考试形式

–考试将是6-8个问题

- 多项选择

–简短答案和冗长答案

–数字

–它将从讲座和练习中汲取

–如何复习?

–查看幻灯片和笔记本解决方案

–没有Python

–有SQL（尤其是SQL的solutions）

week一、Kinds of Data and Exploratory Analysis 数据种类和探索性分析

1，数据种类（Data (File) Formats）（week3）

–数据（文件）格式

–结构化，半结构化与非结构化数据

Structured vs. Semi-Structured vs. unstructured data

–数据种类（ Kinds of Data）

–标称数据 Nominal vs序数数据ordinal   vs间隔数据Interval  vs比率数据 ratio

Spearman’s rho for ratio data, ordinal data, etc (rank-order correlation)

Spearman的rho，用于比率数据，序数数据等（排名相关）

Python pprint模块

–复杂数据结构的漂亮打印

– pprint格式化通过CSV读取的词典，因此更易于阅读

2、Descriptive Statistics 描述性统计

–目标：用于量化价值分布的简单指标

–集中趋势的度量（ central tendency: ）：Measures of central tendency:

        –均值，中位数，众数 mean, median, mode

–分散（或扩散）措施（ dispersion (or spread)）：Measures of dispersion (or spread):

       –计数，最小，最大，范围，百分位数，方差，标准差，IQR，百分位数

count, min, max, range, percentiles, variance, stddeviation, IQR, percentiles

–注意：并非所有描述性统计信息都对每种数据均有效

       –例如 最小值/最大值或百分位数未在标称值上定义，但仅至少在序数数据上定义

min/max or percentiles not defined on nominal, but only at least ordinal data

–重要的是要知道每个统计的含义以及如何解释

– Important to know what each statistic means and how to interpret

3、Data Visualisation ·数据可视化

–目标：直观地汇总数据 Goal: summarise data visually

  –传达价值分配和关系的重要方法

important approach to convey value distributions and relationships

–取决于数据种类 Depending on kinds of data

      –汇总名义数据： Summarising Nominal Data 直方图，条形图

      –汇总序数数据： Summarising Ordinal Data 条形图，箱形图，直方图

      –汇总间隔数据： Summarising Interval Data: 散点图，合并的直方图，频率多边形

–我们介绍了如何在电子表格和Python中完成此操作

       –确切的命令是什么，比知道什么时候可视化最有效以及如何做得更好（例如，清晰的标签，有意义的轴范围等）以及如何解释它们更重要。

4、推荐复习方法（见week3的ppt和 solution）

–查看并理解幻灯片和练习

     -（1）描述性统计

使用collections.Counter的频率分布

–收集模块提供了几种有用的数据结构

–柜台

      –将列表或其他可迭代项作为输入并计算其条目

      –返回每个项目出现频率的计数

用collections.Counter计算模式

–记得模式是最频繁的值

–我们可以使用most\_common（）方法进行计算

清洁数据：转换为正确的类型

– Python csv模块将所有内容读取为字符串类型

–需要进行适当的转换（例如int，float，timestamp）

– int（）创建整数对象，例如-1、101

– float（）创建浮点对象，例如3.14、2.71

– datetime.strptime（）从字符串创建日期时间对象

     （2）描述性数据可视化，例如直方图，箱形图，散点图，

Week二、Clustering 聚类

1、data Clustering 数据聚类

–目标：将对象分组为相似数据的簇

–注意：没有确切答案

聚类类型：

(1)分区集群（Partitional clustering）（K-means选中心点）

      将数据对象划分为不重叠的子集（集群），以便每个数据对象都恰好在一个子集中

(2)分层聚类（Hierarchical Agglomerative Clustering）

一组嵌套的聚类，组织成一个层次树

2、层次聚集聚类 Hierarchical Agglomerative Clustering

层次集群的策略通常分为两种：

a\集结：这是一种“自下而上”的方法：每个对象都在其自己的群集中开始，并且随着一个对象在层次结构中的上移，合并成对的群集。

b\分裂：这是一种“自上而下”的方法：所有对象都在一个群集中开始，并且随着一个对象在层次结构中向下移动，递归执行拆分。

–初始

      –自己集群中的每个点

–重复

      –查找最接近的集群对

              •任何两点之间的最小距离（单联动）

      –将它们合并到一个集群中

      –重新计算新群集与其他群集之间的距离

- 直到

      –所需的剩余群集数，例如 单簇

3、分区聚类：K均值聚类方法 Partitional clustering: K-Means Clustering Method

–给定k，k-means算法的实现分为四个步骤：

      （1）将对象划分为k个非空子集

      （2）将种子点计算为当前分区的群集的质心（质心是群集的中心，即mean point）

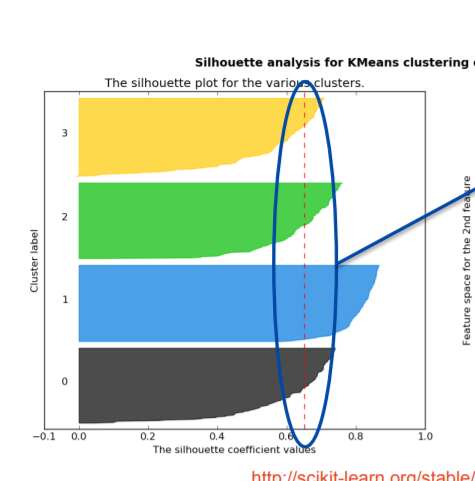
      （3）将每个对象分配给具有最近种子点的群集

      （4）返回步骤2，当分配不变时停止

4、使用Silhouettes选择k Using Silhouettes to choose k

High average silhouette indicates points far away from neighboring clusters

高平均轮廓表示点距邻近群集较远



SEE距离计算：SSE Example

– 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

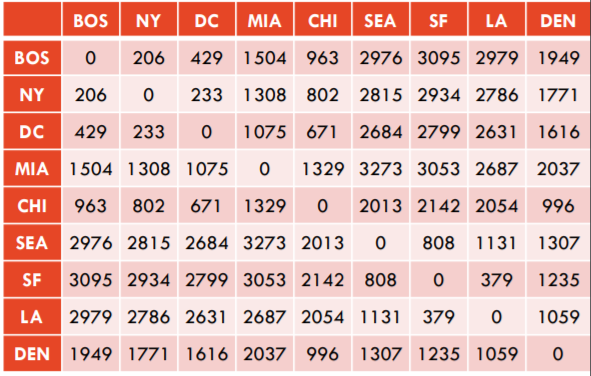
5、练习：（答案在12周）

以下页面跟踪美国城市之间距离的分层聚类（以英里为单位）。

•群集的方法是单链接。

•使用聚集层次聚类（第2）算法对以下数据进行聚类。

•使用报告中的计算表说明合并过程的每个步骤。

•绘制最终树状图

分层算法

两个层次聚类之间的距离计算：Distance Calculation between two hierarchical clusters：

–单联动： – single linkage:

        –每个群集元素之间的最小距离  – The minimum distance between elements of each cluster

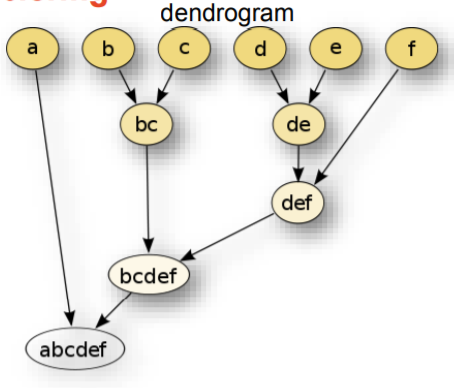
–完整的链接：– complete linkage:

        –每个群集元素之间的最大距离The maximum distance between elements of each cluster

–平均链接：即平均距离计算。– average linkage: i.e. mean distance calculation.

若BC间距最小，则将BC放到一起，如果按照single linkage算，则取A到B和C中最短的距离当做A到BC的距离，如果complete linkage:则取最大， average linkage则取平均数。

要求：在每个步骤中绘制表格，然后绘制最终树状图。



week三、关联规则 Association Rule（week 7）

1、关联规则挖掘

–根据交易中的其他项目预测项目的发生，例如：

{尿布}---{啤酒}

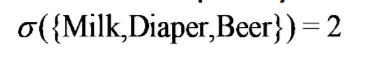
{牛奶，面包}---{鸡蛋，可乐}

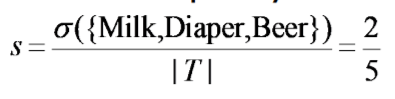
{啤酒，面包}--{牛奶}

–请注意，箭头表示同时出现，而不是因果关系（买尿布同时又买了啤酒）

2、Definition: Frequent Itemses 定义：频繁项集



–支持计数（σ）是项目集频率 

–支持（s）是标准化的项目集频率 

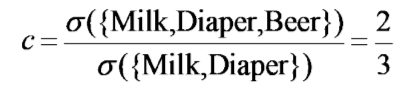
–频繁项集有  s ≥ min\_support

3、定义：关联规则 Definition: Association Rule

–关联规则是X --- Y形式的含义，其中X和Y是项目集

{牛奶，尿布}----{啤酒}

––Confidence (c) measures how often Y occurs in transactions with X

 –置信度（c）衡量与X的交易中Y发生的频率

4、Mining Association Rules

–关联分析 Association Analysis

      –频繁项集     – frequent itemsets

      –关联规则表示频繁项目集之间的含义：X --- Y

       –措施：频率，支持，信心 measures: frequency, support, confidence

–采矿协会规则 Mining Association Rules

     –蛮力枚举在计算上是禁止的，因此：

     – FP-growth或Apriori算法，用于生成频繁项集（两种算法都要会）

5、练习

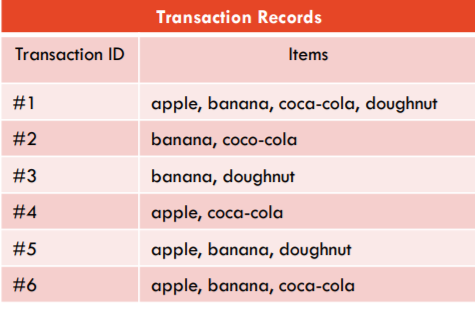
1.使用最小支持min\_sup = 2构建FP树。显示树如何为每个事务演化。

2.使用FP-Growth算法从FP-tree中发现频繁的项目集。

3.对于先前的交易记录，在此数据集上使用Apriori算法，并验证它会生成min\_sup = 2的同一组频繁项集。

4.假设{苹果，香蕉，甜甜圈 Doughnut }是一个常用商品集，请使用min\_confidence = 70％导出其所有关联规则

Suppose that { Apple, Banana, Doughnut } is a frequent item set, derive all its association rules with min\_confidence = 70%



Apriori的性能瓶颈(缺点，disadvantage，不足)

– Apriori的瓶颈：

      –候选人产生：

           •生成庞大的候选集

      –多次扫描数据库

6、推荐复习方式

–查看并理解幻灯片和练习

      –使用k均值的数据聚类

      –分层聚集聚类

      –评估聚类，如何选择k

      –关联规则挖掘：FP增长和Apriori（week7 和8 clustring 和 Association rule）

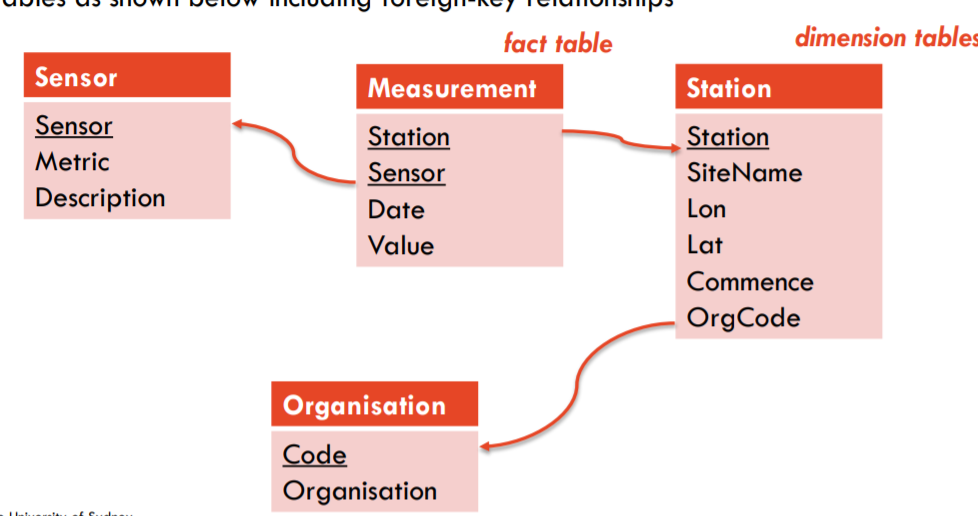
week四、数据处理与管理Data Processing and Management

1、水数据库架构 Water Database Schema（week 4 和 5）

如下所示的四个表包括外键关系

Four tables as shown below including foreign-key relationships

（考试中会给出这样的图表，各个表格间的关系，然后要求写出a query，你只需要table的名称和fears，这就是全部你需要的去写SQL语句，所以要去练习）



2、SELECT语句选项（week 5）（例如找出top5/ 最高的）

SELECT COUNT(\*) FROM T:：计算表T中存储了多少个元组

SELECT \* FROM T：列出表T的内容

SELECT \* FROM T LIMIT n：仅列出表中的n个元组

SELECT \* FROM T ORDER BY a：对结果按属性a进行排序（升序；按降序添加DESC）

（例如找出 最高的）： SELECT Mark From Student（table） order by mark limit 1

3、SQL聚合函数（第五周）

COUNT（attr）; COUNT（\*）：非空位数； 或所有值

MIN（attr）：attr的最小值

MAX（attr）：attr的最大值

AVG（attr）：attr的平均值（算术平均值）

MODE（）WITHIN GROUP（ORDER BY attr）：attr上的模式函数

PERCENTILE\_DISC（0.5）WITHIN GROUP（ORDER BY attr）：attr值的中位数

4、具有GROUP BY和HAVING的查询

在SQL中，我们可以根据一个或多个属性的值将关系“划分”为组：

SELECT [DISTINCT]：]目标列表target-list

FROM  ：    relation-list  关系表

WHERE      qualification   资格

GROUP BY      grouping-list  分组清单

HAVING group-qualification

– A group是一组元组，它们对分组列表中的所有属性具有相同的值。

–注意：聚合函数之外的select子句中的属性必须出现在分组列表中

–直观地，每个答案元组都对应一个组，并且这些属性每个组必须具有单个值。

5、推荐学习

–查看并理解幻灯片和练习

      –使用SQL查询和汇总数据

      – SQL连接运算符

      –使用having子句过滤组

      – DISTINCT（unique vaule  去除重复的）(例如有10000个人，但是只有10门课，我想知道是哪十门，就需要用上DISTINcT)

SELECT DISTINCT  A  FROM B

week五、Statistical Inference统计推断

1、研究问题

–研究问题（问）：

      –询问自变量是否有影响

      –“如果自变量有变化，因变量也有变化吗？”

–空假设（H0）：

      –没有效果的假设

      –“自变量更改时，因变量没有更改。”

假设检验

–根据数据样本提供的证据，我们用它来指定是接受还是拒绝有关人口的索赔。

2、使用p值测试可靠性

–多数测试会计算观察端的p值

–比较显着性水平阈值α

       –α是给定H0时拒绝H0的概率

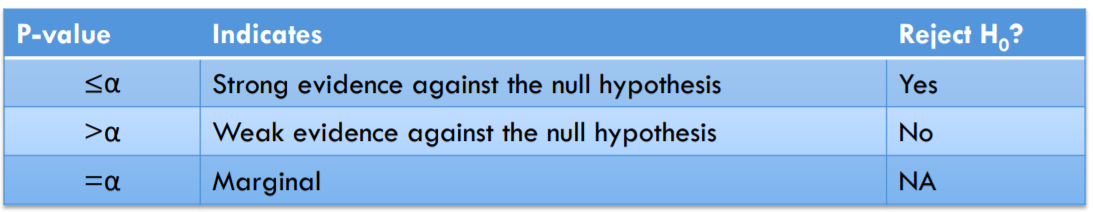
       –通常使用5％或1％的α

– Most tests calculate a p-value measuring observation extremity

– Compare to significance level threshold α

– α is the probability of rejecting H0 given that it is true

– Commonly use α of 5% or 1%



3、Measurement: Accuracy, precision, recall, f1 测量：精度，精度，召回率，f1\*

（week6）

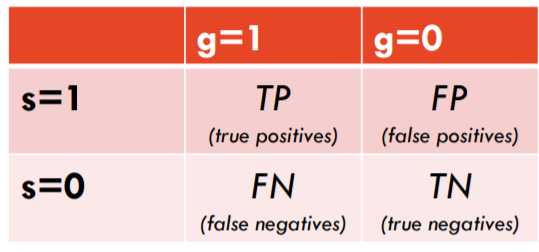
–精度Accuracy：（TP + TN）/ N 在所有实例中正确率％

–精度 Precision：TP /（TP + FP） 正确的系统预测百分比

–召回 Recall：TP /（TP + FN） ％correct gold labels

– F1：2PR /（P + R） 精确度和召回率的谐波平均值

TP:spam被检测为spam，TN：ham被认为是ham，FN：Spam被认为是ham，FP：ham被认为是Spam



4、推荐复习方法

–查看并理解幻灯片和练习

      –示例场景

      - 研究问题

      –空/替代假设

      –统计测试Statistical tests

– Between subjects:

Each subject sees  one and only one  condition

– Within subjects:

Subjects see more  than one or all  conditions

–主题之间：

每个对象只能看到一种情况

–主题内：

受试者看到一个或多个条件

Dependent variable is  the measure of interest

– Independent variable is  manipulated to observe the effect on dependent variable

– Controlled variables are  materials, measurements and= methods that don’t change

因变量是关注度

–对自变量进行操作以观察对因变量的影响

–控制变量是不变的材料，度量和=方法

Increase the power of a significance test

– Obtain a larger sample

– Larger N means more reliable statistics

– Less likely to have errors

– Type I: Reject true H0 – Type II: Fail to reject false H0

增加重要性检验的能力

–获取更大的样本

– N越大表示统计数据越可靠

–不太可能出现错误

–类型I：拒绝为真H0 –类型II：未能拒绝为假H0

week六、Machine Learning机器学习

1、Goal is to build models that generalise to unseen data

目标是建立可概括未见数据的模型（1、Underfitting 2、fit 3、Overfitting）

2、（要知道如何防止）通过正则化防止过度拟合 Prevent overfitting with regularisation

–较大的特征空间会带来过度拟合的问题

–正则化增加了随着系数（β）变大而变大的损失

–我们对误差项给予的权重越大，就越不鼓励大系数

（1、增加正则项 2、增加 trainning set、3、降低模型复杂度）

（讲义：）通过正则化防止过度拟合 Prevent overfitting with regularization

一种自动控制学习模型复杂度的方法

–正则化旨在惩罚较大的系数值（𝜃𝜃𝑗𝑗）

      –可以纳入成本函数

      –我们对误差项给予的权重越大，就越不鼓励大系数

–还可以通过消除功能（手动或通过模型选择）解决过拟合问题

      –较大的特征空间会带来过度拟合的问题

A method for automatically controlling the complexity of the learned model

– Regularization aims to penalize for large values of coefficients (𝜃𝜃𝑗𝑗)

     – Can incorporate into the cost function

     – The more weight we give to the error term, the more we discourage large coefficients

– Can also address overfitting by eliminating features (either manually or via model selection)

     – Large feature spaces introduce problems with overfitting

启示：Mann-Whitney U检验 Significance: Mann-Whitney U test

–未配对t检验的非参数版本 – Nonparametric version of unpaired t-test

–假设 – Assumes

–样本是独立的 – The samples are independent

- 注意 – Note

– N至少应为20 – N should be at least 20

意义：Kruskall-Wallis H检验

–方差分析的非参数版本

–测试不假定您的数据来自特定分布，例如正态分布。

–假设样本是独立的

–有时称为等级单向方差分析

–因为测试中使用的是数据值的排名，而不是实际数据

- 注意：

–不推荐用于小于5的样品

–在统计上不如方差分析

–方差分析和Kruskall-Wallis H检验都是Mann-Whitney检验的扩展

和未配对学生的t检验用于比较两个以上的均值

人口。

七、Information Gain 信息增益（找怎么算H(X)和H(YIX)，要知道怎么计算）

1、信息增益（IG）

– IG根据属性的值做出决策后，计算熵的有效变化。

IG calculates effective change in entropy after making a decision based on the value of an attribute.

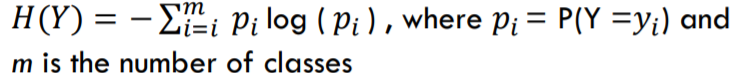
IG(Y|X) = H(Y) – H(Y|X)

2、信息增益（IG）

– IG根据属性的值做出决策后，计算熵的有效变化。

–熵：

–测量与数据相关的不确定性



–解释：

•更高的熵=>更高的不确定性

•较低的熵=>较低的不确定性

–条件熵

𝐻(𝑌|𝑋) = ∑𝑥 𝑝𝑝(𝑥) 𝐻(𝑌|𝑋 = 𝑥)

3、Exercise 练习

– What is the information gain of X relative to these training examples? given that Y is the label

与这些训练示例相比，X的信息收益是什么？ 鉴于Y是标签



八、决策树 Decision trees

1、培训数据：受访者数据

–四个功能：

      •Level，Lang，Tweets，PhD

–类标签：

      •面试得很好 Interviewed well

–我有新申请人A15（Level，Lang，Tweets，PhD）

–想要预测被采访者是对还是错

–很难猜到A15！



2、预测A15是否属于True或False（首先会问你选择那个属性，不过考试会考诉你例如Level有最高的IG，那么就要知道首先选Level进行分类，level分类后，通过肉眼数数，决定接下来怎么分，怎么分才能纯净，可以每个子集选不同的分类，比如A选PhD，C选Tweets）

–分而治之 Divide-and-conquer

–选择属性以将数据拆分为子集

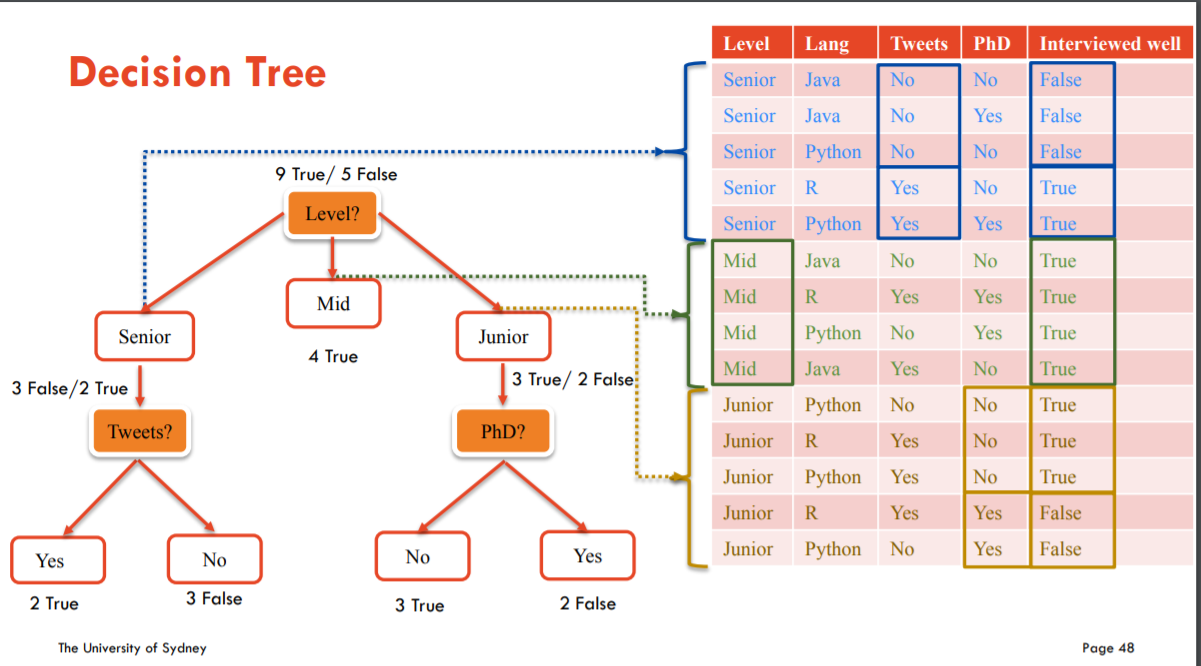
–它们是纯净的吗？（全部为真或全部为假）

–如果是：停止

–否：重复

–选择哪些属性？

–让我们首先选择“级别”属性



max\_depth

      –树的最大深度

–标准

      –熵：选择最小化总不确定性的分割

      –基尼（Gini）：选择尽量减少误分类的拆分

–分离器

      –最佳：为每个功能选择最佳阈值

      –随机：为每个功能选择最佳随机阈值

max\_depth

     – the maximum depth of the tree

– criterion

     – entropy: choose splits that minimise total uncertainty

     – gini: choose splits that minimise misclassification

– splitter

     – best: choose the optimal threshold for each feature

     – random: choose the best random threshold for each feature

9、朴素贝叶斯分类器

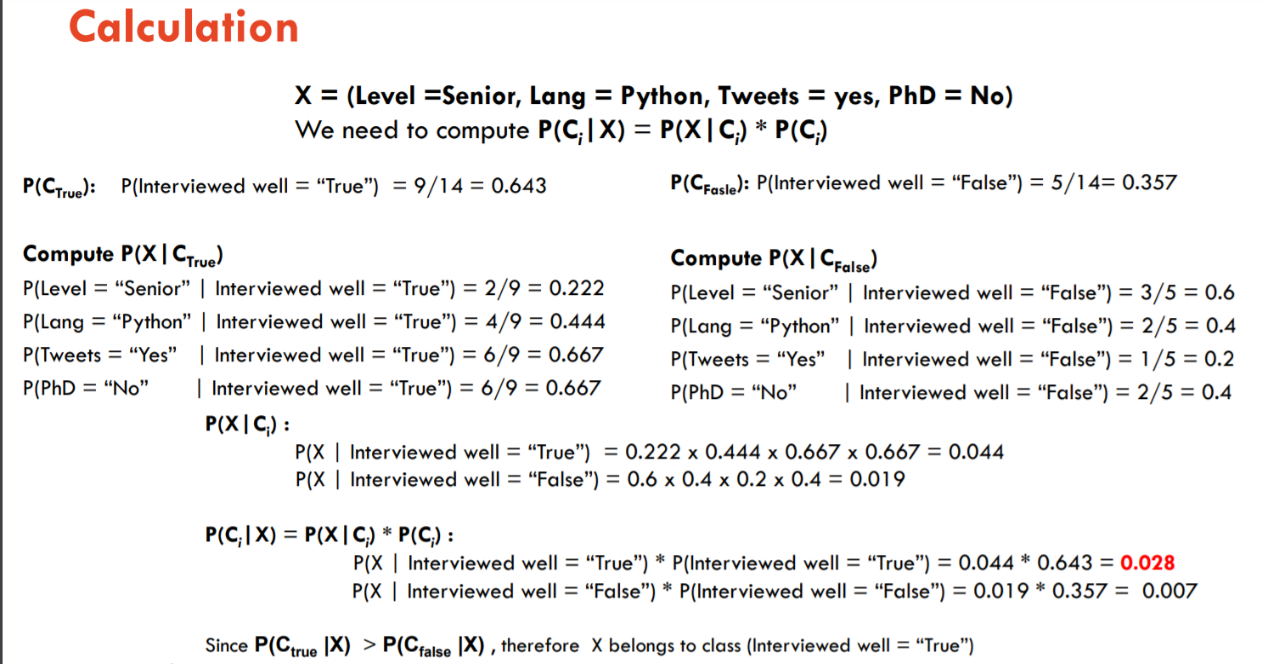
1、一个例子

类：

C1：接受采访=“假”

C2：接受良好采访=“正确”

要分类的数据：X =（级别=高级，郎= Python，推文=是 PhD =否）



1.5\𝒑 𝒄𝒍𝒐𝒔𝒆 𝑺𝒑𝒐𝒓𝒕𝒔 = 𝟎, leading to 𝒑 𝒂 𝒗𝒆𝒓𝒚 𝒄𝒍𝒐𝒔𝒆 𝒈𝒂𝒎𝒆 𝑺𝒑𝒐𝒓𝒕𝒔 = /span>（当其中一个项为零时，需要处理，详情见ppt，week11）

2、推荐复习方式

–查看并理解幻灯片和练习

      –单次和多次线性回归（了解）

      –决策树（怎么获得树，怎么算IG，怎么建立模型，怎么data）

     –朴素贝叶斯分类器 Naïve Bayes classifier

      –信息获取 Information gain

      –评估和设置   – Evaluation and setup

补充：：

Clustering聚类 层次聚集聚类 Hierarchical Agglomerative Clustering（2）（计算）

要求：在每个步骤中绘制表格，然后绘制最终树状图。（比sample 里面的步骤多了一步）

week4 3/4 tables, SQL 的问题

week5 周的sql语句，练习，就够了，应对sql问题

（tutor 的solution）/

15个多选题

5个简答(给两个plot，让解释这个图)

3个问答题（大题，需要图片，可以在纸上写，拍下来）

内容

kinds of data（week10）

how we can masure the data (diff types)

一般是多选题，想要视觉化数据，选择点状图，还是柱状图

K-Means的步骤（4步）

Hierarchical Agglomerative Clustering知道从下到上， 从下到上

Definition: Frequent Itemset两种算法都要知道，会问你用哪种

最小支持会给出

week6 最小 f1，p-vaule，H0假设，

logistic regrssion可能会有多选题

Are all the knowledge points included in the final exam in today's lecture? 90%