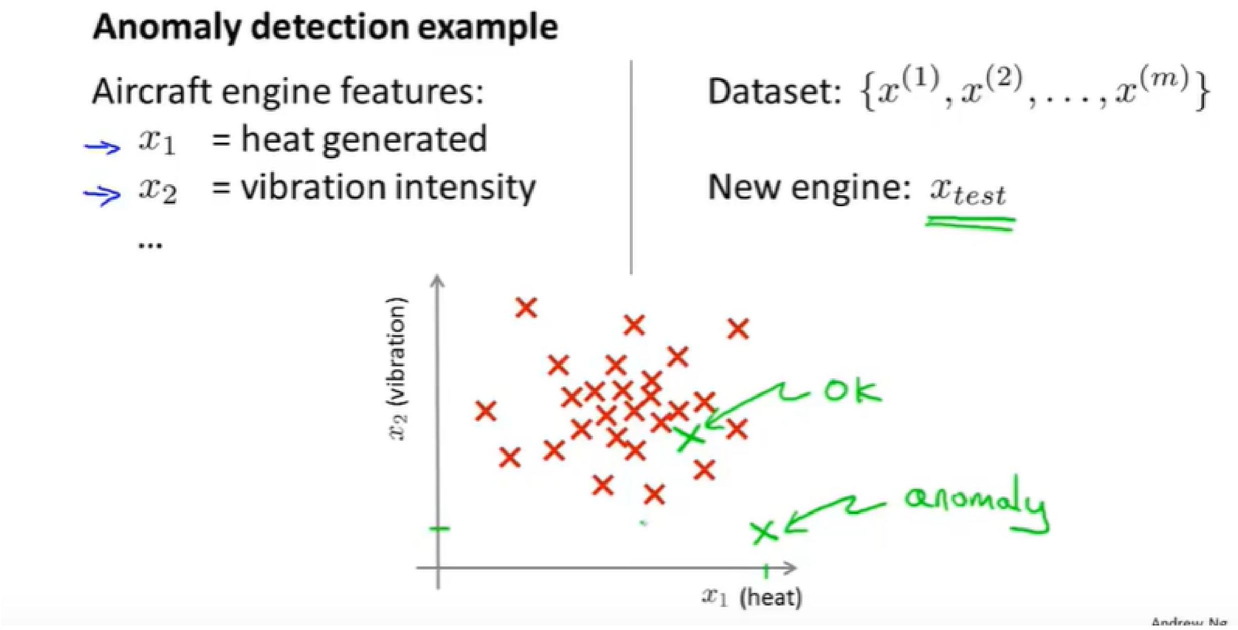
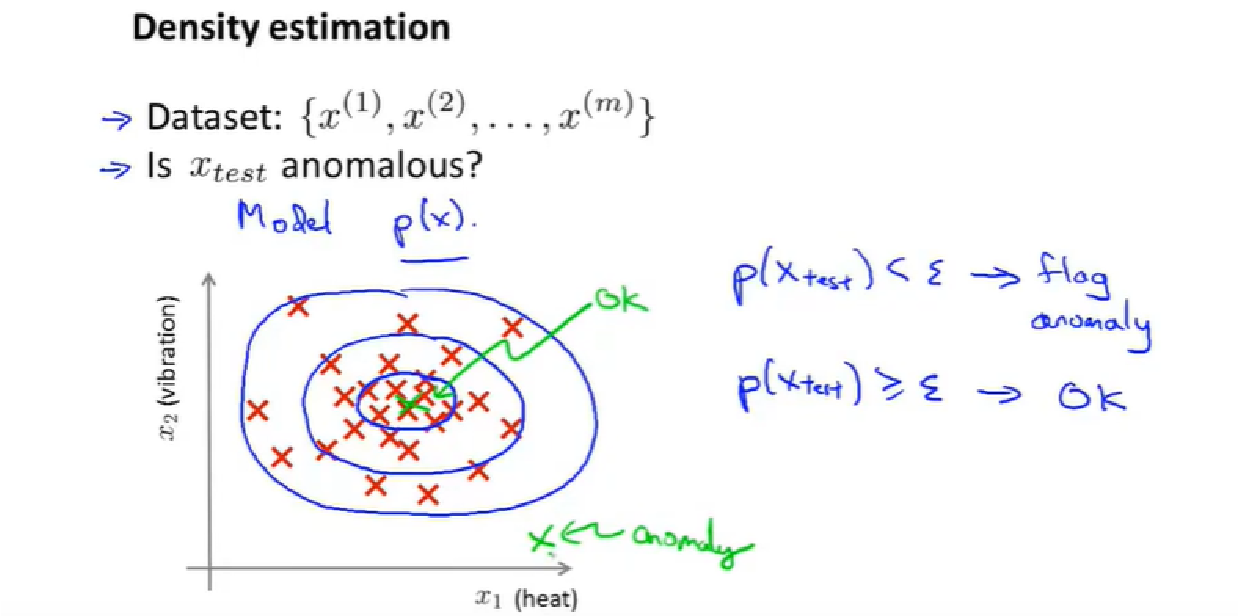
问题动机



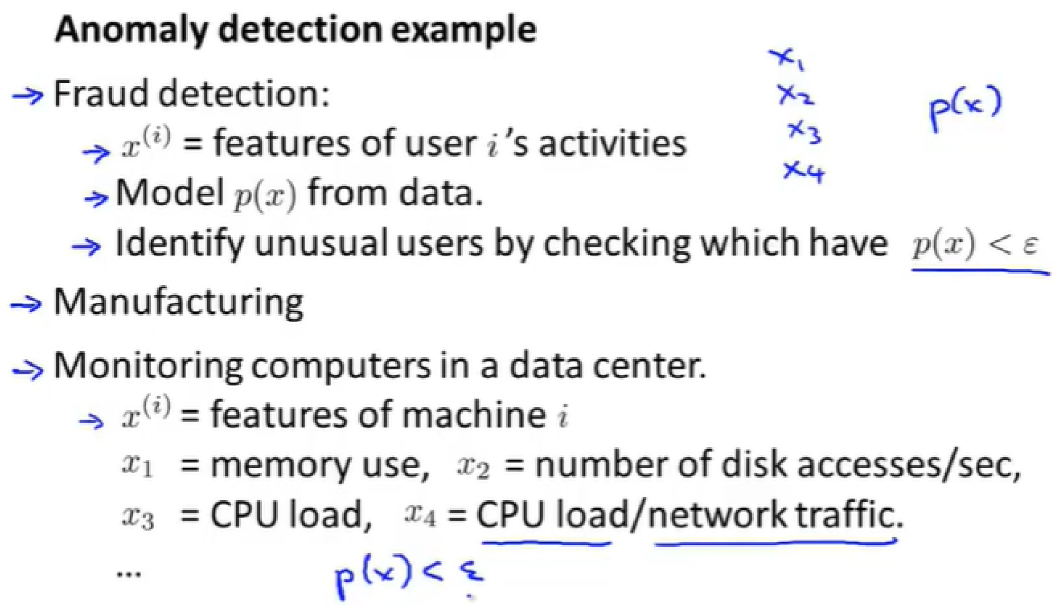
飞机引擎例子中，和其他正常的引擎看起来不同的，被视为异常引擎。



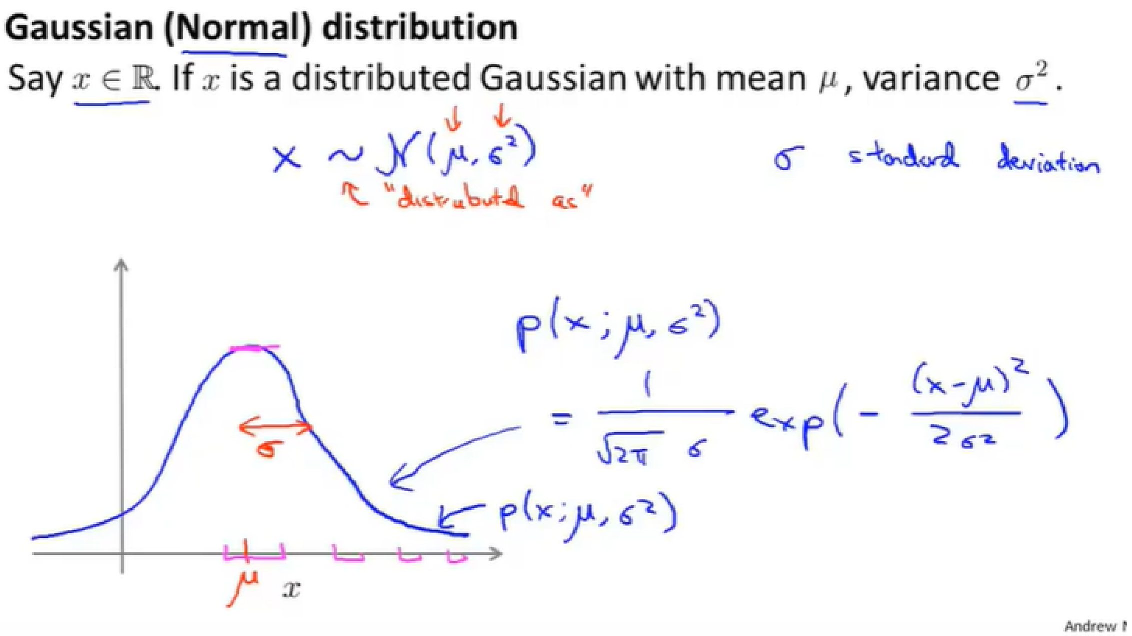
对数据样本建模，计算一个概率并规定阈值，最终蓝色圈中的点，我们将其认为是正常引擎，而之外的被认为是异常。

异常检测案例

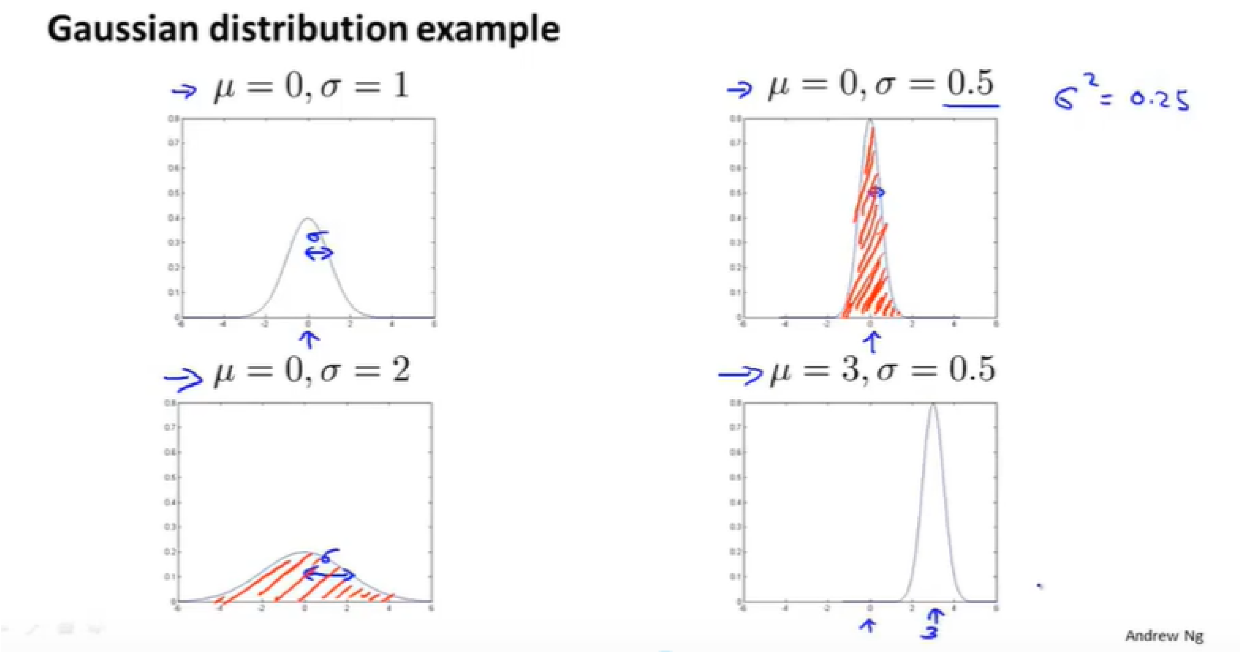
欺诈检测，生产行业，监控计算机



高斯分布(正态分布)

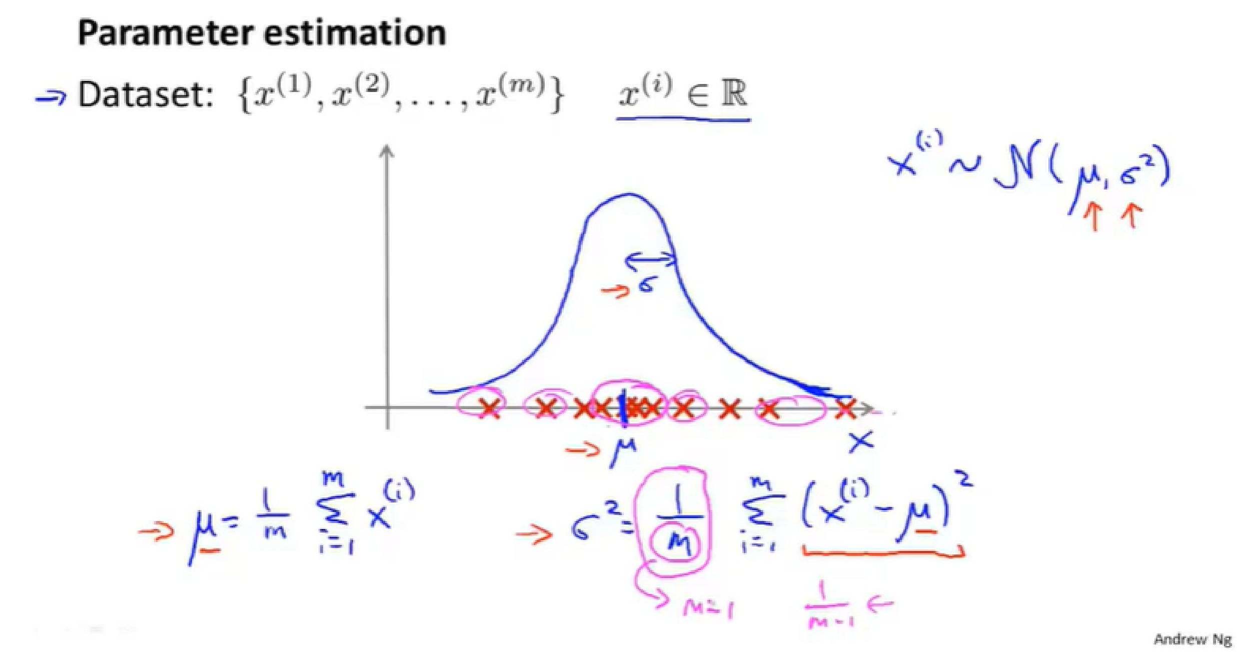


高斯分布中，x是变量，~表示服从什么分布，N表示正态分布，μ是均值，σ^2是方差，σ是标准差，μ决定了钟形的中心位置，σ决定了钟的宽度。



例子中阴影部分面积和为1.

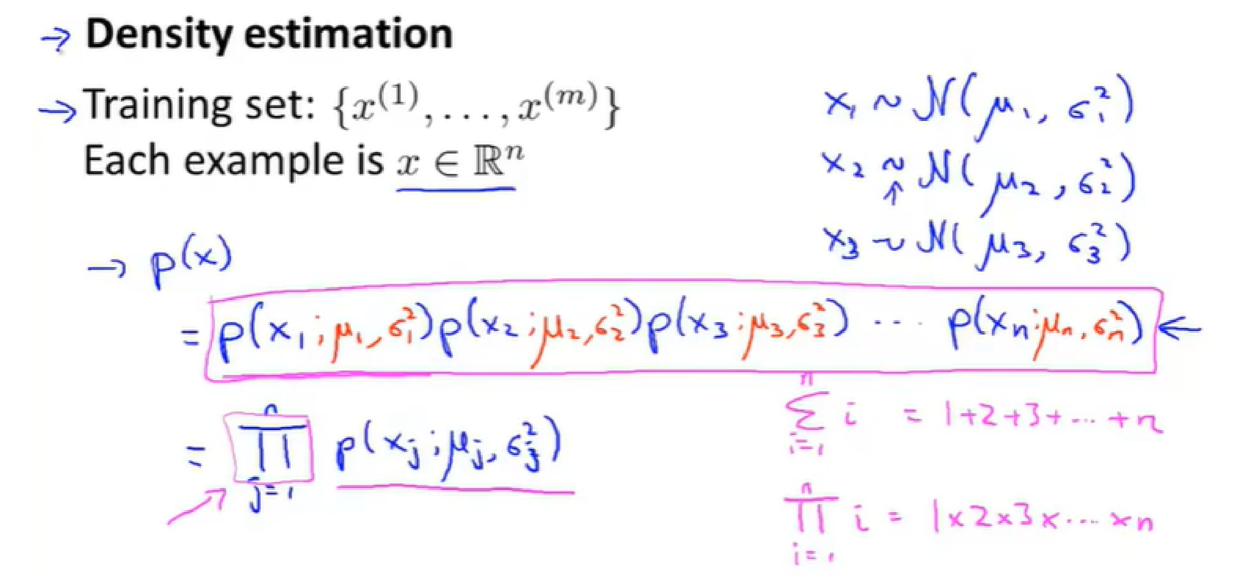
参数估计



μ，σ^2的公式如上，有的版本σ^2公式中的分母写作m-1，其实几乎没有差别。

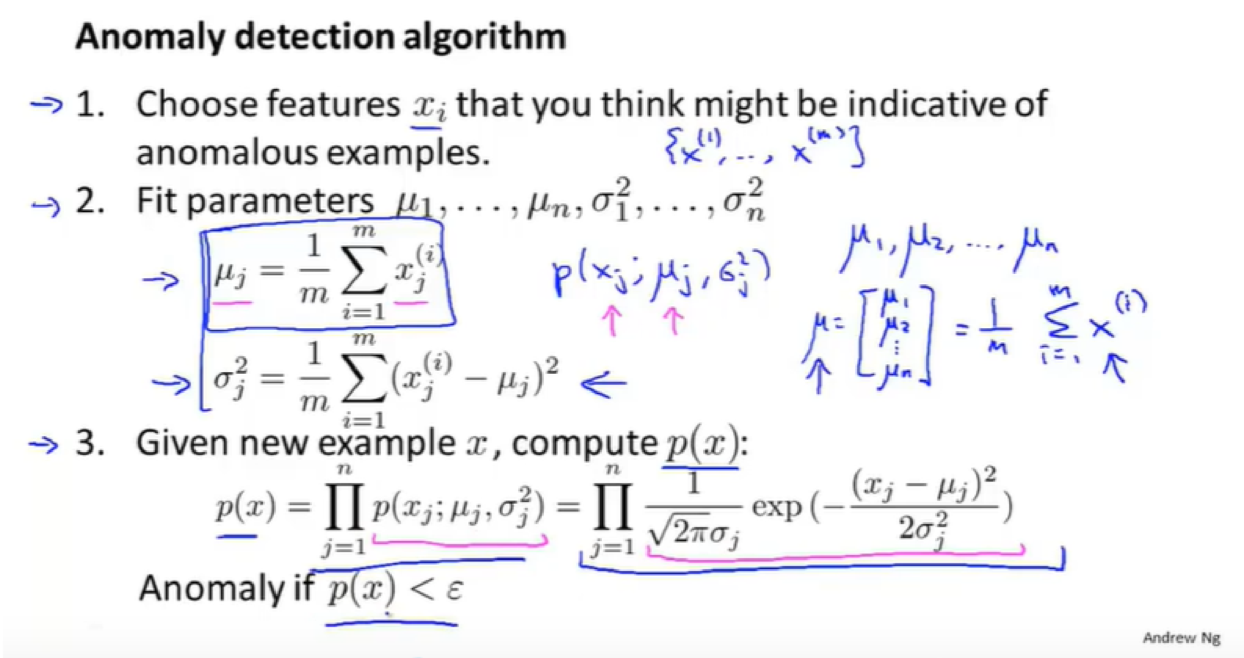
高斯分布构造异常检测算法

密度估计：



密度估计就是将一个数据样本中一个数据中的各个维度的值用高斯分布计算得出的概率进行连乘。

异常检测算法：

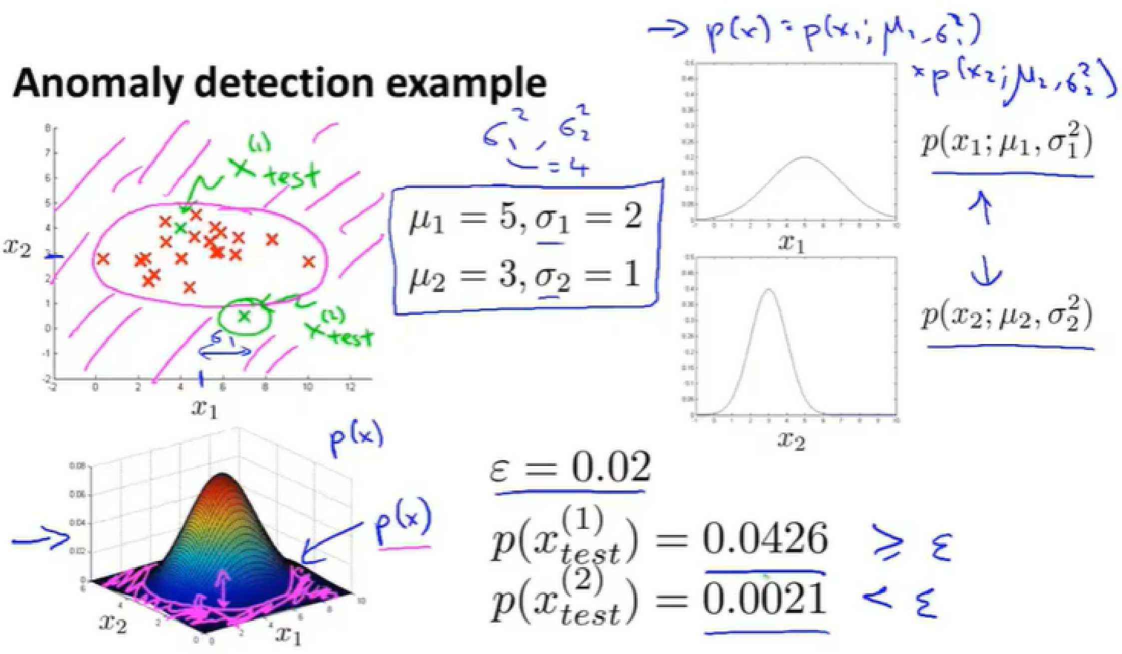


1.选择特征

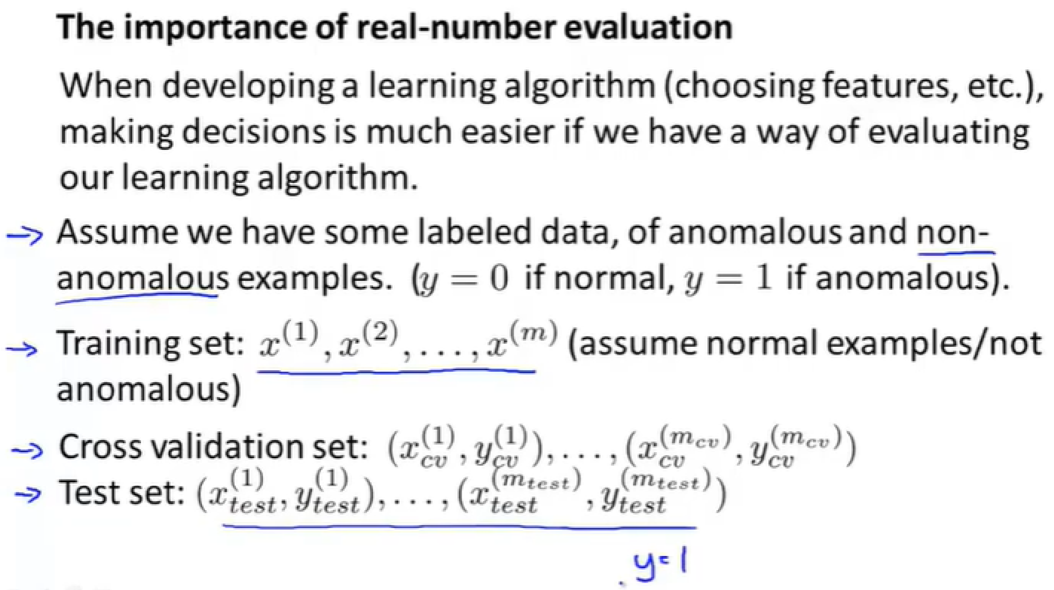
2.拟合参数，计算μ，σ^2

3.密度估计，当计算出的概率低于阈值，视为异常

例子：



开发和评估异常检测系统



y=0表示飞机引擎正常，y=1代表飞机引擎异常。

实数评估：

一个评估学习算法的数字来告诉我们算法的质量，我们可以根据这个实数来调整加入的特征。

例子：

Aircraft engines motivating example

→ $\boxed{10000}$ good (normal) engines

→ $\boxed{20}$ flawed engines (anomalous) $\underline{2-50}$ $\underline{y=1}$

$\mu_1, \sigma_1^2, \ldots, \mu_n, \sigma_n^2$

→ Training set: $\boxed{6000}$ good engines $(y=0)$ $p(x) = p(x_1;\mu_1,\sigma_1^2) \cdots p(x_n;\mu_n,\sigma_n^2)$

CV: $\boxed{2000}$ good engines ($y = 0$), $\boxed{10}$ anomalous ($y = 1$)

Test: $\boxed{2000}$ good engines ($y = 0$), $\boxed{10}$ anomalous ($y = 1$)

训练集中只有正常的引擎。

同样的数据（包括正常，异常引擎）不要同时用于交叉训练集和验证集。

算法评估：

Algorithm evaluation

→ Fit model $p(x)$ on training set $\{x^{(1)}, \ldots, x^{(m)}\}$

→ On a cross validation/test example x, predict

$$y = \begin{cases} 1 & \text{if } p(x) < \varepsilon \text{ (anomaly)} \\ 0 & \text{if } p(x) \geq \varepsilon \text{ (normal)} \end{cases}$$

Possible evaluation metrics:

- True positive, false positive, false negative, true negative
- Precision/Recall
- F_1-score

Can also use cross validation set to choose parameter ε

好的评价指标不能只有单纯的分类正确率。

一个选择阈值epsilon ε的方式是在交叉验证集中使用不同的值，从中选择一个能最大化F1_score或者其他指标的。最后在测试集中进行最后评估。