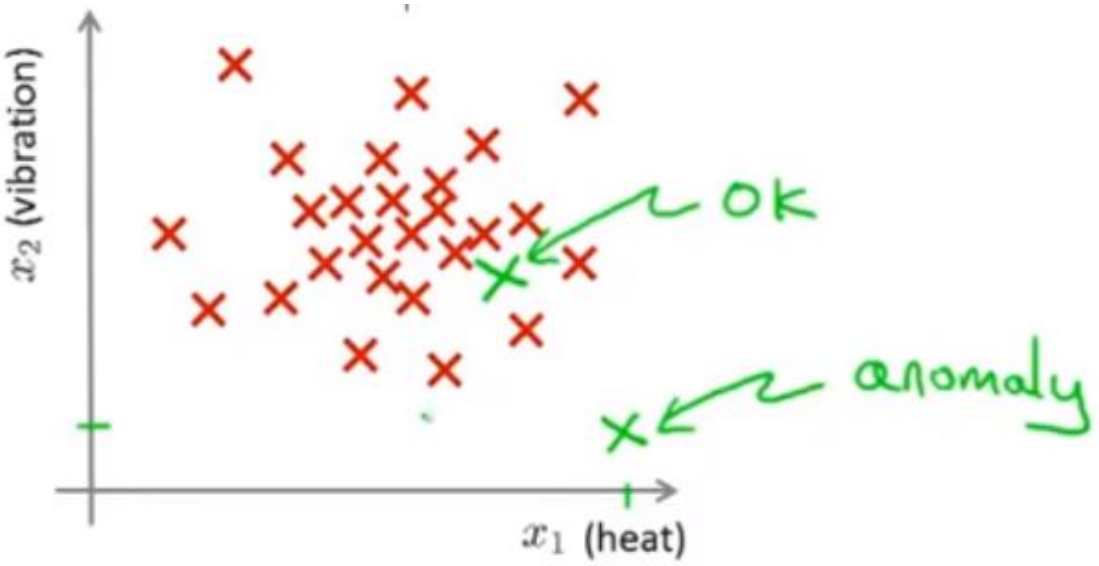
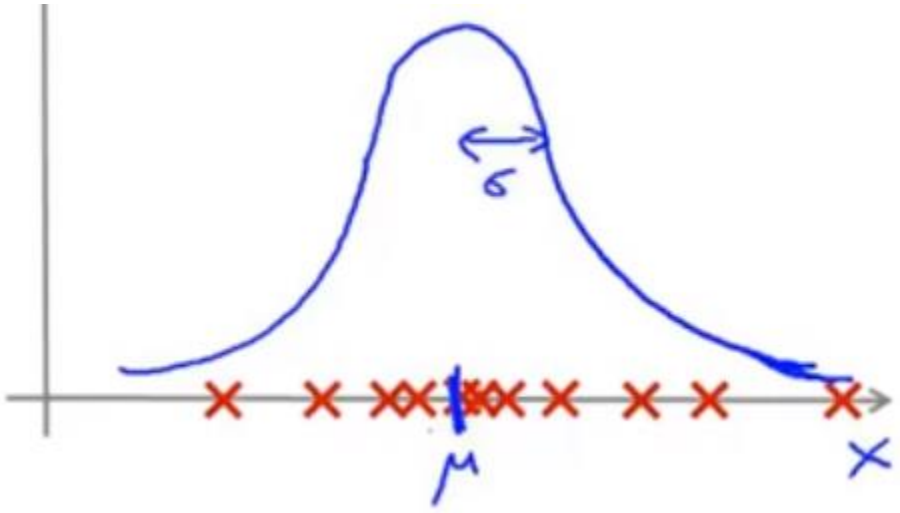
1、异常说明



上方所示为飞机引擎训练集分布情况示意图，可以看到，大部分数据集中在中间那片区域，现要测试一个新的样本是否为异常样本，是否需要进一步检测。右下角那个点明显偏离正常范围，需要进一步检测。那么，如何用数学上的计算得到异常值呢。

2、高斯分布（正态分布）

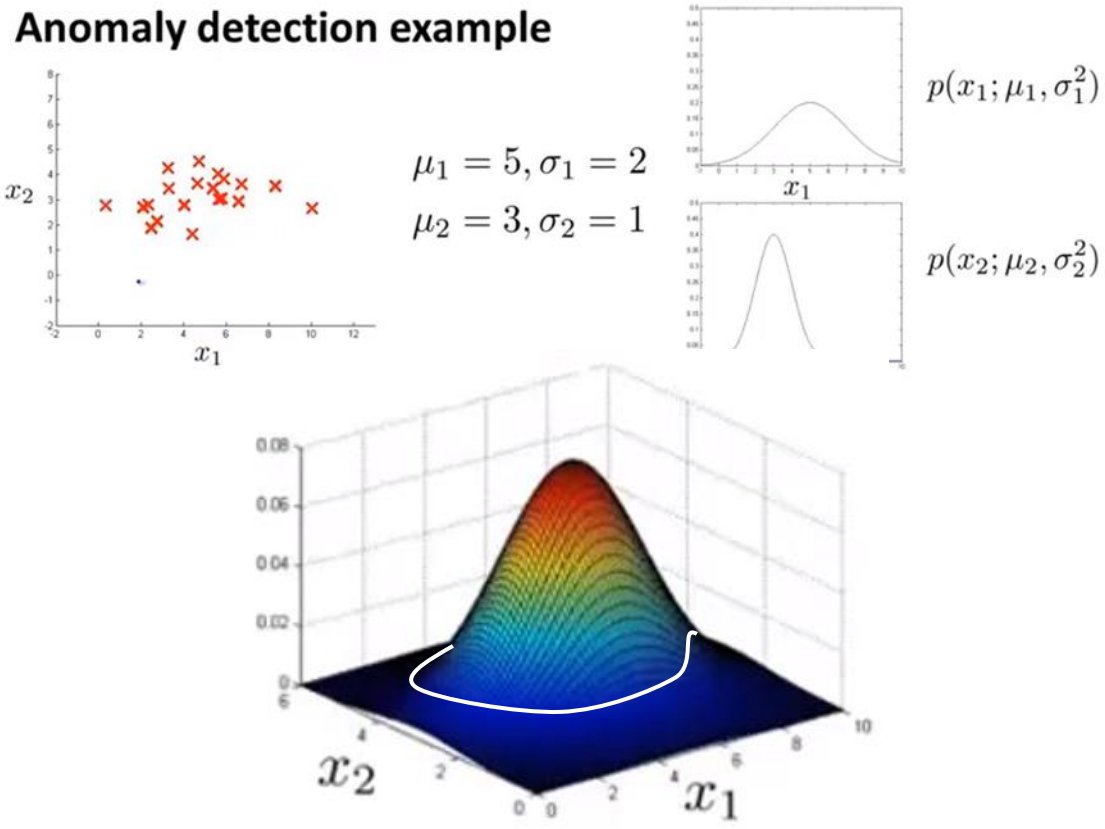


如图所示为正态分布曲线示意图，横轴代表某一特征值分布区间，纵轴代表概率曲线，可以通过极大似然估计的思想得出$X^{(i)} \sim N(\mu, \sigma^2)$中$u$和$\sigma$的值，计算公式为：

$$u = \frac{1}{m}\sum_{i=1}^{n} X^{(i)} \qquad \sigma^2 = \frac{1}{m}\sum_{i=1}^{m}(x^{(i)} - u)2$$

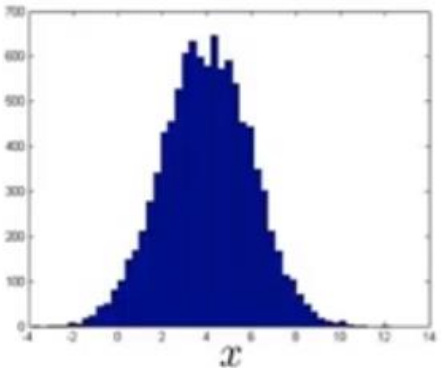
接下来的判别的方法就很简单了，由于每个样本有多个特征值，所以计算出多个u和σ的值，然后将待检测样本的各个特征值代入高斯计算公式，得出多个概率值，最后进行连乘，如果该值<=阈值ε，则认为该值为异常值，换句话说，判断样本是否为异常样本，即估计它的各个特征值在正态分布中的概率，最后算总的概率。

举个例子，左边为有两个特征值的样本分布，计算每个特征值的高斯分布，然后使用概率计算公式可以得到下方所示的概率分布曲面，可以认为规定一个临界线，比如低于下方白色线条，可认为总概率值过低，为异常样本。



3、异常检测 VS 监督学习

有的示例似乎让人觉得它和监督分类学习很相似，但这里给一个通用的抉择条件，当异常样本很多时，可以直接选择监督分类算法，如果异常样本没有那么多时，这时采用异常检测算法。



比如，上方绘制出的数据看似服从高斯分布，因此可以采用异常检测算法。