

# MM5425 商业分析

---

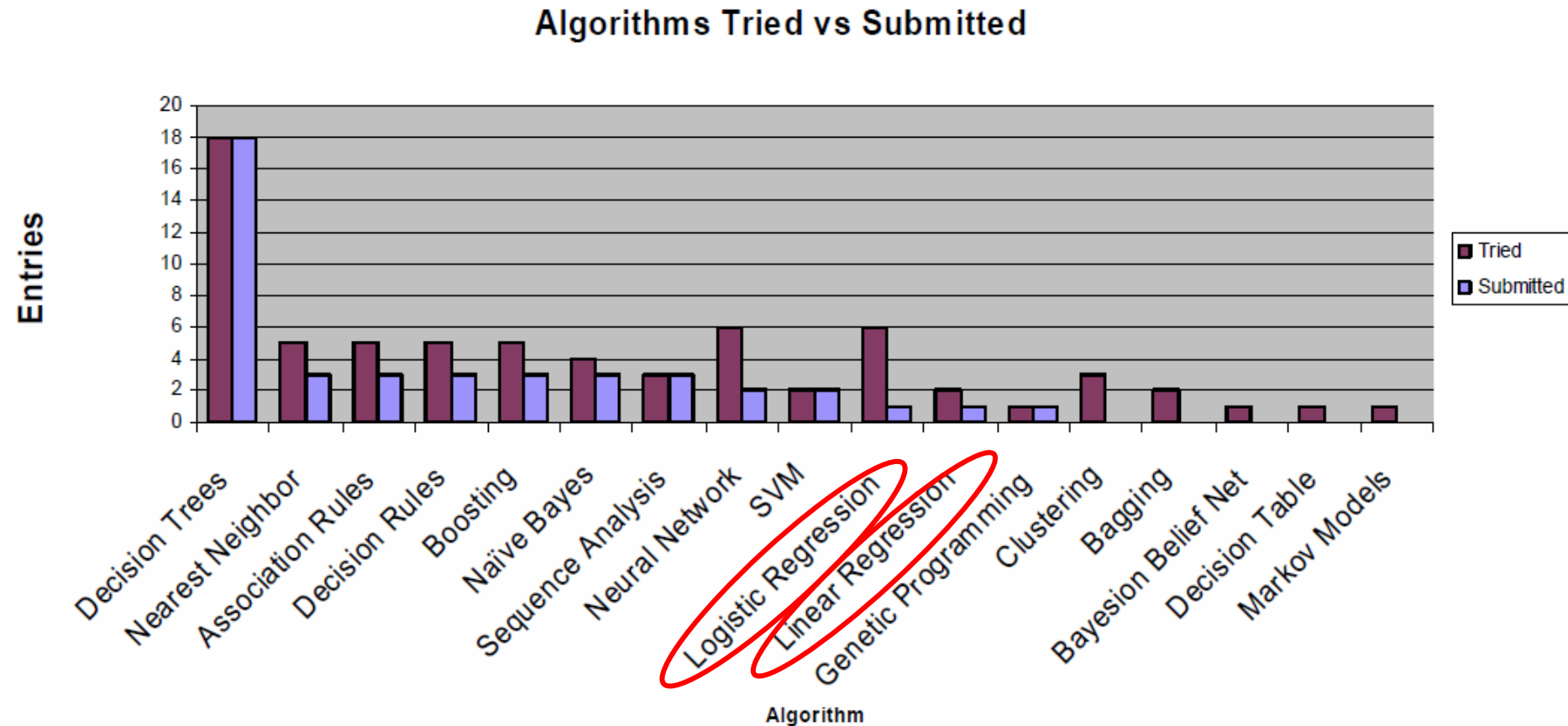
WEEK 7 LECTURE – NONLINEAR REGRESSION

# WK7 目录 Contents

---

- 逻辑回归 Logistic Regression
  - 什么是逻辑回归 What is logistic regression?
  - 逻辑回归解释性 Logistic regression interpretation
  - 逻辑回归 vs. 决策树 Logistic regression vs. decision tree
- 回归树 Regression Tree

# 常用算法Commonly Used Algorithms



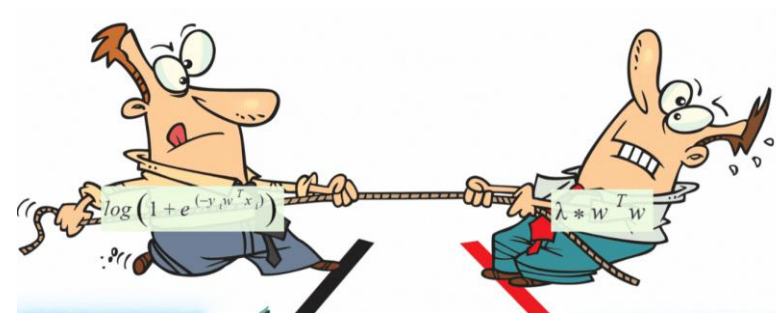
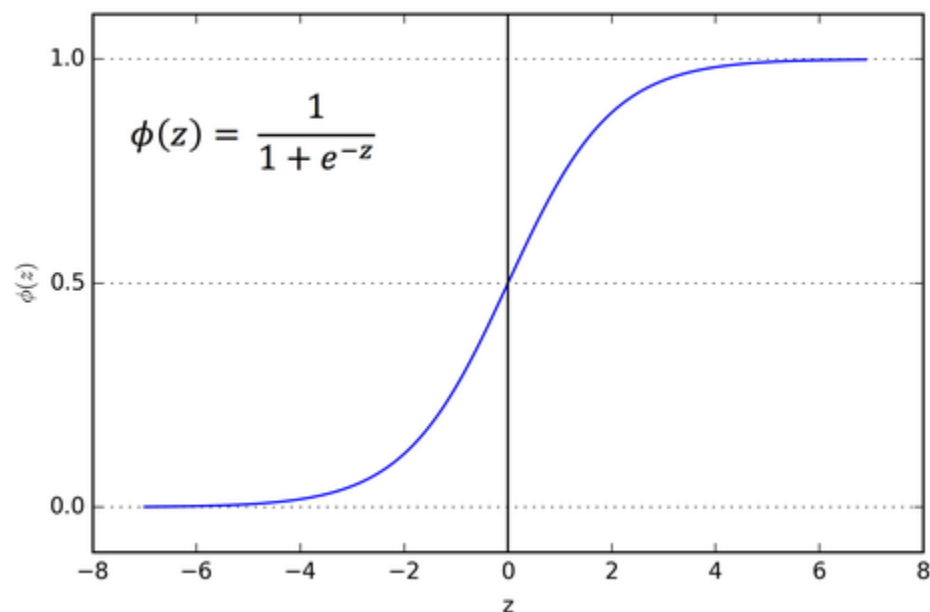


# 消费者如何做出决定

- **消费者偏好**告诉我们，如果假设这些商品组合对消费者是免费提供的，消费者会如何对任意两组商品进行排序。
  - ✓“我更愿意选择法式香草拿破仑星冰乐，而不是摩卡。”
- **效用函数**（也称为价值函数）：衡量消费者从消费特定数量商品中获得的满足感水平。
  - ✓ 消费者可能会从您的产品或服务不同方面获得满足感。。

# 解读效用的力量

对于市场营销经理来说，研究效用函数的目的是识别关键因素，并找到最大化顾客满意度的策略，从而提高顾客购买的可能性。



# 其他例子 Other Examples

- 信用评分 Credit scoring

- 一家金融公司试图预测信用评分，以辅助特定的决策制定。



- 酒店预订

- **Booking.com** 尝试预测用户意图并识别相关实体。你会去哪里，你更喜欢在哪里停留，你计划做什么？有些预测甚至在用户在搜索栏输入任何内容之前就已经完成。

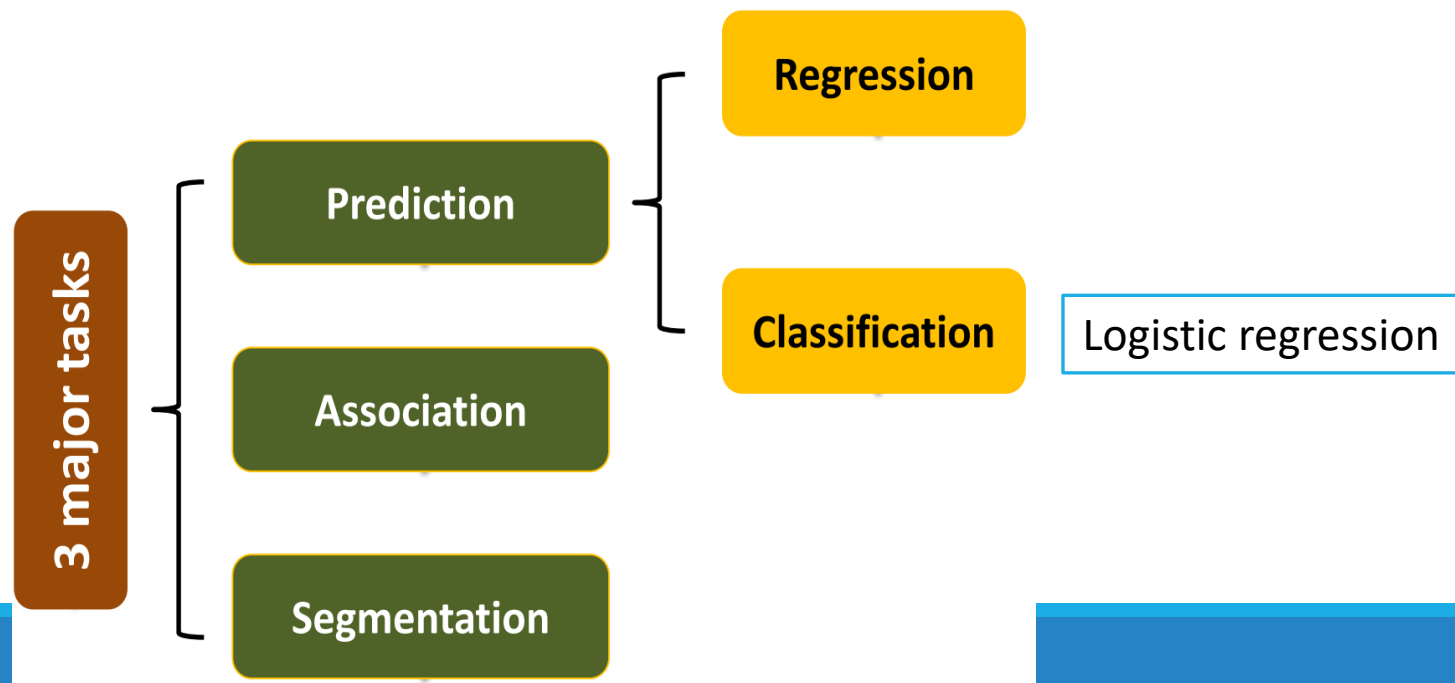
# 逻辑回归 Logistic Regression

---

CLASSIFICATION MODEL

# 什么是逻辑回归 What is a logistic regression?

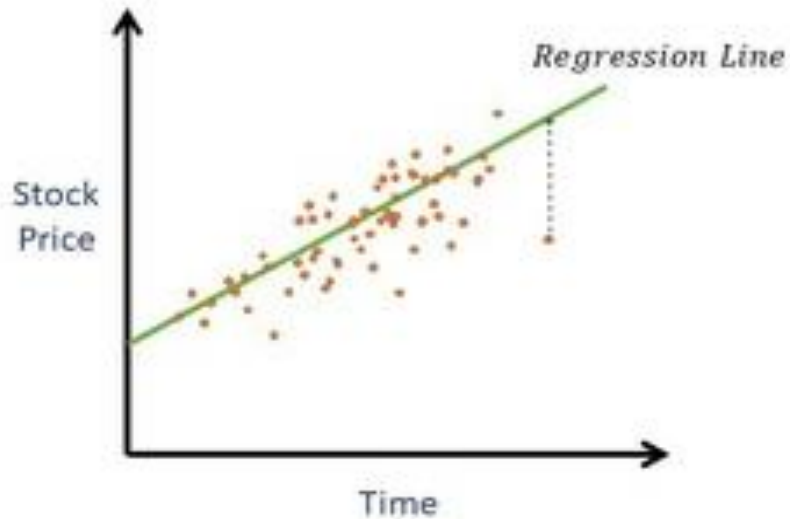
- 逻辑回归是一种分类模型。
- 数据中目标变量的取值是类别型的。
- 该模型会生成一个数值估计——某一特定类别的概率。



# 线性回归 vs. 逻辑回归

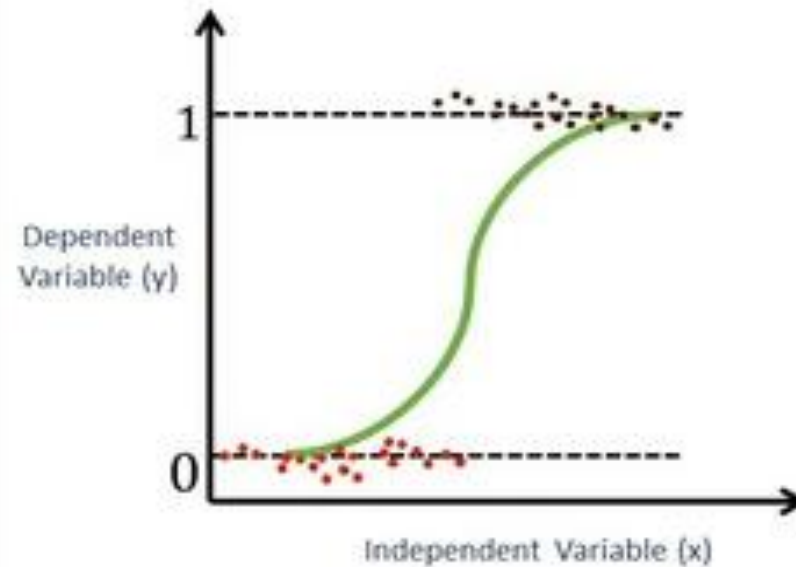
## Linear Regression

- Aim is to predict continuous valued output.
- Output value can be any possible integer number.



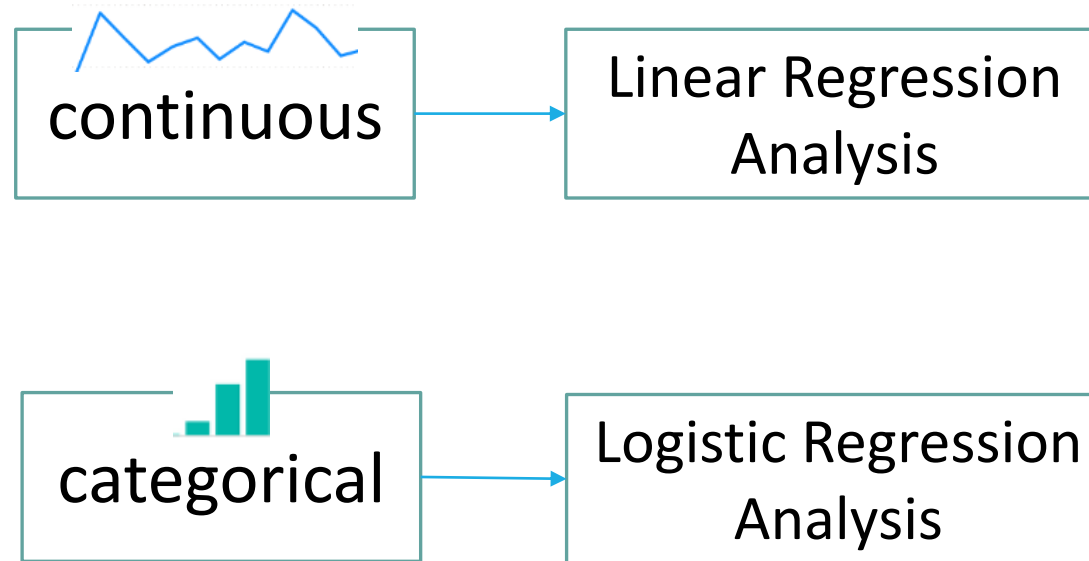
## Logistic Regression

- Aim is to predict the label for input data.
- Output is categorical (Binary) i.e. 0/1, True/False, etc.



# 线性回归 vs. 逻辑回归

---



# 分类 Classification



Binary variable:

- Yes it will happen
- No it will not happen

Probability:

- $p \sim$  the probability of  $y=1$
- $1-p \sim$  the probability of  $y=0$

Logistic regression

$$y = \alpha + \beta x$$

$$p = \frac{1}{1 + e^{-y}}$$

# 逻辑回归 Logistic Regression

---

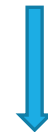
$$y = \alpha + \beta X$$



$$p = \frac{1}{1 + e^{-y}} = \frac{1}{1 + e^{-(\alpha + \beta X)}}$$



$$\frac{p}{1-p} = e^{(\alpha + \beta X)}$$



$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta X$$

# 如何得到 $\alpha$ 和 $\beta$ ?

---

## 最大似然估计

确定能够最大化我们观察到的数据的似然性的“ $\alpha$ 和 $\beta$ ”。

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 \dots + \beta_n x_n$$

- $\text{odd} = \frac{p}{1-p}$
- 如果“ $x_1$ ”增加1个单位，在其他变量保持不变的情况下，对数几率将改变“ $\beta_1$ ”。

# 例子Example 1: 学校表现评测

**问题:** 有一组20名学生为考试学习了0到6小时。我们能否根据学生为考试学习的小时数来预测他们是否会通过考试??

可用数据: 历史学生数据, 包括出勤率、阅读的书籍数量和成绩。

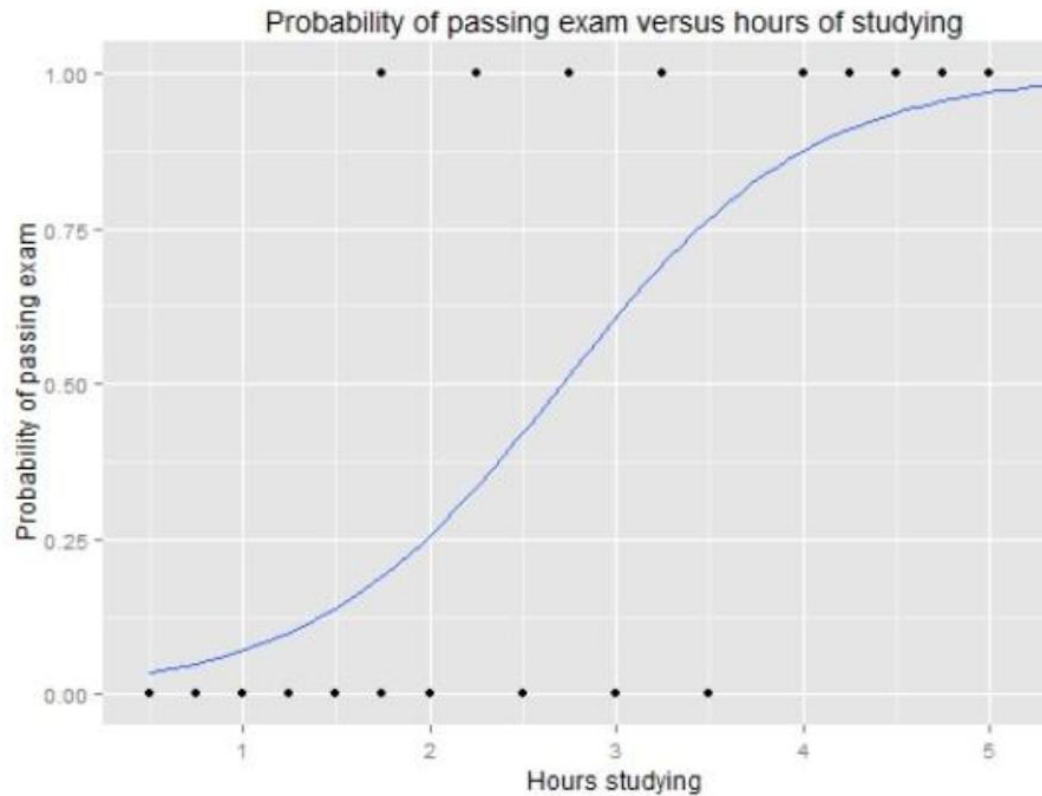
0: failed; 1: passed

Hours	0.50	0.75	1.00	1.25	1.50	1.75	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50	4.00	4.25	4.50	4.75	5.00	5.50
Pass	0	0	0	0	0	0	1	0	1	0	1	0	1	0	1	1	1	1	1	1

目标变量  $y$ : 是否通过考试

$x$ : 学习了多少小时

# 例子 1: 使用逻辑回归



如果一名学生学习了2小时，估算其通过考试的概率为0.26。

如果一名学生学习了4小时，估算其通过考试的概率为0.87。

$$\text{Probability of passing exam} = \frac{1}{1 + \exp(-(1.5046 \cdot \text{Hours} - 4.0777))}$$

# 例子 2: 员工流失

问题:

- 员工是否会离职

数据:

- 员工历史数据

可以使用模型: 逻辑回归



$$\ln\left(\frac{p_{\text{attrition}}}{1-p_{\text{attrition}}}\right) = \alpha + \beta_1 * \text{Overtime} + \beta_2 * \text{Salary}$$

## 例子 2: 员工流失

---

Coefficients:	Estimate	Pr(> z )
■ Intercept	-1.40	5.25e-16***
■ Overtime(Yes)	1.39	3.43e-16***
■ Income	-0.000137	2.62e-07***

员工流失预测函数

$$\ln\left(\frac{p_{attrition}}{1-p_{attrition}}\right) = -1.40 + 1.39 * \text{Overtime} - 0.000137 * \text{Salary}$$

# 逻辑回归 vs. 决策树

---

- 对于较小的训练集规模，逻辑回归比决策树归纳具有更好的泛化准确率。
  - 对于较小的数据集，决策树归纳方法更容易出现过拟合。
- 分类树比线性逻辑回归具有更灵活的模型表示能力。
- 当训练集较大时，决策树归纳的灵活性可以成为一种优势：
  - 树模型能够表示特征与目标之间显著的非线性关系。

# 分类（Classification）与回归（Regression）

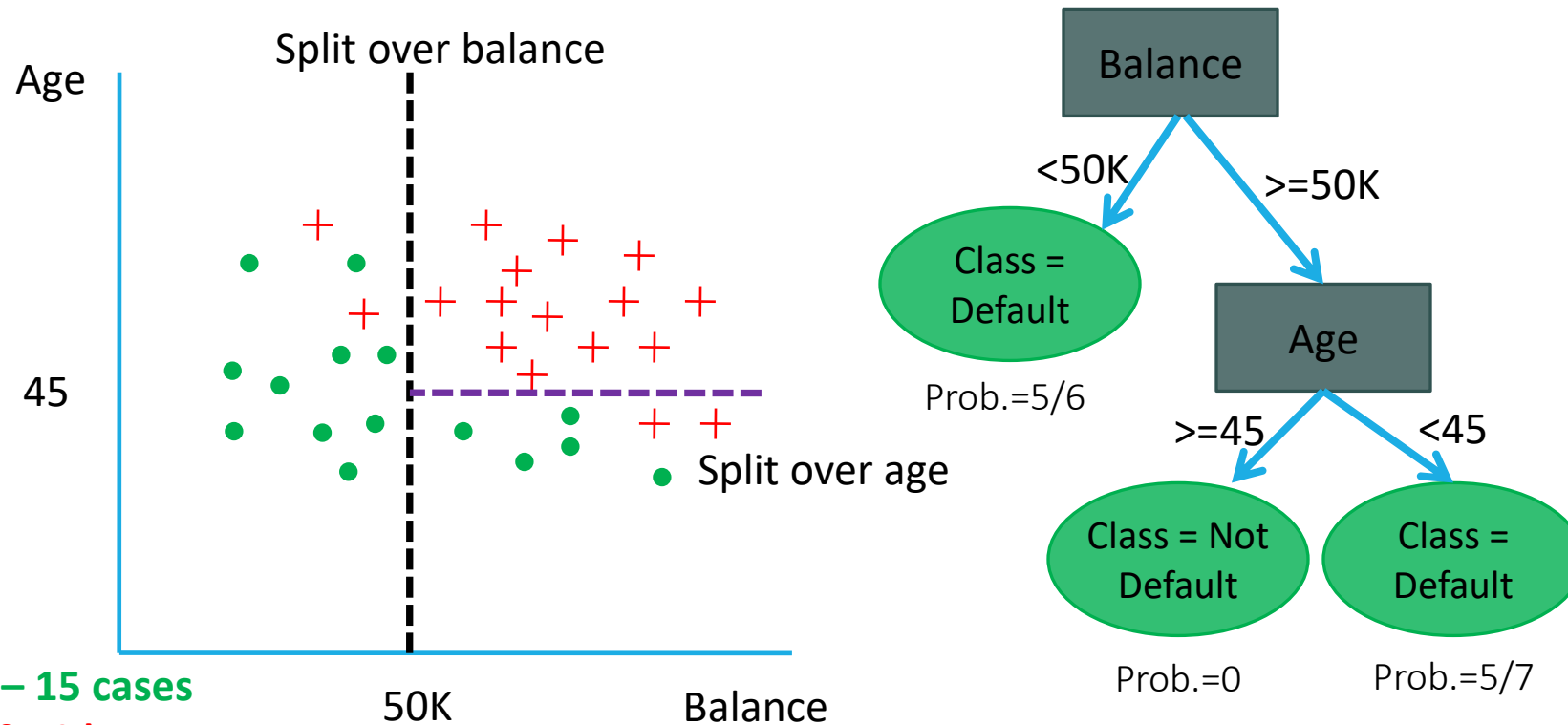
$$y = \alpha + \beta x$$

Classification	Regression
– Predict if a student will pass this course	– Predict the grades of a student
– Predict if the bank client is able to pay the debt	– Predict the credit score of a bank client
– Predict if a movie will become a popular one	- Predict the total sales of a movie

方面	分类 (Classification)	回归 (Regression)
输出	离散类别/标签	连续数值
示例	邮件垃圾识别	房价预测
常用算法	逻辑回归等	线性回归等

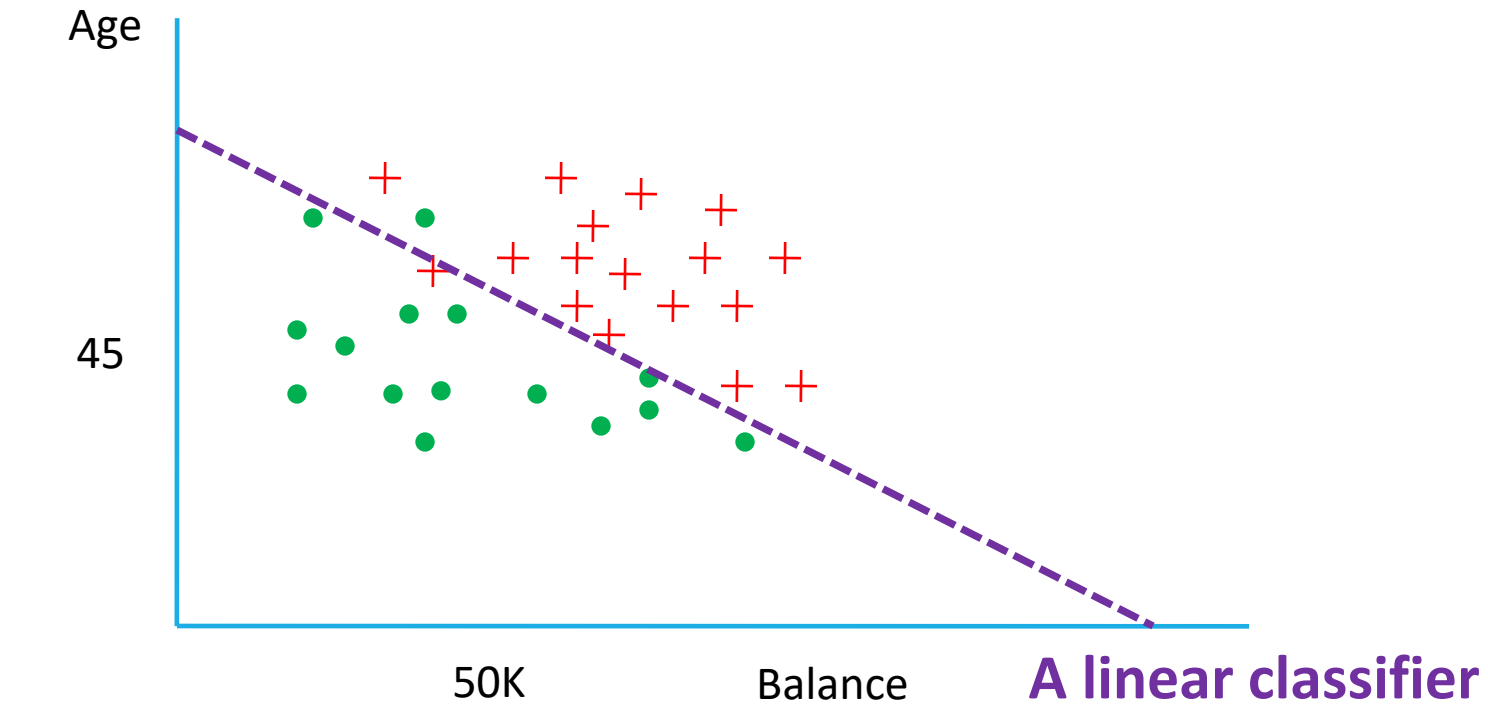
# 回顾：决策树进行分类

分类树通过轴平行的决策边界将样本空间划分为不同的区域。



- **Bad risk (Default) – 15 cases**
- + **Good risk (Not default) – 17 cases**

# 替代划分方式 Alternative Partitioning

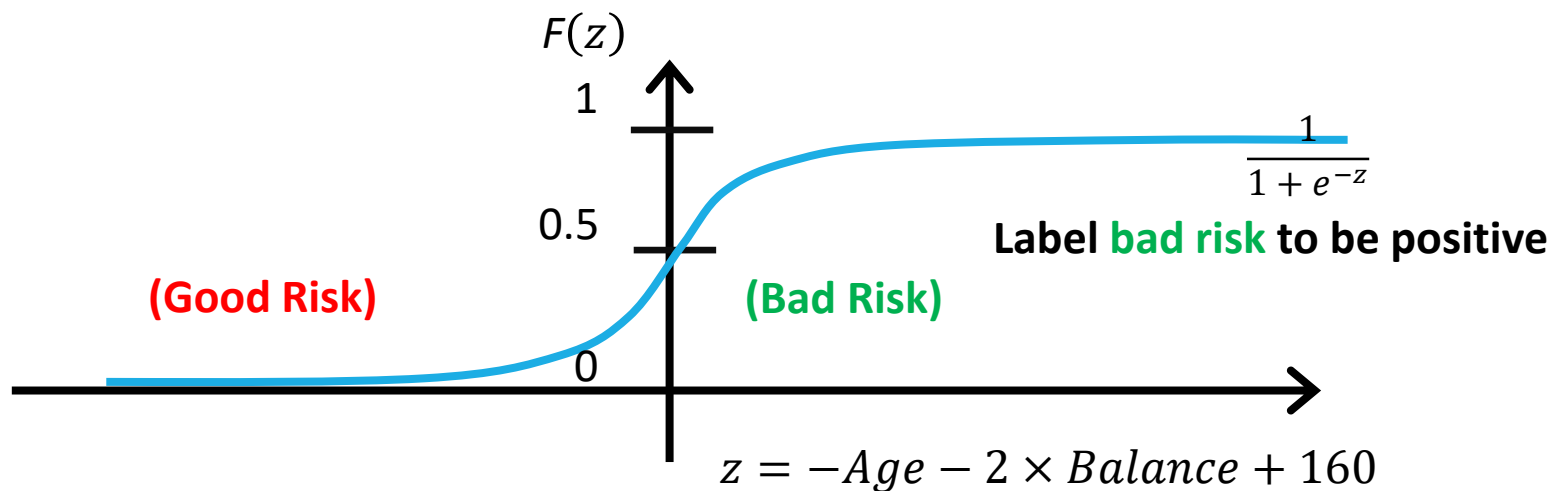


- Bad risk (Default) – 15 cases
- + Good risk (Not default) – 17 cases

# 分类：逻辑回归

默认情况下，我们将阈值设为 0.5:

If  $p = F(z) \geq 0.5$ , 预测 “ $y = 1$ ”; If  $p = F(z) < 0.5$ , 预测 “ $y = 0$ ”



逻辑回归 Logistic function  $F(z) = \frac{1}{1+e^{-z}}$  where  $z = -Age - 2 \times$   
 $Balance + 160$

预测类别 Predict Class =  $\begin{cases} \bullet & \text{if } F(z) \geq 0.5 \\ + & \text{if } F(z) < 0.5 \end{cases}$  or  $\begin{cases} \bullet & \text{if } -Age - 2 \times Balance + 160 \geq 0 \\ + & \text{if } -Age - 2 \times Balance + 160 < 0 \end{cases}$

# 回归树 Regression Trees

---

CONTINUOUS DATA PREDICTION

# 回归树 Regression Trees

---

CART: 分类与回归树

回归树: 当目标变量为数值型时使用

预测输出: 叶节点中训练样本的均值

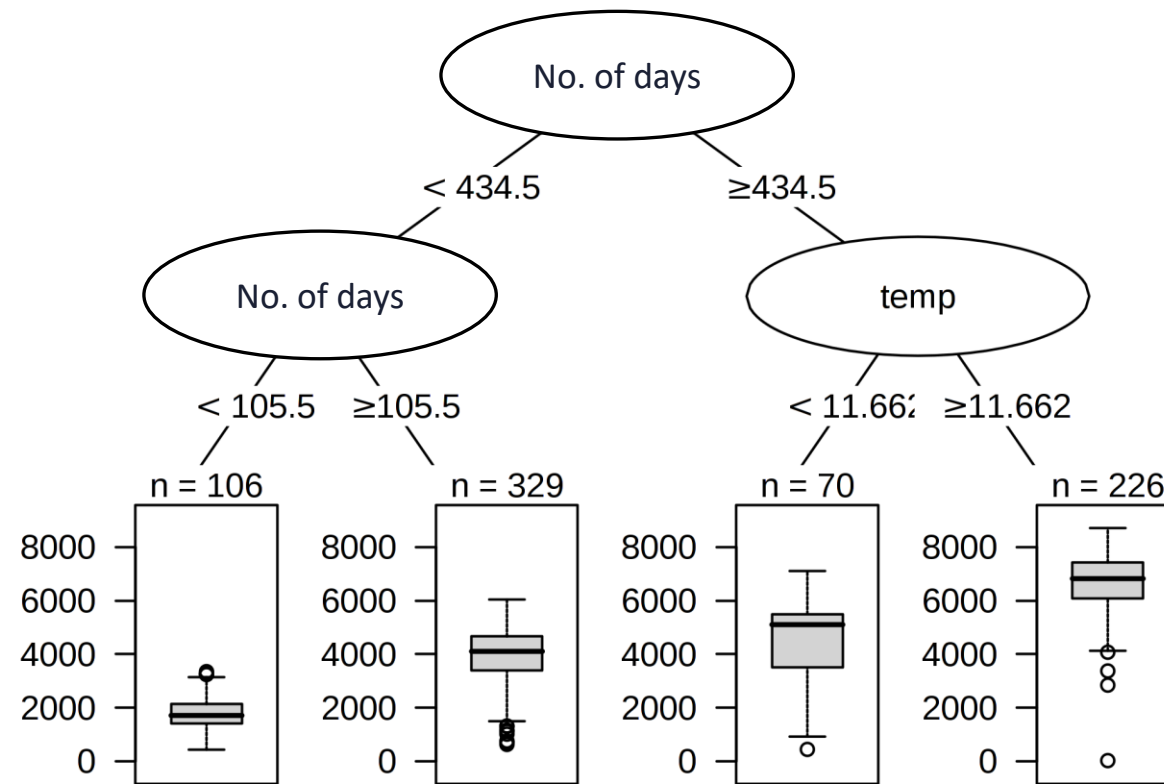
不纯度度量: 该子集内的平方误差之和

$$SS_E = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

划分是基于

Where  $y_i$  is the true value and the  $\hat{y}_i$  is the predicted value

# 回归树例子



# 回归的评估指标

均方误差（Mean Squared Error, MSE）：实际值与预测值之间差值的平方的平均值。

公式如下：

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

均方根误差 Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

平均绝对误差 (MAE)

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

# 回归的评估指标

---

R squared ( $R^2$ ): 决定系数, 表示因变量 ( $y$ ) 的变异中有多少比例可以通过自变量 ( $x$ ) 来预测。

公式如下:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} = 1 - \frac{SS_{residual}}{SS_{total}}$$

- 当所有的预测值都是  $\bar{y}$  (即实际值的均值) 时,  $R^2=0$  (基线情况)
- 比基线 (均值预测) 更差的模型, 其  $R^2$  会为负值.



Any Questions?

**Reference:**

Business analytics: The science of data-driven decision making / U. Dinesh Kumar, New Delhi Wiley India, 2022