# 伯努利分布（Bernoulli Distribution）

伯努利分布是****概率论中最基础的离散概率分布之一****，由瑞士数学家雅各布・伯努利（Jakob Bernoulli）命名。它描述的是****只有两种可能结果的单次随机试验****，例如 “抛硬币正面朝上”“投篮命中” 等场景。

## 一、核心定义

若一个随机试验（称为 “伯努利试验”）只有两种互斥的结果：

****“成功”****（记为 1），概率为 *p*（0<*p*<1）；****“失败”****（记为 0），概率为 1−*p*；

则该试验的结果服从伯努利分布。

## 二、概率质量函数（PMF）

对于随机变量 *X*（取值为 0 或 1），伯努利分布的概率质量函数（描述每个取值的概率）为：  
*P*(*X*=*k*)={*p*1−*p*​若 *k*=1（成功）,若 *k*=0（失败）.​  
也可简写为：*P*(*X*=*k*)=*pk*(1−*p*)1−*k*（*k*=0 或 1）。

## 三、数字特征（期望与方差）

伯努利分布的期望（均值）和方差是其重要特征，计算简单且常用：

****期望（E [X]）****：  
表示 “长期试验中成功的平均概率”，公式为：*E*[*X*]=1⋅*p*+0⋅(1−*p*)=*p*

****方差（Var [X]）****：  
表示结果的离散程度，公式为：Var[*X*]=*E*[*X*2]−(*E*[*X*])2=*p*−*p*2=*p*(1−*p*)

注：由于 *X* 只能取 0 或 1，*X*2=*X*，因此 *E*[*X*2]=*E*[*X*]=*p*。

## 四、适用场景

伯努利分布适用于****单次、二元结果****的随机事件，例如：

* 抛一枚硬币，正面（1）或反面（0）；
* 一次抽奖，中奖（1）或未中奖（0）；
* 检测一个产品，合格（1）或不合格（0）。

## 五、与其他分布的关系

伯努利分布是许多复杂分布的基础：

1. ****二项分布****：若进行 *n* 次独立的伯努利试验，“成功次数” 服从二项分布（可视为伯努利分布的 “多次试验扩展”）。
2. ****几何分布****：若重复伯努利试验，直到首次成功为止，“试验次数” 服从几何分布。
3. ****泊松分布****：当二项分布中 *n* 很大且 *p* 很小时，可近似为泊松分布，而二项分布本身基于伯努利试验。

## 六、示例说明

假设 “投篮命中” 为成功（*X*=1），概率 *p*=0.6，则：

* 命中的概率：*P*(*X*=1)=0.6；
* 未命中的概率：*P*(*X*=0)=0.4；
* 期望：*E*[*X*]=0.6（长期投篮的平均命中率）；
* 方差：Var[*X*]=0.6×0.4=0.24。

## 总结

伯努利分布是概率统计的 “基石” 之一，其简单性使其成为理解更复杂分布（如二项分布、逻辑回归等）的基础。记住其核心：****单次试验、二元结果、概率***p***和**1−*p***。

# 线性回归

# 逻辑回归算法

## 核心概念

虽然名字里有“回归”，但逻辑回归（Logistic Regression）本质上是一种用于解决分类问题的统计学习方法，特别是在二分类问题上应用极为广泛。

它之所以被称为“回归”，是因为它的核心思想源于线性回归，但其输出经过了非线性变换，使其适用于分类任务。

## 核心目的

预测一个离散的类别标签（最常见是二元标签，如 0 vs. 1，是 vs. 否，垃圾邮件 vs. 正常邮件）。更重要的是，它输出的是某个样本属于某个类别的概率（介于 0 和 1 之间）。例如，“这封邮件是垃圾邮件的概率是 85%”。

## 工作原理简述

基础：线性组合像线性回归一样，逻辑回归首先计算输入特征 (X₁, X₂, ..., Xₙ) 的线性加权和： z = β₀ + β₁X₁ + β₂X₂ + ... + βₙXₙ 其中，β₀ 是截距项（偏置），β₁ 到 βₙ 是模型需要学习的权重（系数），对应于每个特征的重要性。

关键转化：Sigmoid 函数 线性回归的输出 z 可以是任意实数（-∞到+∞）。为了将这个值映射到一个表示概率的范围（0 到 1），逻辑回归引入了 Sigmoid 函数（也叫 Logistic 函数）： σ(z) = 1 / (1 + e^(-z))

Sigmoid 函数的作用：

它是一个非线性函数，形状像字母“S”,它将任何实数 z 压缩到 (0, 1) 区间内。

当 z 非常大时，σ(z) 趋近于 1。当 z 非常小（负值很大）时，σ(z) 趋近于 0。

当 z = 0 时，σ(z) = 0.5。输出概率： 模型的最终输出是： P(Y=1 | X) = σ(z) = 1 / (1 + e^(-(β₀ + β₁X₁ + ... + βₙXₙ))) 这个 P(Y=1 | X) 表示在给定输入特征 X 的条件下，样本属于类别 1 的概率。

决策： 为了做出最终的类别预测，通常设定一个阈值（Threshold，默认是 0.5）：

如果 P(Y=1 | X) >= 0.5，则预测类别为 1。

如果 P(Y=1 | X) < 0.5，则预测类别为 0。 （阈值可以根据具体业务需求调整，例如在疾病诊断中，为了减少漏诊，可能降低阈值，如设为 0.3）。

## 重要特点

决策边界是线性的：

逻辑回归通过 Sigmoid 函数将线性回归的输出转化为概率，但其决策边界（即 P(Y=1 | X) = 0.5 的点）在特征空间中仍然是一条直线（二维空间）/一个平面（三维空间）/一个超平面（高维空间）。

这意味着逻辑回归是一个线性分类器。

注意：可以通过引入多项式特征（如 X₁², X₁X₂, X₂²）来拟合非线性的决策边界，但本质上学习到的仍然是原始特征空间转换后的线性关系。

输出是概率：

这是逻辑回归区别于其他分类算法（如支持向量机/SVM、决策树）的一个巨大优势。输出概率提供了预测的不确定性度量，在很多场景（如风险评估、排序推荐）中非常有用。

损失函数：交叉熵损失（Log Loss）

逻辑回归通过极大似然估计来学习最优参数 β。

其对应的损失函数是交叉熵损失或对数损失。它的设计目标是：对于预测正确的样本，损失很小；对于预测错误且置信度很高的样本（例如模型以 99% 的概率预测错误），惩罚非常大。

公式为：Loss = - [y \* log(p) + (1 - y) \* log(1 - p)]，其中 y 是真实标签（0 或 1），p 是预测为类别 1 的概率。

优化方法：

为了最小化交叉熵损失并找到最优的权重 β，通常使用梯度下降或其变种（如随机梯度下降/SGD、批量梯度下降）等迭代优化算法。

优点

计算高效： 训练和预测的速度都很快，尤其适用于大数据集。

易于实现和解释：

模型的权重 β 有直观的解释：权重绝对值的大小表示特征重要性，正负号表示特征与目标类别（通常指 Y=1）的关系是正向的还是负向的（在特征标准化后）。

输出概率易于理解。

正则化： 容易应用 L1（Lasso）或 L2（Ridge）正则化来防止过拟合，提高模型泛化能力（通过调整 C 或 λ 超参数）。

良好的基线模型： 由于其简单、快速和可解释性，逻辑回归通常是解决二分类问题的首选基线模型。

缺点/局限

本质是线性分类器： 难以直接捕捉特征之间复杂的非线性关系（需要通过特征工程来克服，这会增加复杂性）。

对特征相关性和多重共线性敏感： 高度相关的特征或共线性会影响权重估计的稳定性和可解释性。

需要较大的样本量： 样本量太小可能导致模型不稳定。

特征工程重要： 其效果很大程度上依赖于特征的质量和合适的预处理（如特征缩放、缺失值处理、特征编码、特征选择）。

可能被“更好”的模型替代： 对于非常高维、复杂非线性或需要非常高精度的问题，像随机森林、梯度提升树（如 XGBoost, LightGBM）或神经网络等模型可能表现得更好，但它们通常牺牲了可解释性和训练速度。

总结

逻辑回归是一种强大、简单且应用极其广泛的统计学习方法，主要用于解决二分类问题。 它的核心在于利用 Sigmoid 函数 将线性回归模型的输出转换为一个表示概率的值（0 到 1 之间）。它的优点在于高效、易于实现和解释（提供特征重要性）、易于正则化以及能够直接输出概率。虽然它是一个线性模型，但通过精心设计的特征工程，它在许多实际问题中仍然表现出色，并且是构建分类模型时一个非常重要的基准点。

简单来说：逻辑回归 = 线性回归 + Sigmoid函数 -> 概率输出 -> 用于分类。