2}{7} - \frac{5}{7}\log_2\frac{5}{7}$

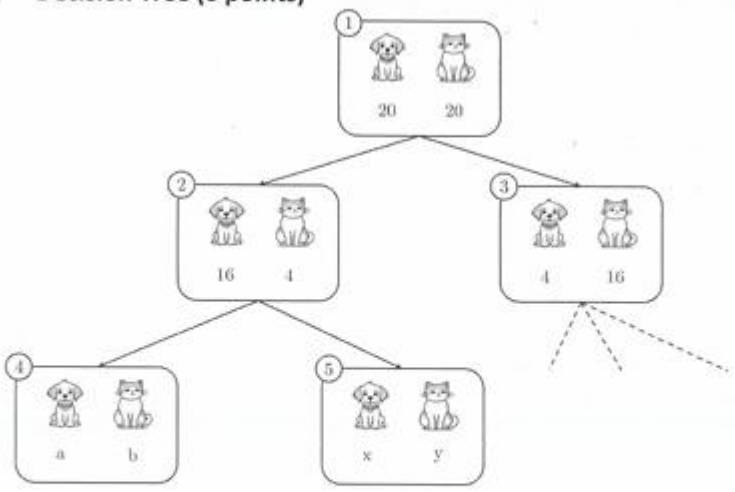
$\approx 0.52 + 0.35 = 0.87$




这里的 0.87 是由 a 得来的

More Exercise

**Question 7. Decision Tree (6 points)**




The figure above is an example of the decision tree for a classification problem. The numbers in each node represent the number of samples in each class (Dog/Cat) at that node. For example, in the original dataset (node number (1)), there are 20 Dogs and 20 Cats.

A split at node (1) gives us nodes (2) and (3) with a given number of samples in each class. In fact, the values in node (3) can be inferred based on the values in node (2) and vice versa. Suppose we have some split (D) at node (2), which gives us nodes (4) and (5), and the model decides that no split will be implemented at (4) and (5).

Entropy H(S) of node S is computed by:

$$H(S) = -\sum_i P_i . log_2 P_i$$

With $P_i$ be the proportion of class $i$ in node S

Information Gain of the split at the node S is computed by:

$$Gain(S) = H(S) - \frac{n_k}{n_S}\sum_k H(k)$$

a) Compute the information gain for the split at node (1) (2 points)

$H(S) = -\frac{20}{40}\log_2\frac{20}{40} - \frac{20}{40}\log_2\frac{20}{40} = 1$

$H(S_{left}) = -\frac{16}{20}\log_2\frac{16}{20} - \frac{4}{20}\log_2\frac{4}{20} = 0.723$

$H(S_{right}) = -\frac{16}{20}\log_2\frac{16}{20} - \frac{4}{20}\log_2\frac{4}{20} = 0.723$

$Gain = 1 - 0.723 \times \frac{1}{2} - 0.723 \times \frac{1}{2} = 0.277$

上面不是 0.723,而是 0.7219

b) Compute the information gain for the split (D) at node (2), with
a = 8, b =2
a = 16, b = 0
(2 points)

这里的 T1 是已知的，为 0.7219 (从 1 中得到)

Case1: a=8, b=2 -> x=8, y= 2；即左右 2 边的 entropy 相等

$$-\frac{8}{10}\log\frac{8}{10}-\frac{2}{10}\log\frac{2}{10}=0.7219$$

所以 case1 的 gain 是 0.7219-(10/20)*0.7219 - (10/20)*0.7219 =0

Case2: a=16,b=0-> x=0,y=4；即左右 2 边的 entropy 相等且都是 0

gain 是 0.7219-0=0.7219

c) From the values obtained in question b), derive your conclusion about the meaning of information gain. (2 point)

The information gain represents the goodness of split, if the split does not change the class distribution, the gain is zero. Otherwise, the gain is equal to the entropy of the node where the split occurs. It can also be used for attribute selection

# Week11

## Naïve without ZERO-FREQUENCE

Class:
C1: Interviewed well = 'False'
C2: Interviewed well = 'True'

Data to be classified:
X = (Level =Senior,
Lang = Python,
Tweets = yes
PhD = No)

14

| Level | Lang | Tweets | PhD | Interviewed well |
|---|---|---|---|---|
| Senior | Java | No | No | False |
| Senior | Java | No | Yes | False |
| Mid | Java | No | No | True |
| Junior | Python | No | No | True |
| Junior | R | Yes | No | True |
| Junior | R | Yes | Yes | False |
| Mid | R | Yes | Yes | True |
| Senior | Python | No | No | False |
| Senior | R | Yes | No | True |
| Junior | Python | Yes | No | True |
| Senior | Python | Yes | Yes | True |
| Mid | Python | No | Yes | True |
| Mid | Java | Yes | No | True |
| Junior | Python | No | Yes | False |

**X = (Level =Senior, Lang = Python, Tweets = yes, PhD = No)**
We need to compute $P(C_i \mid X) = P(X \mid C_i) * P(C_i)$

$P(C_{True})$: P(Interviewed well = "True") = 9/14 = 0.643

$P(C_{Fasle})$: P(Interviewed well = "False") = 5/14= 0.357

**Compute $P(X \mid C_{True})$**
P(Level = "Senior" | Interviewed well = "True") = 2/9 = 0.222
P(Lang = "Python" | Interviewed well = "True") = 4/9 = 0.444
P(Tweets = "Yes" | Interviewed well = "True") = 6/9 = 0.667
P(PhD = "No" | Interviewed well = "True") = 6/9 = 0.667

**Compute $P(X \mid C_{False})$**
P(Level = "Senior" | Interviewed well = "False") = 3/5 = 0.6
P(Lang = "Python" | Interviewed well = "False") = 2/5 = 0.4
P(Tweets = "Yes" | Interviewed well = "False") = 1/5 = 0.2
P(PhD = "No" | Interviewed well = "False") = 2/5 = 0.4

**$P(X \mid C_i)$ :**
P(X | Interviewed well = "True") = 0.222 x 0.444 x 0.667 x 0.667 = 0.044
P(X | Interviewed well = "False") = 0.6 x 0.4 x 0.2 x 0.4 = 0.019

**$P(C_i \mid X) = P(X \mid C_i) * P(C_i)$ :**
P(X | Interviewed well = "True") * P(Interviewed well = "True") = 0.044 * 0.643 = **0.028**
P(X | Interviewed well = "False") * P(Interviewed well = "False") = 0.019 * 0.357 = 0.007

Since $P(C_{true} \mid X) > P(C_{false} \mid X)$, therefore X belongs to class (Interviewed well = "True")

## Naïve with ZERO-FREQUENCE

| Text | Category |
|---|---|
| "A great game" | Sports |
| "the election was over" | Not sports |
| "Very clean match" | Sports |
| "A clean but forgettable game" | Sports |
| "It was a close election" | Not sports |

Our goal is to build a Naïve Bayes classifier that will tell us which category the sentence "A very close game" belongs to.

Using Laplace we have (因为在 sports class 中没有 close 这个 token)

$$p(\text{close} \mid \text{Sports}) = \frac{(0+1)}{(11+14)}$$

$$\{a, great, game, the, election, was, over, very, clean, match, but, forgettable, it, close\}$$

会发现所有的分母都加上了 14，即 distinct word 的数量

## Calculation

p(Sports | a very close game) =?
p(Not Sports | a very close game) =?

| w | $p(w \mid \text{Sports})$ | $p(w \mid \text{Not Sports})$ |
|---|---|---|
| a | $\frac{(2+1)}{(11+14)}$ | $\frac{(1+1)}{(9+14)}$ |
| very | $\frac{(1+1)}{(11+14)}$ | $\frac{(0+1)}{(9+14)}$ |
| close | $\frac{(0+1)}{(11+14)}$ | $\frac{(1+1)}{(9+14)}$ |
| game | $\frac{(2+1)}{(11+14)}$ | $\frac{(0+1)}{(9+14)}$ |

Let Z = A Very Close Game ; S = Sport ; NS = Non Sport

Therefore probability of "A Very Close Game" being a Sport => P(S|Z) = P(Z|S) P(S)

P(A|S) x P(V|S) x P(C|S) x P(G|S) x P(S) = 3/25 x 2/25 x 1/25 x 3/25 x 3/5 = 0.0004608 x 0.6 = 0.000027648

And the probability of "A Very Close Game" being a Non Sport => P(NS|Z) = P(Z|NS) P(NS)

P(A|NS) x P(V|NS) x P(C|NS) x P(G|NS) x P(NS) = 2/23 x 1/23 x 2/23 x 1/23 x 2/5 = 0.00001429 x 0.4=0.00000571

Then "a very close game" belongs to the Sports class

## 其它 Naïve example

| outlook | temp. | humidity | windy | play |
|---|---|---|---|---|
| sunny | hot | high | false | no |
| sunny | hot | high | true | no |
| overcast | hot | high | false | yes |
| rainy | mild | high | false | yes |
| rainy | cool | normal | false | yes |
| rainy | cool | normal | true | no |
| overcast | cool | normal | true | yes |
| sunny | mild | high | false | no |
| sunny | cool | normal | false | yes |
| rainy | mild | normal | false | yes |
| sunny | mild | normal | true | yes |
| overcast | mild | high | true | yes |
| overcast | hot | normal | false | yes |
| rainy | mild | high | true | no |

问题为

Use Naïve Bayes to predict the class (yes or no) of the new example:

| outlook | temp. | humidity | windy | play |
|---|---|---|---|---|
| sunny | cool | high | true | ? |

- $E$: 指 $E_1 = outlook = sunny, E_2 = temp = cool, E_3 = humidity = high, E_4 = windy = true$
- $H$: 指 play = yes 或者 play=no

Step 1: 明确公式

$$P(yes|E) = \frac{P(E|yes)P(yes)}{P(E)}$$

$$P(no|E) = \frac{P(E|no)P(no)}{P(E)}$$

Step 2：求$P(E|yes)$和$P(E|no)$

这里我们以$P(E|yes)$举例

因为$E$之间是independent的所以可以进行如下操作

$$P(yes|E)=\frac{P(E_1|yes)\,P(E_2|yes)\,P(E_3|yes)\,P(E_4|yes)\,P(yes)}{P(E)}$$

$$P(no|E)=\frac{P(E_1|no)\,P(E_2|no)\,P(E_3|no)\,P(E_4|no)\,P(no)}{P(E)}$$

这里我们拿$E_1$举例

P(E1|yes)=P(outlook=sunny|yes)=?/9=2/9

P(E2|yes)=P(temp=cool|yes)=3/9

P(E3|yes)=P(humidity=high|yes)=3/9

P(E4|yes)=P(windy=true|yes)=3/9

P(yes)=?

| outlook | temp. | humidity | windy | play |
|---|---|---|---|---|
| sunny | hot | high | false | no |
| sunny | hot | high | true | no |
| overcast | hot | high | false | yes |
| rainy | mild | high | false | yes |
| rainy | cool | normal | false | yes |
| rainy | cool | normal | true | no |
| overcast | cool | normal | true | yes |
| sunny | mild | high | false | no |
| sunny | cool | normal | false | yes |
| rainy | mild | normal | false | yes |
| sunny | mild | normal | true | yes |
| overcast | mild | high | true | yes |
| overcast | hot | normal | false | yes |
| rainy | mild | high | true | no |

proportions of days when outlook=sunny for days with play=yes
i.e. the probability of outlook=sunny, given that play=yes

Step 3：求$P(yes)$和$P(no)$

there are 9 rows which play = yes; and there are 14 rows in total. Hence

$P(yes)=9/14; P(no)=1-9/14=5/14$

Step 4：对比$P(yes|E)$和$P(no|E)$

$$P(yes|E)=\frac{\frac{2}{9}\frac{3}{9}\frac{3}{9}\frac{3}{9}\frac{9}{14}}{P(E)}=\frac{0.0053}{P(E)}$$

Similarly we can calculate the probability for play=no:

$$P(no|E)=\frac{\frac{3}{5}\frac{1}{5}\frac{4}{5}\frac{3}{5}\frac{5}{14}}{P(E)}=\frac{0.0206}{P(E)}$$

$P(yes|E)>P(no|E)$，所以 predicts "play=no" for the new example。

## 其它 Naïve exampleD

A national park has created a dataset to help hikers determine if a reptile they encounter could be venomous.

| | Head | Eyes | Size | Venomous |
|---|---|---|---|---|
| 1 | Triangle | Elliptical | Small | Yes |
| 2 | Round | Round | Small | No |
| 3 | Narrow | Elliptical | Small | No |
| 4 | Narrow | Round | Large | No |
| 5 | Narrow | Elliptical | Large | Yes |
| 6 | Triangle | Round | Small | Yes |
| 7 | Narrow | Round | Large | No |
| 8 | Round | Elliptical | Large | No |
| 9 | Triangle | Elliptical | Small | Yes |

Use Naïve Bayes to predict if the following example is venomous or not:
*Head=narrow, Eyes=elliptical, Size=Large*
Show the working for your calculations.

E1 is head=narrow, E2 is eyes=elliptical, E3 is size=Large

P(yes) = 4/9

P(E1|yes) = ¼

P(E2|yes) = ¾

P(E3|yes)=1/4

$$P(yes|E) = \frac{\frac{1}{4}*\frac{3}{4}*\frac{1}{4}*\frac{4}{9}}{P(E)} = 0.021 / P(E)$$

P(no)=5/9

P(E1|no) = 3/5

P(E2|no) = 2/5

P(E3|no)=3/5

$$P(no|E) = \frac{\frac{3}{5}*\frac{2}{5}*\frac{3}{5}*\frac{5}{9}}{P(E)} = 0.08 / P(E)$$

⇨ The prediction will be that the example is not venomous

https://zhuanlan.zhihu.com/p/26329951

# GPT 错题

6. In a star schema, what type of table is the fact table?

A. A table containing dimension attributes

B. A table storing numerical measures

C. A table with only categorical data

D. A table used only for normalization

> *Answer: B*

7. Briefly describe the basic steps of K-means clustering.

> *Randomly select K centroids, assign data points to nearest centroid, update centroids by calculating means of assigned points, repeat until convergence.*

20. Which measure is used to describe the spread in skewed data?

A. Mean

B. Mode

C. Median

D. Range

> *Answer: C*

8. How do you interpret a data set with a high standard deviation?

> *Data points are widely spread from the mean; high variability.*

14. What type of data is best displayed by a pie chart?

A. Categorical data showing parts of a whole

B. Continuous data distribution

C. Time series data

D. Correlation between variables

> *Answer: A*

8. Why is the median shown in a box plot instead of the mean?

> *Because median is resistant to outliers and better represents the central tendency in skewed data.*

9. Describe a situation where a pie chart is not appropriate.

> *When there are too many categories or the exact value comparison is important; a bar chart is better.*

9. What is the purpose of a Star Schema in a data warehouse?

A. To organize fact tables connected directly to dimension tables

B. To normalize all tables

C. To create complex many-to-many relationships

D. To use only one table for all data

> *Answer: A*

10. Which of the following is true about Snowflake Schema?

A. Dimension tables are normalized

B. Fact tables contain only keys

C. Dimension tables are denormalized

D. There are no foreign keys

> *Answer: A*

3. What happens if you compare NULL with any value using = operator?

A. Returns TRUE

B. Returns FALSE

C. Returns UNKNOWN (NULL)

D. Causes an error

> *Answer: C*

6. Which SQL function can calculate percentiles?

A. PERCENT_RANK()

B. PERCENTILE_CONT()

C. NTILE()

D. All of the above depending on the DBMS

> *Answer: D*

8. Why do you need to handle NULL values carefully in aggregation?

> *Because NULLs are ignored, which may skew the results if not accounted for.*

6. What is p-hacking?

A. Using inappropriate data analysis to get significant p-values.

B. Using cross-validation to avoid overfitting.

C. Correctly calculating p-values.

D. Running ANOVA.

> *Answer: A*

9. Which test would you use to compare means of two related groups?

A. Paired Student's t-test

B. Unpaired Student's t-test

C. Mann-Whitney U test

D. ANOVA

> *Answer: A*

16. What does precision measure?

A. Proportion of predicted positives that are true positives.

B. Proportion of actual positives correctly identified.

C. Total proportion of correct predictions.

D. The false negative rate.

> *Answer: A*

B 是 recall

9. Why is cross-validation used in machine learning?

> *To evaluate model performance and generalization by repeatedly training and testing on different subsets of data.*

10. What is the purpose of Ward's linkage in hierarchical clustering?

A. Minimizing the variance within clusters.

B. Maximizing cluster distances.

C. Using single point linkage.

D. Initializing centroids.

> *Answer: A*

这就是选择最小的 2 个进行合并

14. Which PCA step involves finding eigenvalues and eigenvectors?

A. Standardizing the data.

B. Computing the covariance matrix.

C. Projecting data onto components.

D. Clustering the data.

> *Answer: B*

15. What does SSE tend to do as the number of clusters increases in K-means?

A. Increase

B. Decrease

C. Remain constant

D. Randomly vary

> *Answer: B*

随着聚类数（k）增加:

✅ 每个簇会变得更小、点到质心的距离更短。

✅ 由于更多的簇可以更好地拟合数据，**SSE 会持续减小**。

但是，过多的簇（比如簇数等于样本数）虽然 SSE 很低，但并不一定代表合理的聚类。

18. The k-means algorithm tries to minimize:

A. Total intra-cluster variance (SSE).

B. Total inter-cluster variance.

C. Number of clusters.

D. The dimensionality of the data.

> *Answer: A*

19. Which distance measure is most sensitive to outliers in clustering?

A. Euclidean distance

B. Manhattan distance

C. Cosine similarity

D. Jaccard distance

> *Answer: A*

欧几里得距离（Euclidean distance）是基于平方差的距离度量。由于平

方操作会放大大数值的影响，因此它对离群点（outliers）特别敏感。

8. Describe the role of eigenvectors and eigenvalues in PCA.

> *Eigenvectors define principal component directions; eigenvalues measure variance captured by each component.*

10. Why might one choose hierarchical clustering over K-means?

> *Hierarchical clustering does not require predefining cluster number and reveals data hierarchy via dendrograms.*

19. The intercept α in linear regression is also called:

A. Bias term

B. Weight

C. Loss function

D. Gradient

> *Answer: A*

5. What are the basic steps of the gradient descent algorithm?

> *Initialize parameters, compute the gradient of the loss function, update parameters by subtracting learning rate times the gradient, and repeat until convergence.*

. What does Information Gain (IG) measure in attribute selection?

. The reduction in entropy after splitting by the attribute.

. The frequency of an attribute.

. The total number of unique values of the attribute.

. The length of the attribute's name.

> *Answer: A*

## 4. In TF-IDF, what does TF stand for?

A. Total Frequency

B. Term Frequency

C. Text Frequency

D. Token Frequency

> *Answer: B*

12. Which metric is often used along with Information Gain in attribute selection?

A. Gini Index

B. Mean Squared Error

C. Cosine Similarity

D. Euclidean Distance

> *Answer: A*

## 14. Naive Bayes is often used for which type of data?

A. Time series data

B. Text data

C. Image data

D. Unstructured data

> *Answer: B*

1. What is Information Gain, and why is it used in attribute selection?

> *Information Gain measures how much an attribute decreases the entropy (uncertainty) in the target variable, helping to select the best attribute to split data in decision trees.*

10. Describe one advantage and one limitation of using Naive Bayes for text classification.

> *Advantage: Fast and works well with high-dimensional data. Limitation: The independence assumption may not hold, potentially reducing accuracy.*