逻辑回归(Logistic Regression)是一种借鉴统计领域的机器学习技术。它是二类分类问题的首选方法。在这篇文章中，您将发现用于机器学习的逻辑回归算法。在读完这篇文章后，您将了解到:

* 描述逻辑回归时的许多专有名称和术语(如 log odds和logit)。
* 逻辑回归模型中的表达式。
* 用于从数据中学习逻辑回归模型系数的技术。
* 如何使用学习好的逻辑回归模型进行实际的预测。

这篇文章是为对机器学习的应用、预测建模感兴趣的开发人员编写的，您不需要具备线性代数或者统计学的背景知识。

# 1.Logistic函数

逻辑回归以在该模型中使用的核心函数命名，即逻辑函数。逻辑函数也被称为Sigmoid函数，是统计学家发明的用于描述生态中人口增长、环境承载力迅速上升的特性。它是一条S型曲线，将任何值映射到0和1之间，但是很难取到0或者1.公式如下：

**1 / (1 + e^-value)**

其中,e是自然对数的基数，value是你想要转换的实际数值。下图是使用逻辑函数将-5和5之间的数字转换到0和1之间。



**逻辑函数**

现在我们知道了逻辑函数是什么，让我们看看它在逻辑回归中是如何使用的.

# 2.逻辑回归中使用的表达

逻辑回归使用方程式来表达，与线性回归非常类似。使用权重或者系数值(称为希腊大写字母Beta)线性组合输入值(x)来预测输出值(y).与线性回归的主要区别在于，模型的输出值是二进制值(0或者1)，而不是连续的数值。

以下是逻辑回归方程示例：

**y = e^(b0 + b1\*x) / (1 + e^(b0 + b1\*x))**

其中，y是模型的预测输出，b0是偏置项或者截距项，b1是输入(x)的系数或者权重参数。输入数据中的每一列都有一个相关的b系数，我们模型要学习的就是该参数。

# 3.逻辑回归预测概率

逻辑回归对默认类或(标签)的概率进行建模。例如，如果我们根据身高来对性别(男性或者女性)进行建模，那么，默认类就可能是男性。那么逻辑回归的模型可以写成给定身高是男性的概率:

**P(X) = P(Y=1|X)**

概率预测必须被转换为二进制值(0或者1)，以便在实际中进行概率预测。

逻辑回归是一种线性方法，但需要使用逻辑函数将预测转换成概率。我们不能再将预测理解为输入的线性组合。模型可以表示为：

**p(X) = e^(b0 + b1\*X) / (1 + e^(b0 + b1\*X))**

将上式进行转换：

**ln(p(X) / 1 – p(X)) = b0 + b1 \* X**

上式中，右边的计算再次变成线性的，左边的输入是默认类概率的对数。左边的这个比率被称为默认类的几率。几率是将事件发生概率和事件不发生概率相除计算而得，例如:0.8/(1-0.8)=4,则几率为4。因此，上述表达式可以写为:

**ln(odds) = b0 + b1 \* X**

因为上面的几率进行了对数转换，所有我们称之为对数几率。

将指数右移并将其写为：

**odds = e^(b0 + b1 \* X)**

上面的这些有助于我们理解模型确实仍然是输入的线性组合，但这种线性组合与默认类的对数几率相关。

# 4.逻辑回归模型的学习

我们需要根据数据来估算逻辑回归算法的系数(Beta值),通常使用最大似然估计来完成。

最大似然估计是各种机器学习算法常用的学习算法，它对数据的分布做出了假设。最佳的系数将导致模型预测接近1的值为男性，接近0的值为女性。逻辑回归的最大似然性是在搜索过程中寻求系数的值，使预测的类与真实的类误差最小。

# 5.使用逻辑回归模型进行预测

使用逻辑回归模型进行预测很简单，只需将输入数据带入逻辑回归方程计算即可。

下面通过一个具体的例子来说明:

假设我们有一个模型可以根据他们的身高来预测一个人是男性还是女性。给定身高为150cm.

我们已经学习到系数b0=-100和b1=0.6.使用上面的等式可以计算出男性身高150cm以上的概率P：

**y = e^(b0 + b1\*X) / (1 + e^(b0 + b1\*X))**

**y = exp(-100 + 0.6\*150) / (1 + EXP(-100 + 0.6\*X))**

**y = 0.0000453978687**

表明该人是男性的概率接近于0.

在实践中我们可以直接这样使用:

**0 if p(male) < 0.5**

**1 if p(male) >= 0.5**

# 6.逻辑回归数据准备

逻辑回归中关于数据的分布和关系的假设与线性回归中的假设大致相同。在预测建模机器学习项目中，我们需要专注于准确的预测，而不是解释结果。因此，只要模型健壮且性能良好，就可以打破一些假设。

* 二进制输出变量：逻辑回归是针对二进制分类问题的。
* 去除噪声：Logistic回归假设输出变量(y)中没有错误，请考虑从训练数据中删除异常值和可能错误分类的实例。
* 高斯分布:Logistic回归是一种线性算法（对输出进行非线性变换）。 它确实假设输入变量与输出之间存在线性关系。 输入变量的数据转换可以更好地暴露这种线性关系，从而可以生成更准确的模型。 例如，您可以使用log，root，Box-Cox和其他单变量转换来更好地表达此关系。
* 删除相关输入：与线性回归一样，如果您有多个高度相关的输入，模型可能会过度拟合。 考虑计算所有输入之间的成对相关性并去除高度相关的输入。
* 未能收敛：学习系数的期望似然估计过程可能无法收敛。 如果数据中有许多高度相关的输入或数据非常稀疏（例如输入数据中有很多零），就会发生这种情况。