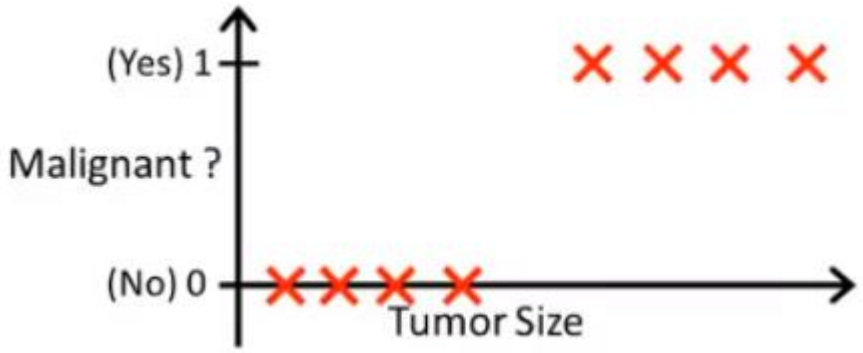
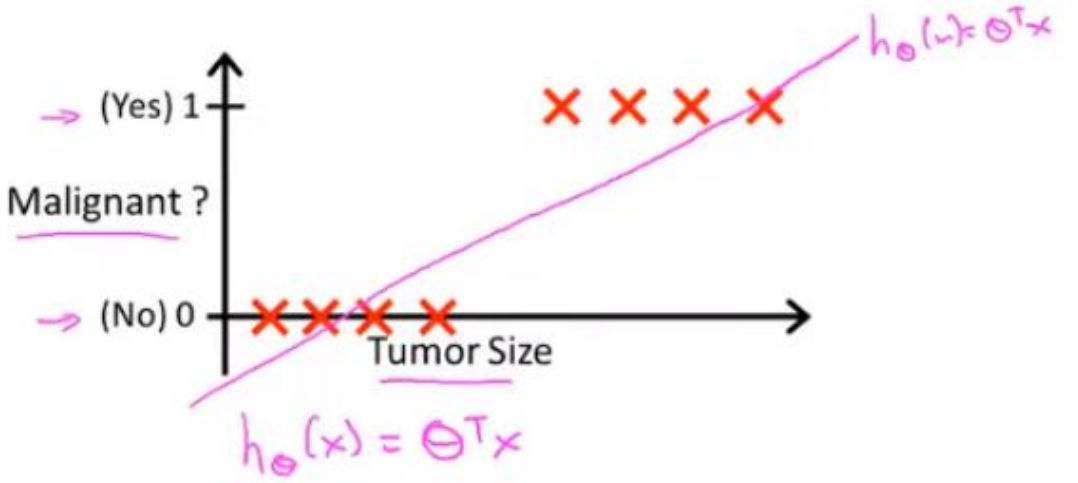
1、分类学习算法概述

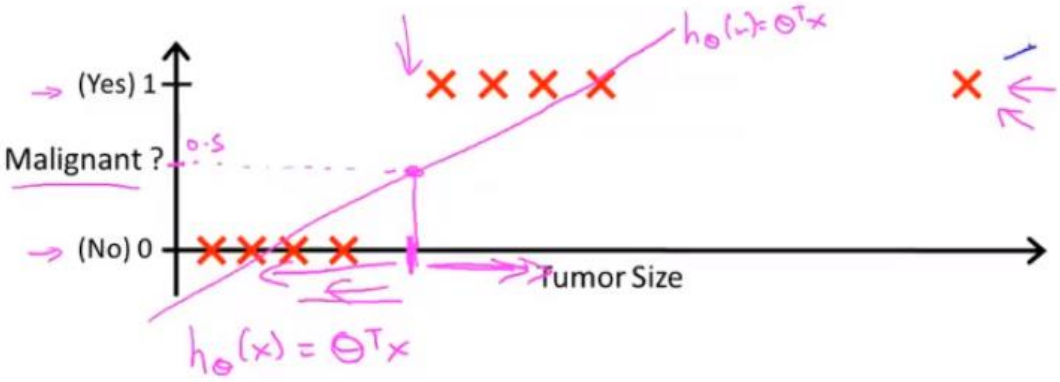
基于监督学习的分类算法，前面已经提到过，分类算法实质上是给定一些样本，并提前告知你哪些属于这一类，哪些属于那一类，为简单处理，我们这里仅考虑输出结果为两种情况，分别表示 0、1。举一个例子，根据肿瘤尺寸判断是阴性还是阳性，这是给定了的数据，可以看到，取值离散，且明显有区分。



如果任像之前的线性回归算法，拟合出一条曲线来，那么计算机很有可能拟出这样一条来，看似结果还算正确，给定一个阈值，输出结果大于 0.5 为阳性，小于 0.5 为



阴性，如果训练集是这样的又如何处理呢？可以看到，远处离散的一个点，使得整个

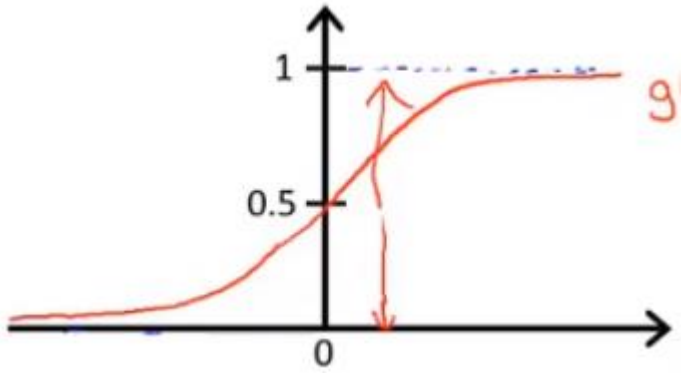


拟合出来的曲线有偏差。这里需说明的一点是，由于回归拟合中，是保证代价差总体保证较小，因此在数据密集且连续的情况下，少量点不再拟合线上，一般也不认为拟合的不好。而在分类问题中，由于数据离散，因此少量的点被错误划归分组也很难让人觉得拟合得较好。另外，由于取值离散，这里仅为 0，1 那些远大于 1 和远小于 0 的点该作何处理？总觉得用用回归的思想去分类数据，是在用大炮打蚊子。

2、logistic 分类

是时候引入 logistic 分类算法了，这一算法保证预测出来的结果总是一边趋于 0，一边趋于 1，在线性回归中，我们用$h_\theta(x)=\theta^T x$来预测结果的，然而之前讲了这并不适用。因此正式引入 logistic 分类预测模型，形如:

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



作出来的图形如上所示，它的取值总是趋于 0~1，当:

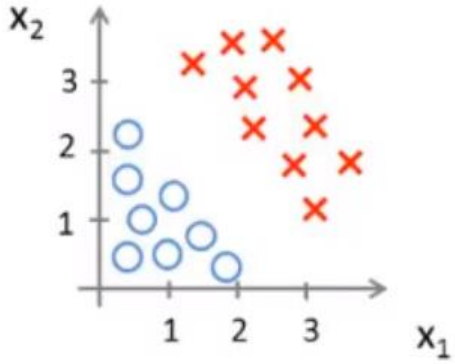
$-\theta^T x = 0$，$h_\theta(x)=0.5$ ——> $\theta^T x=0$，任意

$-\theta^T x > 0$，$h_\theta(x)<0.5$ ——> $\theta^T x<0$，输出 0

$-\theta^T x < 0$，$h_\theta(x)>0.5$ ——> $\theta^T x>0$，输出 1

可见，这样一个预测函数作为分类问题算法似乎是可行的，所有预测结果都将取值为 0 或 1，如果θ值取值合理的话，那么预测模型堪称完美!

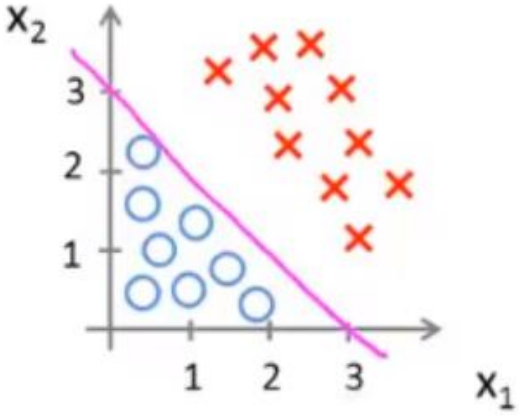
3、模型预测



$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$

假设我们有如上所示的数据集，要进行分类，首先是得选择好的预测函数，之前讲到用 logistic 模型预测效果会好些，只是关于参数的求解还没进行介绍，这里假设已经得到能够较好预测模型的参数值：$\theta_0 = -3$，$\theta_1 = 1$，$\theta_2 = 1$。

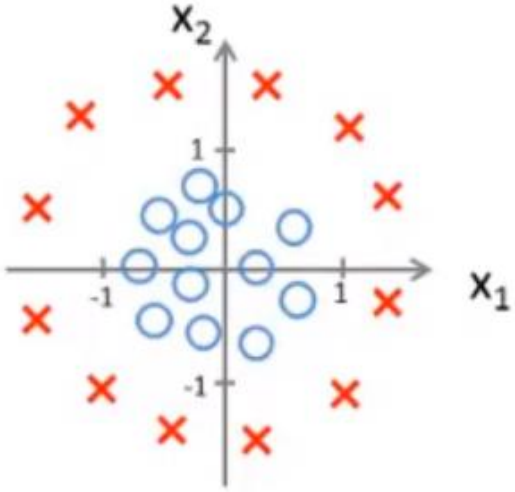
由于已经给出结论，所以这里直接给出决策边界的定义，$-3 + x_1 + x_2 = 0$，画出来的图形是这样的，可以明显看出，决策边界已经将图形划分为了两个组别。因此，

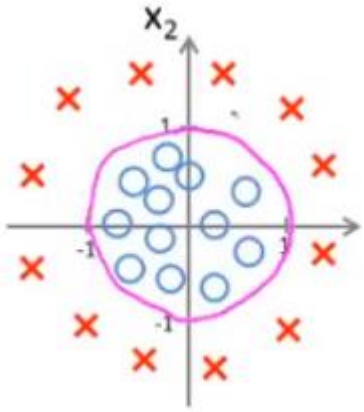


我们可以说，用 $h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$模型很好的对训练数据进行了分类。

关于决策边界，它是预测模型中固有属性，与训练集不直接关联，一旦算出了θ，那么由 $\theta^T x$确定的决策边界也就确定了。

还有更复杂一些的模型，下方这幅图得出的$\theta = [-1, 0, 0, 1, 1]$，决策边界为$\theta^T x = -1 + x_1^2 + x_2^2$，因此得到的分组曲线为：





好了，关于用什么模型去预测，以及 logistic 算法原理就讲这么多了，接下来，就是关于参数θ的计算了。