

6.1 分类问题

参考文档: 6 - 1 - Classification (8 min).mkv

在分类问题中，你要预测的变量 y 是离散的值，我们将学习一种叫做逻辑回归 (Logistic Regression) 的算法，这是目前最流行使用最广泛的一种学习算法。

在分类问题中，我们尝试预测的是结果是否属于某一个类（例如正确或错误）。分类问题的例子有：判断一封电子邮件是否是垃圾邮件；判断一次金融交易是否是欺诈；之前我们也谈到了肿瘤分类问题的例子，区别一个肿瘤是恶性的还是良性的。

Classification

- Email: Spam / Not Spam?
- Online Transactions: Fraudulent (Yes / No)?
- Tumor: Malignant / Benign ?

我们从二元的分类问题开始讨论。

我们将因变量(dependent variable)可能属于的两个类分别称为负向类 (negative class) 和正向类 (positive class)，则因变量 $y \in \{0, 1\}$ ，其中 0 表示负向类，1 表示正向类。

The diagram shows two mathematical expressions. The first, 'Classification: $y = 0 \text{ or } 1$ ', has blue arrows pointing up to the '0' and '1' to indicate they are discrete values. The second, ' $h_{\theta}(x)$ can be > 1 or < 0 ', has blue arrows pointing up to ' > 1 ' and down to ' < 0 ' to indicate values outside the [0, 1] range. The third, 'Logistic Regression: $0 \leq h_{\theta}(x) \leq 1$ ', has a blue arrow pointing up to the entire inequality to indicate the output is constrained to the [0, 1] range.

$$\text{Classification: } y = 0 \text{ or } 1$$
$$h_{\theta}(x) \text{ can be } > 1 \text{ or } < 0$$
$$\text{Logistic Regression: } 0 \leq h_{\theta}(x) \leq 1$$

如果我们要用线性回归算法来解决一个分类问题，对于分类， y 取值为 0 或者 1，但如果你使用的是线性回归，那么假设函数的输出值可能远大于 1，或者远小于 0，即使所有训练样本的标签 y 都等于 0 或 1。尽管我们知道标签应该取值 0 或者 1，但是如果算法得到的值远大于 1 或者远小于 0 的话，就会感觉很奇怪。所以我们在接下来的要研究的算法就叫做逻辑回归算法，这个算法的性质是：它的输出值永远在 0 到 1 之间。

顺便说一下，逻辑回归算法是分类算法，我们将它作为分类算法使用。有时候可能因为这个算法的名字中出现了“回归”使你感到困惑，但逻辑回归算法实际上是一种分类算法，它适用于标签 y 取值离散的情况，如：1001。

6.2 假说表示

参考视频: 6 - 2 - Hypothesis Representation (7 min).mkv

此前我们说过，希望我们的分类器的输出值在 0 和 1 之间，因此，我们希望想出一个满足某个性质的假设函数，这个性质是它的预测值要在 0 和 1 之间。

我们引入一个新的模型，逻辑回归，该模型的输出变量范围始终在 0 和 1 之间。逻辑回归模型的假设是： $h_{\theta}(x) = g(\theta^T X)$

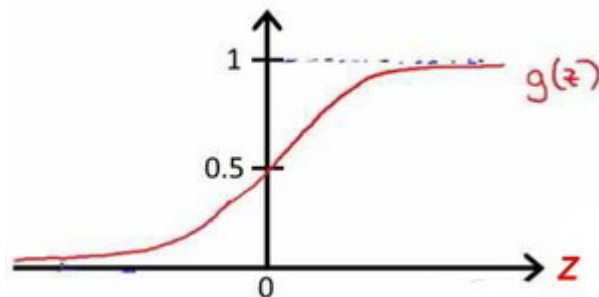
其中： X 代表特征向量 g 代表逻辑函数 (logistic function) 是一个常用的逻辑函数为 S 形函数 (Sigmoid function)，公式为： $g(z) = \frac{1}{1+e^{-z}}$ 。

python 代码实现：

```
import numpy as np

def sigmoid(z):
    return 1 / (1 + np.exp(-z))
```

该函数的图像为：



合起来，我们得到逻辑回归模型的假设：

对模型的理解： $g(z) = \frac{1}{1+e^{-z}}$ 。

$h_{\theta}(x)$ 的作用是，对于给定的输入变量，根据选择的参数计算输出变量=1 的可能性 (estimated probability) 即 $h_{\theta}(x) = P(y = 1|x; \theta)$

例如，如果对于给定的 x ，通过已经确定的参数计算得出 $h_{\theta}(x) = 0.7$ ，则表示有 70% 的几率 y 为正向类，相应地 y 为负向类的几率为 $1-0.7=0.3$ 。

6.3 判定边界

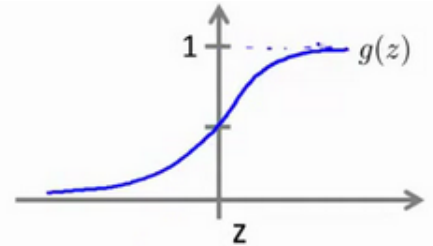
参考视频: 6 - 3 - Decision Boundary (15 min).mkv

现在讲下决策边界(decision boundary)的概念。这个概念能更好地帮助我们理解逻辑回归的假设函数在计算什么。

Logistic regression

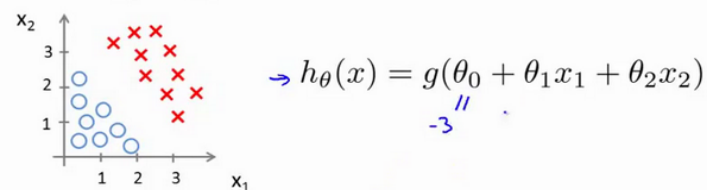
$$\rightarrow h_{\theta}(x) = g(\theta^T x)$$

$$\rightarrow g(z) = \frac{1}{1+e^{-z}}$$

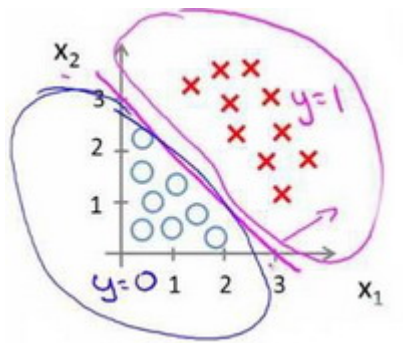


现在假设我们有一个模型：

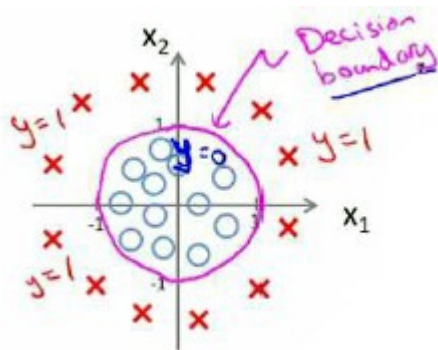
Decision Boundary



并且参数 θ 是向量 $[-3 \ 1 \ 1]$ 。则当 $-3 + x_1 + x_2 \geq 0$ ，即 $x_1 + x_2 \geq 3$ 时，模型将预测 $y = 1$ 。我们可以绘制直线 $x_1 + x_2 = 3$ ，这条线便是我们模型的分界线，将预测为 1 的区域和预测为 0 的区域分隔开。



假使我们的数据呈现这样的分布情况，怎样的模型才能适合呢？



因为需要用曲线才能分隔 $y = 0$ 的区域和 $y = 1$ 的区域，我们需要二次方特征：
 $h_{\theta}(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2)$ 是 $[-1 \ 0 \ 0 \ 1 \ 1]$ ，则我们得到的判定边界恰好是圆
 点在原点且半径为 1 的圆形。

我们可以用非常复杂的模型来适应非常复杂形状的判定边界。

