Week1:

Data Mining is the process of discovering patterns, understanding trends, and extracting useful information from large datasets

Data Visualization is the process of designing visual representations of data.

Data Types：

Categorical分类Data：

data that can be divided into groups（Gender, race, age group, and educational level）

Quantitative定量Data：

data that can be represented numerically, including anything that can be counted, measured, or given a numerical value（Age, height, number of students in a lecture）

Discrete离散Data：

data that can only assume specific values that cannot be subdivided再分

Continuous Data：

any numeric value and can be meaningfully split into smaller parts

分类数据必是离散的，定量数据可离散可连续

Levels of Measurement：

数据根据精度分为四个等级：Nominal→Ordinal→Interval→Ratio

Nominal：

Unordered classes无序类，数据只能被归类categorized

可被编码为数字'dummy variables哑变量'

如：性别，种族，学位项目

Ordinal：

Ordered classes有序类，可以被归类并排名

也可以被编码为numeric variables数值变量

如：年龄组，教育程度

Interval：

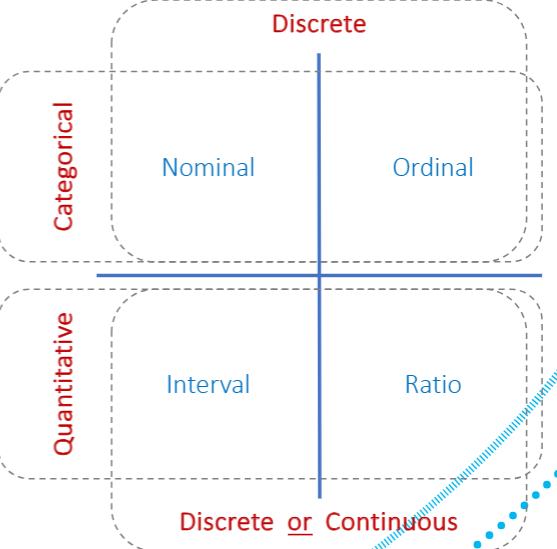
Numerical (quantitative)数值数据，有相等间隔但是没有绝对0

如：温度，年

Ratio：

有相等间隔还有绝对0的Numerical (quantitative)data

如：年龄，体重身高



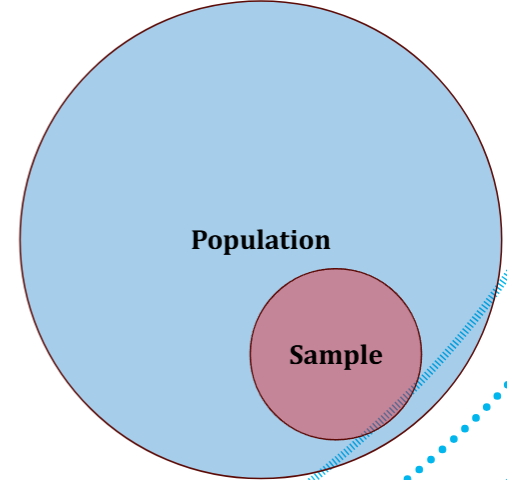
Exploratory Data Analysis (EDA)：

process of exploring your data，通常以无计划unplanned and 调查investigative 的方式进行 帮助深入了解数据

Descriptive Statistics：

Population Vs Sample：

一个给定的数据集可以代表某个总体，或者该总体的一个样本



Population：

an entire group of people/objects/items etc. of interest

一般很难获得population数据

Sample：

感兴趣研究的总体的一个子集subset

可以用来估计总体，但是有误差

Descriptive Statistics描述性统计：

provide a way of summarizing our data

让我们能够了解数据集的性质properties和特征characteristics，以及数据的“外观”

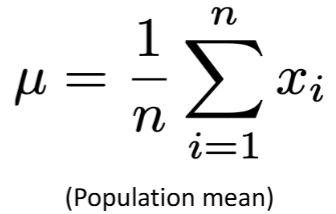
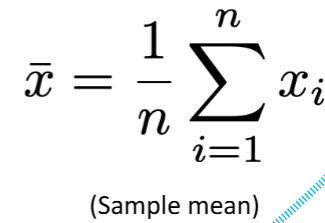
通常我们关注数据的**central tendency中心趋势**和**variability变异性**

Measures of Central Tendency：

The central tendency of a set of data试图通过其中心位置来描述数据

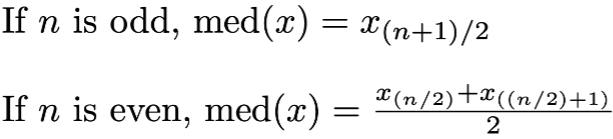
目标是使用一个值来准确地描述整个数据集，最常见的如平均值

Arithmetic Mean算数平均数：

 总体的和样本的不一样

提供了一个有效的中心趋势度量，但是很容易受到偏斜skewed数据和异常值outliers的影响，也可能无法准确反映数据

Median中位数：

将一组值上半部分和下半部分分开的值

在处理有序数据的时候非常有用

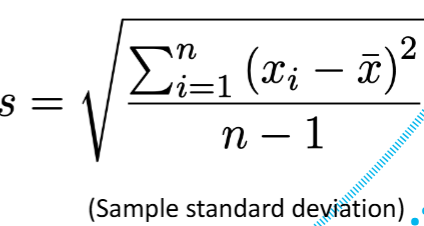
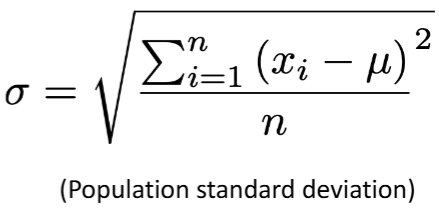
Mode众数：

出现最多的数

Measures of Variability：

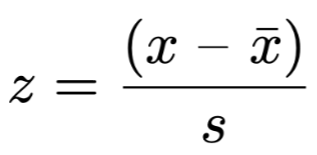
Standard Deviation标准差：

数据与平均值之间的平均分散程度



Z-Score标准分数：

给定的数据点与平均值相差的标准差数量



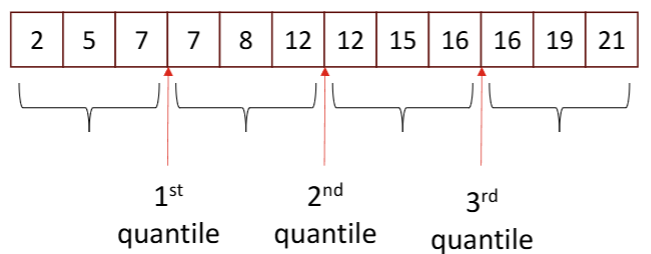
0：该值与平均值相同

1：该值比平均值高1个标准差

-2：该值比平均值低2个标准差

Quantiles分位数：

将样本分成q个相等组的一种方式，其中第k个分位数表示在k/q的数据以下的值

分四组所以q=4

分位数名称：percentiles (100 groups), deciles (10 groups),  
quintiles (5 groups), and quartiles (4 groups)

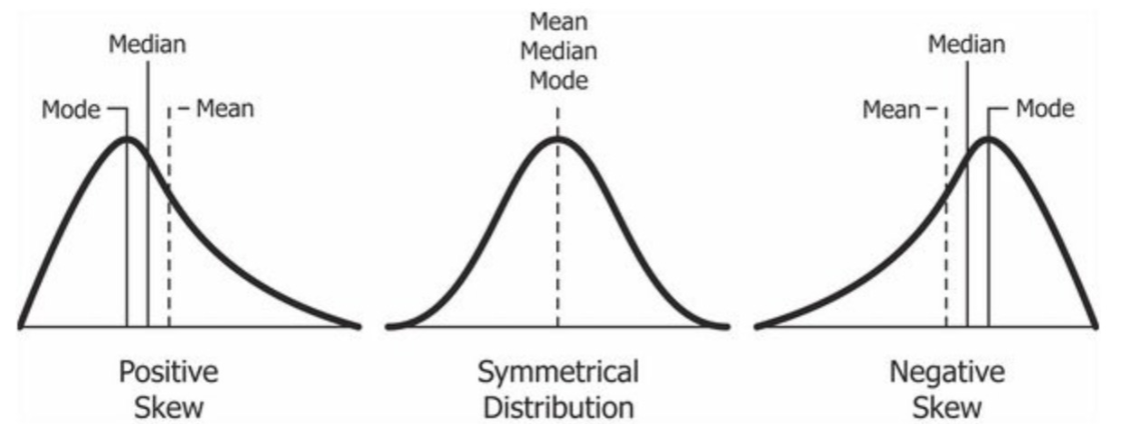
中位数与提到第50百分位数、第5十分位数和第2四分位数是相同的

Interquartile Range (IQR)四分位数：

使用四分位数来衡量的变异性度量，它测量第一四分位数（Q1）和第三四分位数（Q3）之间的差异——即 IQR = Q3 - Q1

Skew偏度 & Kurtosis峰度：

Skewness：衡量分布非对称性的一个指标



Kurtosis：衡量分布“尾部特征tailedness”的一个指标

Cleaning Data：

Missing Data：

可以复制相似行中的值，例如使用其“最近邻”，或者使用其他方法来查找相似行

可以复制平均数，中位数等

Outliers异常值：

与其它观察结果显著不同的数据点

可以选择删除该行数据

向数据中插补其他值（例如，均值上方2个标准差）

但是都是有风险的

Data Pre-Processing：

Dummy Coding 哑变量编码：

将分类变量转换为表示相同数据的一组数值（布尔值）

在处理名义nominal变量或有序ordinal变量时特别有用

如：描述高度时不直接表示高度，而是用tall，medium，short 1/0表示

Bucketing分箱：

有时我们可能希望将数值（区间或比例）数据转换为有序类别

当我们要处理共享某些相似数值属性的组时，这很有用

将数值数据放入“箱”或“桶”中

如：[15, 17, 24, 28, 33, 45]年龄，可以转换为[11-19, 20-29, 30-39, 40+]

[青少年, 青少年, 成年人, 成年人, 成年人, 成年人]

Class Imbalance：

想预测数据集中很少有示例的罕见事件

Data Downsampling下采样：

对多数类进行采样，使其数量等于少数类的数量

好处是类别平衡。但缺点是我们现在拥有的数据少得多

Data Upsampling上采样：

复制少数类的实例，使其数量等于多数类的数量

获得了更多的数据，并且我们没有丢弃任何数据

然而可能导致数据过拟合overfitting

Split–Apply–Combine：

[Split]: 根据某些标准将数据集[分割]成组

[Apply]: 对每个组[应用]一些函数

[Combine]: 将结果[组合]成一个新转换的数据集

箭头

AI 生成的内容可能不正确。

Relationships in Data:

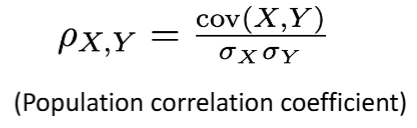
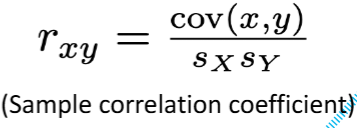
当人们谈论相关系数时，他们通常指的是Pearson Correlation Coefficient (PCC)

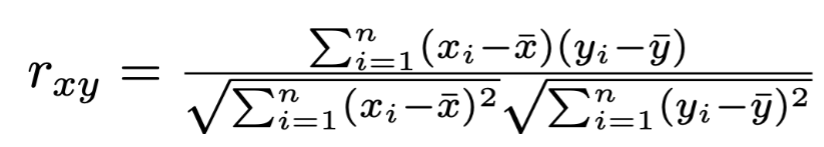
介于 -1 和 1 之间表示关联的强度,可以强或者若，正或者负

0表示不相关

1表示最强相关

-1表示最反相关

 总体和样本的相关系数计算方式



Limitations of Correlation：

*“correlation does not imply causation“*

Correlation是衡量两个变量之间关联强度的指标(x变时y变多少)

Causation是一个事件导致或产生另一个事件的过程(x变会不会导致y变)

Direct correlation：两个或多个变量直接相连

Indirect correlation：两个变量不直接相连，但有一个或多个变量将它们直接连接

Data Visualization：

数据可视化（DataVis）是设计易于理解的视觉表现的过程

Types：

Bar Charts：允许我们探索分类categorical变量和定量quantitative变量间的关系

Histogram直方图：允许我们查看分类变量或定量变量的分布

Boxplots箱线图：用于展示定量变量在不同分类变量上的分布

Scatter Plot散点图：用于探索两个定量变量之间是否存在关系

Line Graph折线图：用于展示两个定量变量之间的关系

The Grammar of Graphics：

提供了一种结构化和模块化modular的方法，用于设计和理解可视化的核心组件

Components：

Data, Aesthetics美学属性, Geometries, Scales比例尺, Facets面元, Coordinate坐标系 systems, Statistical transformations

Data：起点始终是确定我们想要可视化的原始数据（或数据集）

Aesthetics：x轴y轴，颜色，字体

Geometries：数据在图形中呈现的形状或形式。它决定了创建的图形类型

Scales：控制了从数据值到可视化美学属性的映射

Facets：可用于根据一个或多个分类变量的值生成多个子图

Coordinates：控制数据点在图形上的布局

Transformation：总结数据来转换数据(可能包括平滑数据smoothing the data、将数据分成几个桶buckets、添加少量随机噪声random noise等)

A/B Testing：

or ‘split testing’，一种广泛用于进行受控随机实验controlled, randomised experiments的方法（比较两个或多个版本的某样东西（例如界面/设计/产品等）确定一个版本是否比另一个版本更好）但是有幸存者偏差，如何避免？

Null Hypothesis Statistical Testing：

零假设统计检验（NHST）是指一组用于确定数据是否支持特定假设的统计方法

null hypothesis(H0零假设)指正在研究的效果不存在

alternative hypothesis(H1;Ha备择假设)指你想要测试的内容

idea:

使用统计测试计算一个“p 值”，其中 0 < p < 1

如果这个 p 值低于某个阈值（通常是 0.05），我们可能会说结果是“具有统计学意义statistically significant’”的

P值：零假设为真的情况下，观察到至少与实际结果一样极端的测试结果的概率

Effect Size：

两个或多个变量或组之间关系强度

Sample Size：

NHST 的结果还将取决于样本量

随着样本量增加，总体更加准确的反映

图片包含 图示

AI 生成的内容可能不正确。

P值固定的情况下：较强的effect size将需要较小的Sample Size

较高的Sample Size将能够检测到较弱的effect

sample size固定的情况下：较强的effect size将导致较低的 p 值

较弱的effect size将导致较高的 p 值  
effect size固定的情况下：较大的Sample Size会降低 p 值

较低的Sample Size会导致较高的 p 值

当 p < 0.05 是我们固定的截止点时，我们可以从最小的效应量中找到具有统计学意义的结果——只要我们有足够的数据

Independent T-test：

公式：手机屏幕的截图

AI 生成的内容可能不正确。分子：样本均值的差 分母：样本标准误（即方差的度量）

Paired T-test：

如果两个组不是独立的话

ANOVA(Analysis of Variance)方差分析：

允许我们比较三个或更多的组

Family-Wise Error Rate(FWER家族错误率):

随着你运行的 NHST 数量的增加，你至少犯一个错误发现（“第一类错误”）的概率也会增加

Bonferroni correction：你将 p 值截止点除以你运行的测试数量，例如，在运行我们的三个成对独立 t 检验时，我们可能会使用 0.05/3 (= 0.0167) 的 p 值截止点。

Parametric tests：

做出假设，包括关于基础人群的形状和分布，以及关于数据类型，在违反这些假设的数据集上使用参数检验可能导致误导和不准确的结果

与以下概念相关：

正态分布Normally distributed的数据

连续（即区间或比例水平）数据

方差的同质性Homogeneity of variance（每个组的方差应该大致相等）

观测值应该是独立的

每个参数检验都有自己的特定假设，上述假设在不同的上下文（即不同的测试）中意味着不同的东西

Non-parametric tests：

不对数据的底层形状和分布做相同的假设

当我们的数据违反这些(参数检验的)假设时，通常可以使用这些测试的非参数等效测试来替代

Week2:

Supervised Learning：

一种机器学习类型，它使用标记的训练数据来训练模型

dependent variable(DV)因变量

independent variable (IV)自变量

Prediction：

“典型”的机器学习环境中，我们可能希望确定我们的模型在预测新的（未见过的unseen）数据方面的效果如何，我们可能首先训练一个模型，使用保留的withheld数据集对其进行测试，然后将其部署来预测新的数据。

Interpretation解释：

然而，在进行数据分析时，我们可能更感兴趣的是从模型中可以推断出什么。可能想要了解数据中潜在的关系可以教给我们什么

实践中既要Interpretation又要Prediction

Regression problems回归问题：其中我们的因变量是定量变量quantitative

Classification problems分类问题：其中我们的因变量是分类变量categorical

Regression Analysis回归分析：

一组用于量化变量之间关系的统计方法

回归允许我们拥有任意多的自变量

误差平方和Sum of Squared Errors (SSE)：

线性回归模型中的系数通常使用“普通最小二乘法”(Ordinary Least Squares OLS)来估计

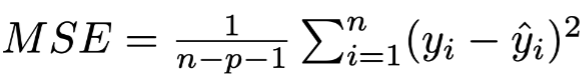
对于每个数据点，我们计算实际值和估计值之间的差的平方，然后加起来



均方误差Mean Squared Error (MSE)：

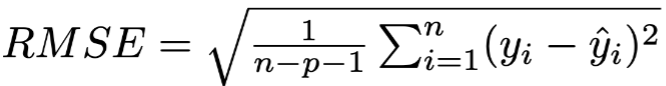
SSE 适合用于比较不同回归模型与数据集，但 SSE 值会随着 n 的增加而增加

所以可以使用MSE将SSE乘以1 / (n – p – 1) p是预测变量（自变量）的数量



Root Mean Squared Error (RMSE)：

MSE 代表平均平方误差，如果我们找到这个值的平方根，我们就找到了平均误差



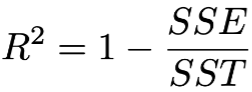
总平方和Sum of Squares Total (SST)：



R2 Value：

告诉我们回归模型解释了总变异性的比例

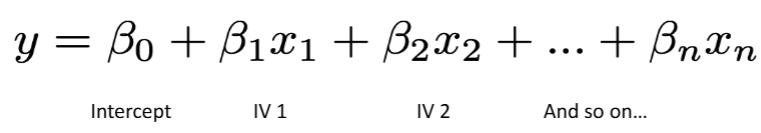
用于评估回归模型拟合程度的指标。它表示因变量的变异性中有多少可以被模型解释



R2告诉我们回归模型解释了总变异性比例，RMSE 告诉我们预测值和实际值之间的平均误差

Multiple Linear Regression多重线性回归：

多个自变量（预测变量）对因变量的影响，通常被称为“多重线性回归”，并且在数学上，理念是相同的：



Non-Linear Regression非线性回归：

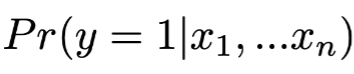
可能导致更好的“拟合”。然而，可解释性变得更加具有挑战性，也面临着过度拟合数据的风险

Logistic Regression：

目标是构建一个模型，以便我们可以从我们的自变量中估计出因变量

逻辑回归让我们可以使用自变量来估计因变量为1（或0）的概率

我们可以将这种方法应用于某些分类任务——只要我们的自变量是二元的（它只有两个可能的结果）

我们试图估计：在给定自变量的情况下，因变量被估计为特定类别（在这种情况下，y = 1）的概率，通常在概率大于0.5时估计y = 1

使用我们的逻辑回归模型：

可以使用模型来预测Y为类别1的概率（预测）

理解自变量如何影响Y值（解释）

Confusion Matrix混淆矩阵：

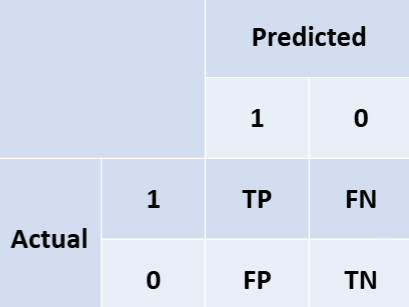
使用二元分类模型进行预测时，你可以有四种类型的结果：

True positive真阳性（正确预测1为1）

False positive假阳性（错误预测0为1）

True negative真阴性（正确预测0为0）

False negative假阴性（错误预测1为0）



Accuracy = (TP + TN) / (TP + FP + FN + TN)预测对的概率

Precision = TP / (TP + FP)被预测为{1}的结果中，预测正确的比例known as Positive Predictive Value阳性预测值

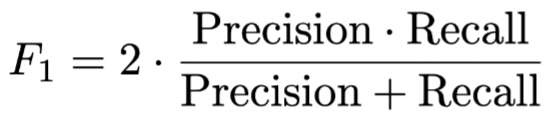
Recall / Sensitivity召回率 = TP / (TP + FN)实际上是{1}的结果中，预测正确的比例，也称为真阳性率True Positive Rate

Specificity特异性 = TN / (TN + FP)实际上是{0}的结果中，预测正确的比例，Also known as True Negative Rate

False Positive Rate假阳性率 = FP / (FP + TN)被预测为{1}的结果中，预测错误的比例，Also known as Type I Error Rate

False Negative Rate假阴性率 = FN / (FN + TP)被预测为{0}的结果中，预测错误的比例Also known as Type II Error Rate

F1分数是一个很好的“总体”指标，它是精确率和召回率的调和平均值：

图片包含 图示

AI 生成的内容可能不正确。

ROC(Receiver Operating Characteristic接收者操作特征) Curve：

一种直观展示二元分类器预测性能的方法。它绘制了不同分类阈值下的真阳性率与假阳性率。理想情况下，你希望ROC曲线接近左上角

Tree：

为什么用：trees are a non-linear data structure，通常在计算效率上更高，可以进行更快的遍历，从而更方便地访问数据

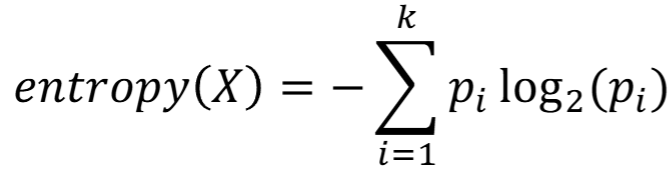
Decision Trees：

Entropy熵：衡量系统中不纯度impurity或不可预测性的机器学习指标，是衡量正在处理的信息中的混乱程度的指标

如果一个数据集中包含相等数量的正负数据点，则熵为 1

如果一个数据集中只包含正数据点或只包含负数据点，则熵为 0

熵越低，意味着可预测性越高

计算：pi 是类别 i 元素的比例

Information Gain：

数据集中观察到的模式，以及熵的减少（一个衡量属性“好”和“坏”的标准）

Information Gain = (Entropy of parent node) - (Entropy of child node)

本质上，我们是在比较在某个属性上分裂后节点的熵与分裂前的熵，换句话说，我们通过分裂获得了多少信息

决策树的每一步，我们都要最大化信息增益：

1. 从一组属性中选择信息增益最高的属性作为父（根）节点。
2. 然后为该属性的每个值构建子节点。
3. 然后，我们迭代地重复这个过程，直到构建出整个树。

Best attribute = highest information gain

Pruning剪枝：

通过移除对预测能力贡献较小的树枝来减小决策树规模的技术

一种正则化regularization方法，可以降低最终模型的复杂度，从而减少过拟合overfitting

two main methods of pruning:

Pre-pruning: 在算法完全分类数据之前停止构建树

Post-Pruning: 允许算法完全构建树，然后在验证误差得到改善的情况下，将一些非叶节点替换为叶节点

Pre-pruning:

涉及在训练过程早期停止：

如果满足某些条件，则当前节点不会被分割（即使节点不是 100% 纯净的）

然后，该节点将变为叶节点，并带有当前集中多数类的标签

当将不纯的节点转换为叶节点时，我们甚至可以使用类分布作为预测置信度值

Post-Pruning:

在后剪枝中，如果可以降低验证误差，我们会以自底向上的方式剪枝节点

ID（Iterative Dichotomiser迭代二分器）3 Algorithm：

1. 使用类分布确定整体数据集的熵
2. 对于每个特征：
   1. 计算分类值的熵
   2. 评估特征每个唯一分类值的信息增益
3. 选择生成最高信息增益的特征
4. 迭代应用以上所有步骤来构建决策树结构

C4.5 Algorithm：

extension of ID3，有改进：

能处理分类categorical（离散discrete）和数值numerical（连续continuous）属性

通过找到最佳分割阈值来分割连续属性

能够在训练training和推理inference阶段处理缺失值

在训练期间，计算信息增益时不使用缺失值

在推理期间，通过探索所有相应的分支来处理缺失值

能够处理具有不同成本的属性

以自底向上的方式进行后剪枝，以移除那些降低验证误差的分支（即增加泛化generalization能力）

Handling Numerical Attributes：

一种方法是找到最佳的分割值：即我们计算如果我们在值 t 处分割数值属性 a 的信息增益

Handling Missing Values at Training Time：

1. Set them to the most common values
2. Set them to the most probable value given the label
3. Add a new instance for each possible value
4. 保留它们为未知，但在评估该属性的增益时丢弃样本
5. 在其他所有属性（包括标签）上构建一个决策树来预测缺失值

Generative vs Discriminative Classifiers:

当我们想要根据某些可观察变量（X）对某个分类变量（Y）进行分类时，我们希望估计：P(X|Y)

在处理分类任务时，我们有时会区分两种主要方法：

Discriminative classifiers判别式分类器

Generative classifiers生成式分类器

Discriminative Classifiers:

旨在模拟给定可观察变量x的目标Y的条件概率：P(Y|X=x)

它们通过直接从训练数据中估计的参数P(Y|X)来实现这一点

Generative Classifiers:

相比之下，生成式分类器旨在模拟给定目标y的可观察变量X的条件概率：P(X|Y=y)

它们通过估计参数来实现P(X|Y),P(Y)

然后使用贝叶斯定理进行计算P(Y|X=x)

Bayes’ Theorem:

文本

AI 生成的内容可能不正确。

Naïve Bayes:

朴素贝叶斯是指一类用于分类的监督学习算法

假设在给定类别的情况下，所有特征都是条件独立的

‘Bag of Words’ Model:

取文档中的每一个词，把它放入一个“袋”中，然后摇一摇。我们丢弃所有的顺序、上下文等。这本质上就是假设所有特征彼此完全独立的意思

图表, 散点图

AI 生成的内容可能不正确。

Support Vector Machines:

支持向量机（SVM）是一种监督学习算法，通过寻找最佳分离边界来对数据进行分类

通过最大化决策边界来实现这一点，从而最大化每个类别之间的距离

为什么要最大化边界:

1. 更宽的间隔意味着我们的线性分隔器linear separator离任何不确定的点更远，这使我们对其分类更有信心2. 由于这个原因，我们的分隔器通常能更好地泛化generalize到未见过的数据，3. 我们的模型更加鲁棒和稳定，因为我们的间隔只依赖于更少的训练数据点，换句话说，该模型对噪声noise更鲁棒robust，并且不受immune任何非支持向量删除的影响4. 最大化间隔是一个凸convex优化问题。因此，当我们最大化间隔时，我们保证能找到一个最优解，而没有局部最小值

Hard Margin硬间隔 SVMs：

我们的数据点将被间隔完美地分隔开，不允许任何数据点落在我们的间隔内

Advantages：

**简单性Simplicity**：硬间隔 SVM 确保类别被间隔完美地分隔开

**计算效率Computational Efficiency**：找硬间隔 SVM 的最优间隔简单且数学上高效

Disadvantages：

**对异常值Outliers敏感**：单个数据点可能会显著影响决策边界的位置

**不适用于非线性non-linear数据**：如果数据不能线性分隔，硬间隔 SVM 将会失败

Soft Margin 软间隔SVMs：

允许我们的 SVM 在允许一些数据点被错误分类方面更加灵活，允许一些数据点落在间隔内，甚至落在分类区域的错误一侧，这样可以有更宽的间隔，代价是少量错误分类

Advantages：

**对异常值的鲁棒性**：软间隔 SVM 比硬间隔 SVM 对异常值和噪声更加鲁棒

**适用于非线性数据**：由于允许一些错误分类，软间隔 SVM 可以处理无法线性分隔的数据

Disadvantages：

**参数调整Parameter tuning**：更多的参数可能需要仔细调整以找到最佳值

**更容易过拟合overfitting**：如果软间隔 SVM 允许太多错误分类misclassifications，这可能导致模型过拟合

Revision:

Briefly explain what you understand by the term ‘Data Mining’:

Data Mining is the process of discovering patterns and extracting useful information from data

Briefly explain what you understand by the term ‘Data Visualization’:

Data Visualization is the process of designing visual representations of data. It can also be used to refer to these visual representations themselves

Briefly explain what you understand by the term ‘Exploratory Data Analysis’:

Exploratory Data Analysis is the process of exploring your data. This might involve summarizing or visualizing aspects of data, identifying outliers, cleaning, and pre-processing

Briefly explain what you understand by the term ‘a weak positive correlation’:

Correlation is a measurement of the strength of relationship between two variables.

A weak positive correlation would have a relatively small positive correlation coefficient. In other words, as variable x increases, variable y will have a slight tendency to increase.

Briefly explain what you understand by the term ‘a strong negative correlation’:

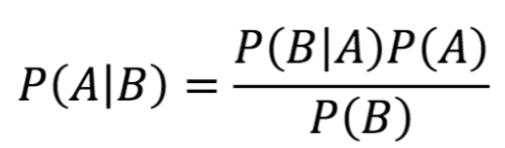
A strong negative correlation would have a relatively large inverse correlation coefficient, representing two variables which are tightly and inversely associated with each other. In other words, as variable x increases, variable y will have a strong tendency to decrease

Briefly explain what you understand by the term ‘no correlation’:

No correlation would refer to a correlation coefficient close to 0. In other words, there is little to no relation between variables x and y

Briefly explain what you understand by the term ‘correlation does not imply causation’:

The term ‘correlation does not imply causation’ refers to the fact that correlation and causation are distinct terms. Correlation is a measure of the strength of an association between two variables, whereas causation is the process of one event causing or producing another event. From a given correlation coefficient, we cannot know whether causation exists or not



**P(A|B**) 是**后验Posterior概率**：在观察到事件 B 的条件下，假设 A 成立的概率

**P(B|A**) 是**似然Likelihood概率**：在假设 A 成立的条件下，数据 B 出现的概率

**P(A**) 是**先验Prior概率**：在观察到证据之前，假设 A 成立的概率（与数据无关）

**P(B**) 是**边缘Marginal概率**：证据出现的概率（与假设无关）