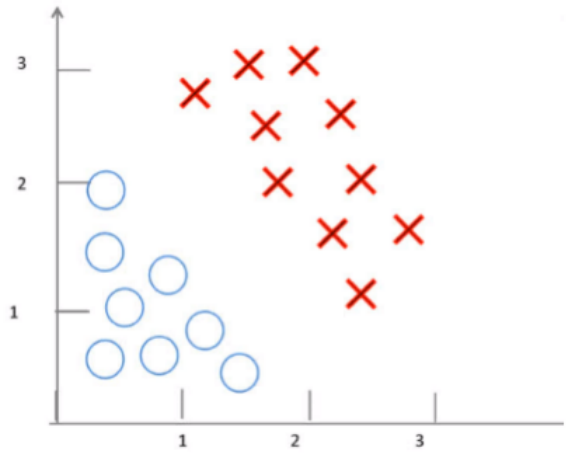
问题引入

实际问题

我们想对两种鸢尾花进行分类，现在我们可以观察到鸢尾花的两个特征：花瓣个数和直径大小。那么我们该如何构建一个自动分类器，让算法可以根据这两个特征识别出当前的鸢尾花属于哪个类别呢？不妨先把已经获得的数据画出来看一下，假设如下如所示



基于线性回归的思考

是否可以训练一个线性回归模型，通过判断输出值的大小来决定是什么类别？

考虑二分类问题，假设我们认为线性回归模型输出值大于等于0.5时为1，小于0.5时为0，因此我们希望存在一个函数 $h_\theta(x)$ ，使得对于线性回归模型的值，能够映射在0-1范围内，表示如下：

$$f(x) = \begin{cases} 1, h_\theta(x) \geqslant 0.5 \\ 0, h_\theta(x) < 0.5 \end{cases}$$

思考一下如果形象的理解（不那么数学的理解）“基于线性回归的思考”这一句话。 refer ipad 1.5 逻辑回归的引入

我们回头看这个公式：

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$$

由上式可得：

$$\ln \frac{h_\theta(x)}{1 - h_\theta(x)} = \theta^T x$$

如果我们把 $h_\theta(x)$ 视为样本x为正例的可能性，那么 $1 - h_\theta(x)$ 即为负例可能性，两者的比值的对数，称为对数几率，这也是为什么逻辑回归也称为对数几率回归。同样，反过来，我们通过对样本为正反例可能性的伯努利分布推导（和指数族定义），也能得到上式的结果，因此，选择**sigmoid**的原因在于他符合描述样本**x**为正反例的可能性，而逻辑回归则是对该可能性建模，求解参数极大似然估计的过程

小总结： sigmoid 引入的其中两点原因：（为了理解，可能还有其他原因，感兴趣的话可以自己去查资料）

1. 损失函数是Non-convex, 求解会比较复杂
2. 可以对可能性建模 （满足值域[0,1]）

损失函数

jupyter lab Logistic_regression.ipynb