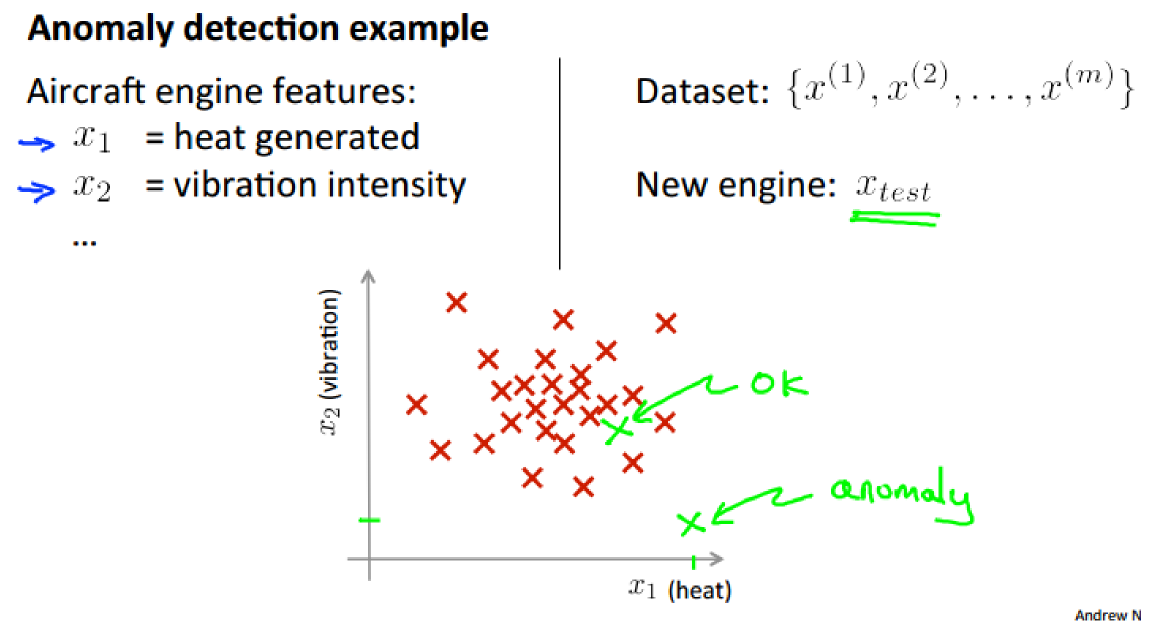
Anomaly Detection

Problem motivation:

首先描写叙述异常检测的样例：飞机发动机异常检测



Anomaly detection example

Aircraft engine features:

→ x_1 = heat generated

→ x_2 = vibration intensity

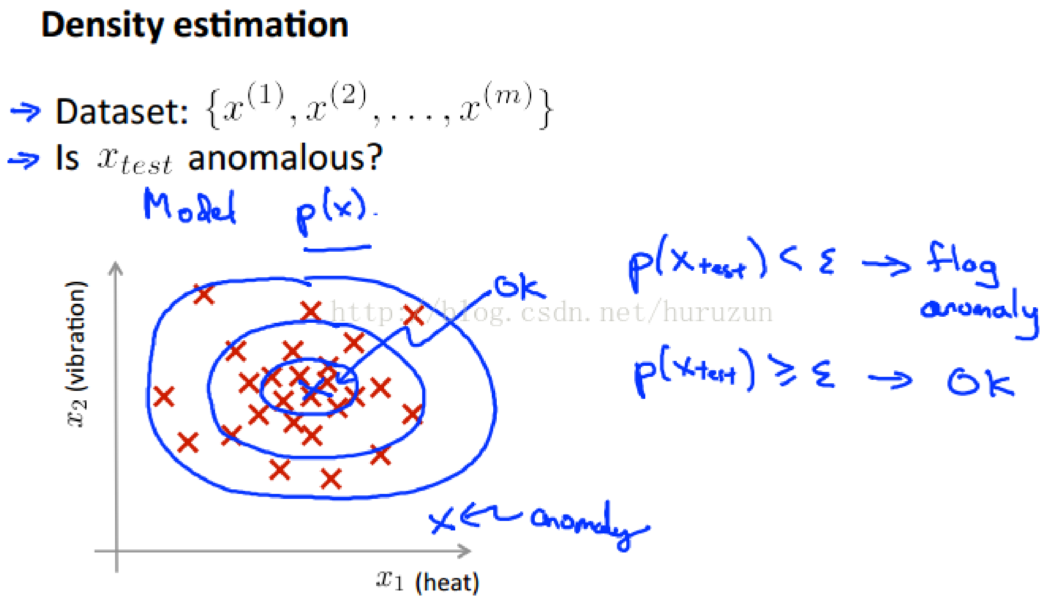
...

Dataset: $\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$

New engine: x_{test}

直观上发现，假设新的发动机在中间，我们非常大可能觉得是OK的。假设偏离非常大。我们就须要很多其它检测确定是否为正常发动机。

以下进行数学形式上的描写叙述，通过概率密度进行预计。例如以下图：



Density estimation

→ Dataset: $\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$

→ Is x_{test} anomalous?

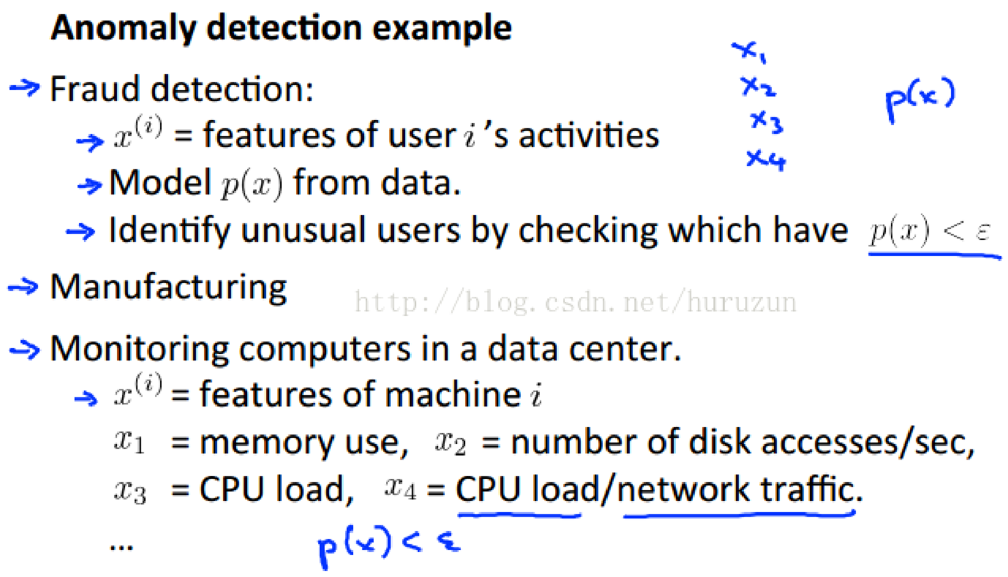
Model $p(x)$.

$p(x_{test}) < \varepsilon$ → flag anomaly

$p(x_{test}) \geq \varepsilon$ → OK

对正常的数据进行建模。求X$_{test}$的概率。当处于中心位置时概率比較大。而且大于设定的阈值，我们判定为OK状态，在远离中心状态。概率比較小，小于设定阈值我们判定为anomaly点。

Anomaly detection常见应用：

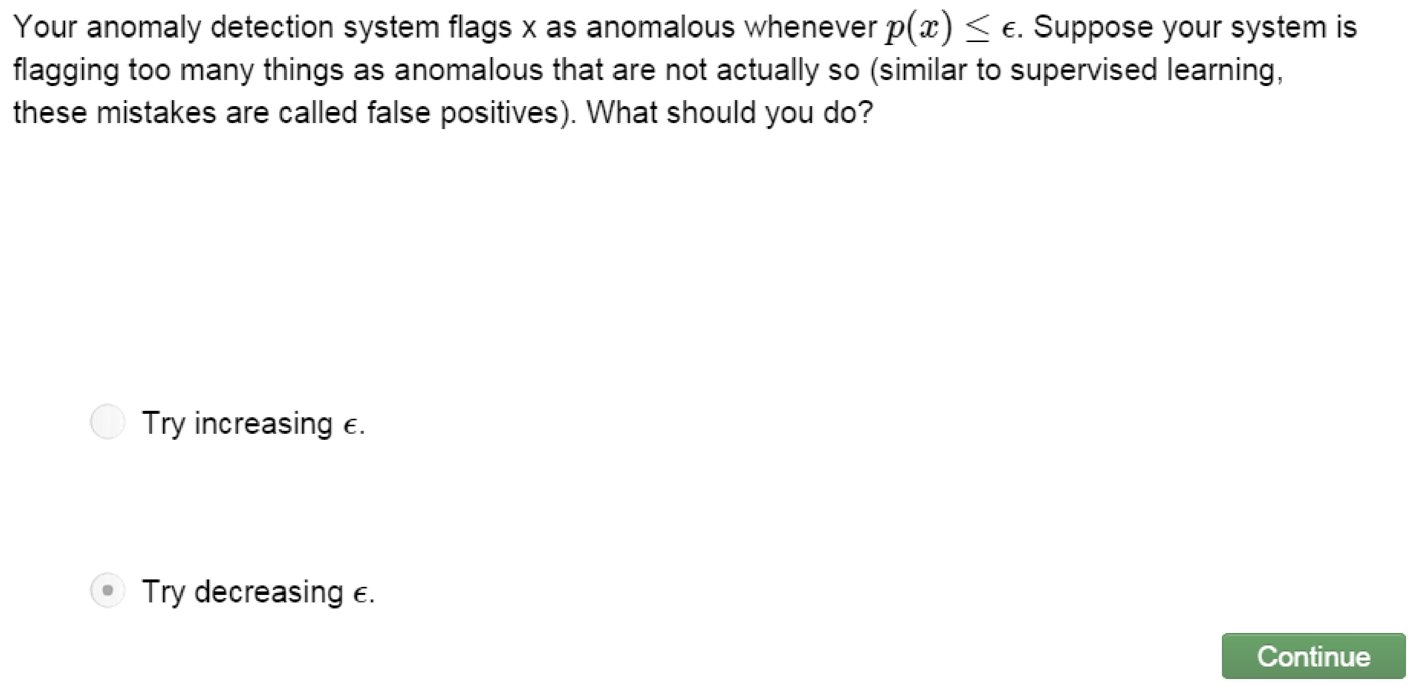


Anomaly detection example

- Fraud detection:
  - $x^{(i)}$ = features of user i 's activities
  - Model $p(x)$ from data.
  - Identify unusual users by checking which have $\underline{p(x) < \varepsilon}$
- Manufacturing
- Monitoring computers in a data center.
  - $x^{(i)}$ = features of machine i
  - x_1 = memory use, x_2 = number of disk accesses/sec,
  - x_3 = CPU load, x_4 = $\underline{\text{CPU load}}/\underline{\text{network traffic}}$.
  - …

NG课上提到了三个应用方向，第一个是最開始举例的飞机引擎。然后是欺诈发现，这个在信用卡和购物站点上得到广泛应用。最后一个是产业界应用，我们须要监视一个计算机系统。我们通过正常执行系统的 内存使用、CUP load等建模。当系统某个值不在正常范围就可以能是计算机系统中有电脑出现异常状态。

习题：当我们系统建模后，导致把异常状态推断为正常状态，这时须要减少阈值避免误判。



Your anomaly detection system flags x as anomalous whenever $p(x) \leq \epsilon$. Suppose your system is flagging too many things as anomalous that are not actually so (similar to supervised learning, these mistakes are called false positives). What should you do?

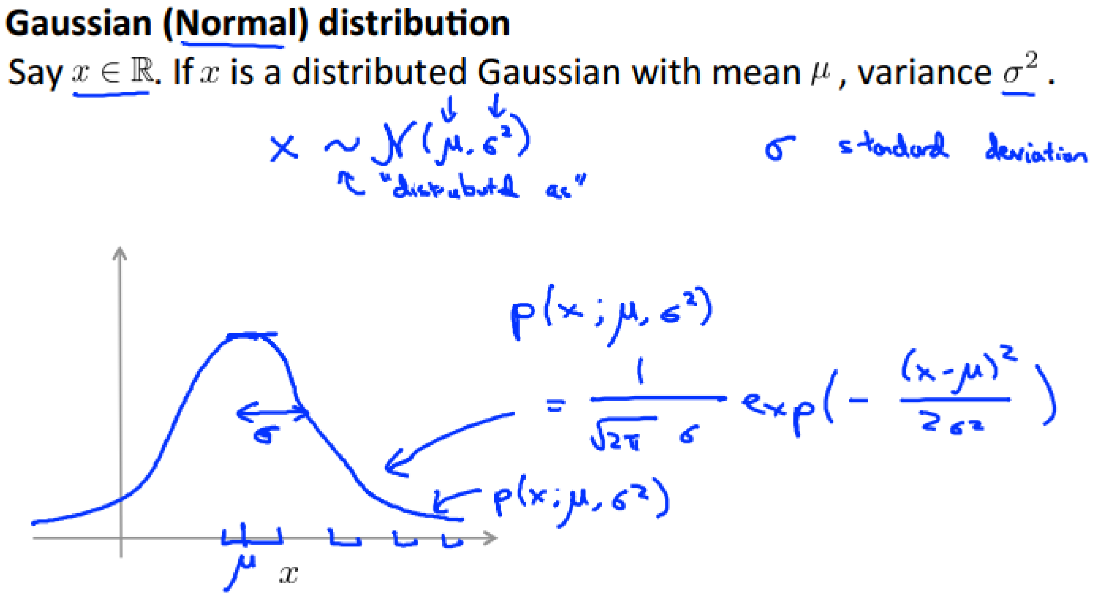
○ Try increasing ϵ.

● Try decreasing ϵ.

Correct!

GaussianDistribution：

复习高斯分布一些内容，比較熟悉能够直接跳过。



Gaussian (Normal) distribution

Say $x \in \mathbb{R}$. If x is a distributed Gaussian with mean μ, variance σ^2.

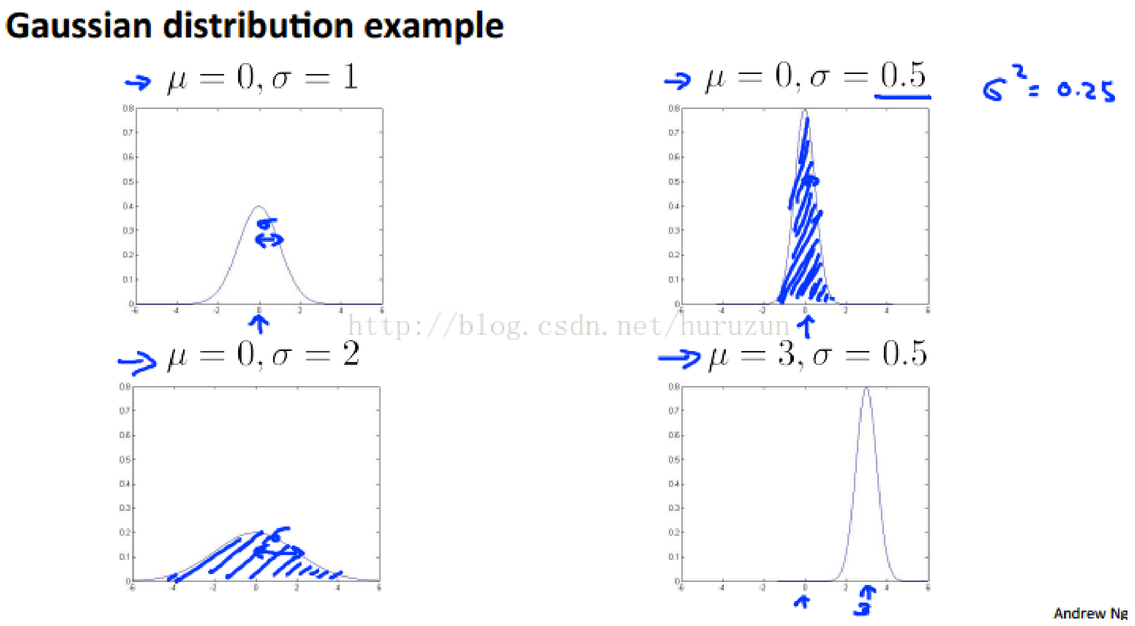
$x \sim \mathcal{N}(\mu, \sigma^2)$

"distributed as"

σ standard deviation

$$p(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

图型和概率分布函数。



Gaussian distribution example

$\mu = 0, \sigma = 1$

$\mu = 0, \sigma = 0.5$ $\sigma^2 = 0.25$

$\mu = 0, \sigma = 2$

$\mu = 3, \sigma = 0.5$

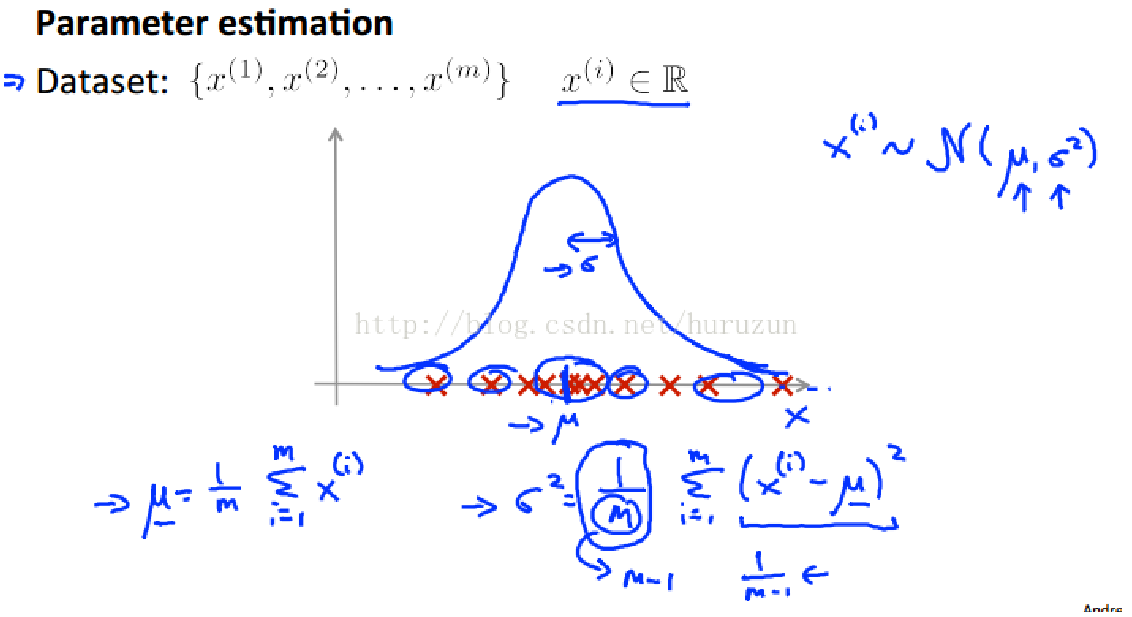
Andrew Ng

上图均值方差表如今高斯分布图型上的差异。

Parameter estimation：

简单的说就是预计均值和方差。下图中写出的公式事实上能够通过极大似然预计进行数学上的求解证明。这里就不具体说（翻开数理统计课本能够找到）。求方差公式中能够选择m或者m-1这都无所谓，由于往往数据集非常大。这样最后计算结果没什么差别，在机器学习中通常选择m而在统计学中往往选择m-1。

选择m还是m-1在理论上有非常大差别，可是实际应用上没什么太大差别。



Parameter estimation

Dataset: $\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$ $\underline{x^{(i)} \in \mathbb{R}}$

$x^{(i)} \sim \mathcal{N}(\mu, \sigma^2)$

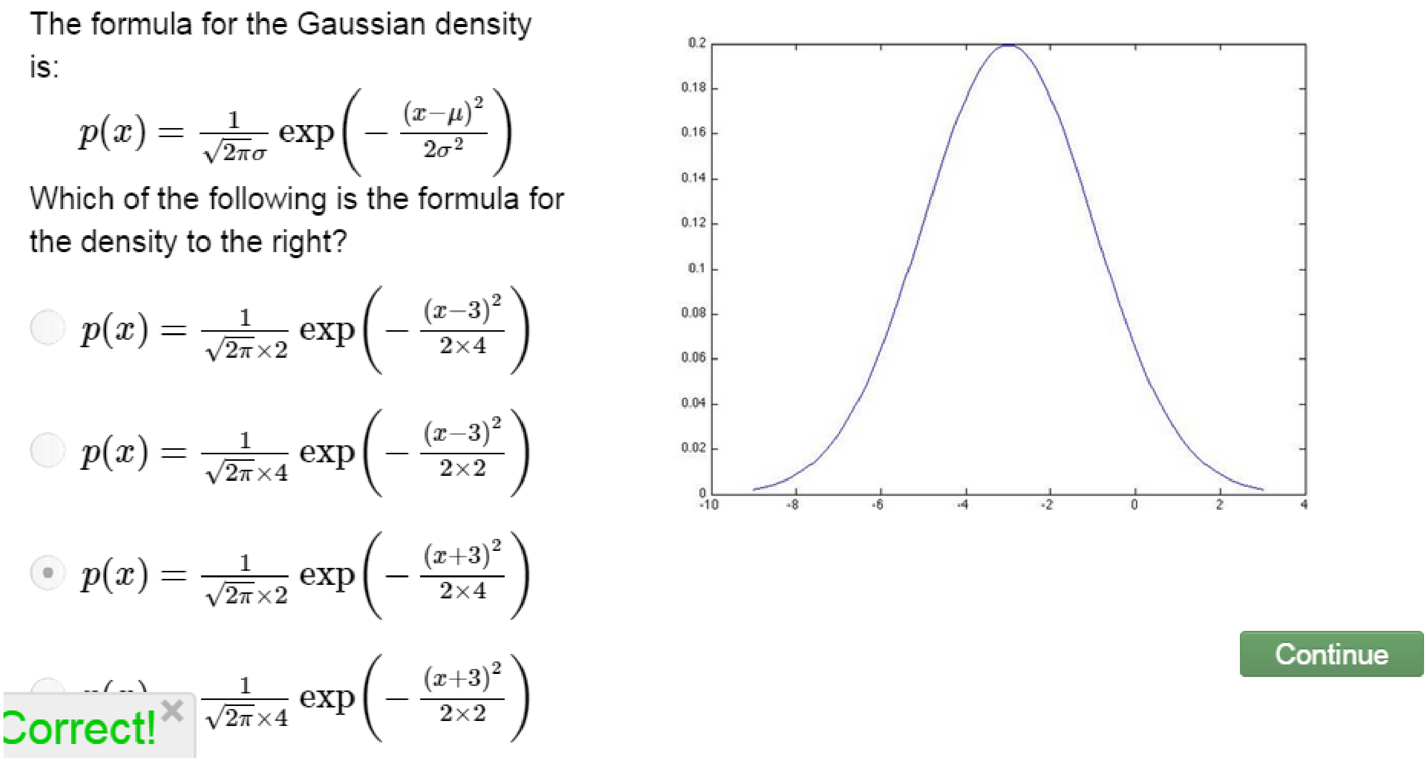
$\mu = \frac{1}{m} \sum_{i=1}^{m} x^{(i)}$

$\sigma^2 = \boxed{\frac{1}{m}} \sum_{i=1}^{m} \underline{(x^{(i)} - \mu)^2}$

$m-1$ $\frac{1}{m-1}$

Andre

习题：高斯分布密度函数求解



The formula for the Gaussian density is:

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Which of the following is the formula for the density to the right?

○ $p(x) = \frac{1}{\sqrt{2\pi}\times 2} \exp\left(-\frac{(x-3)^2}{2\times 4}\right)$

○ $p(x) = \frac{1}{\sqrt{2\pi}\times 4} \exp\left(-\frac{(x-3)^2}{2\times 2}\right)$

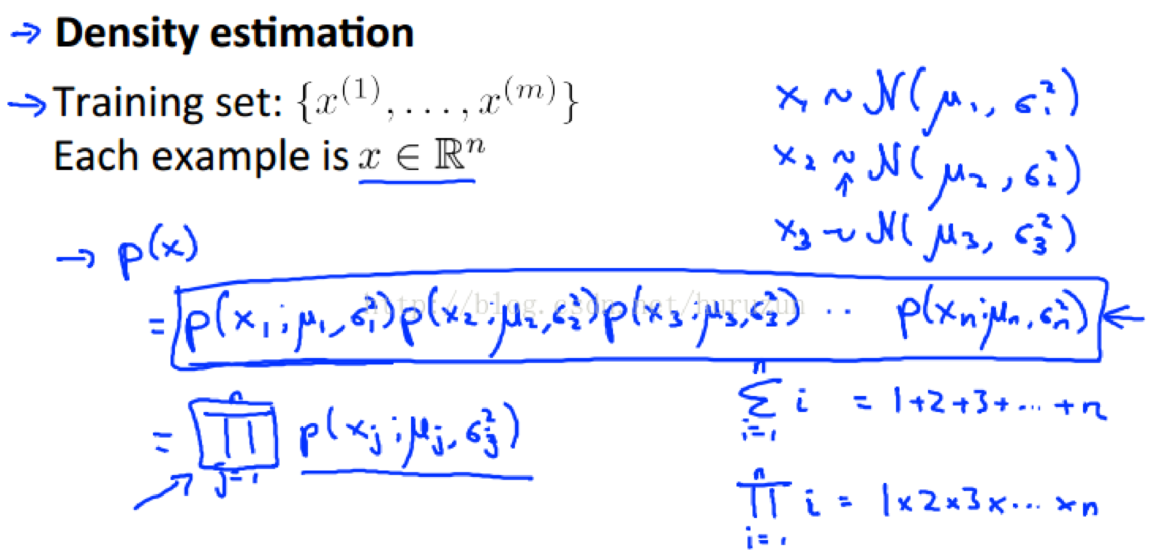
● $p(x) = \frac{1}{\sqrt{2\pi}\times 2} \exp\left(-\frac{(x+3)^2}{2\times 4}\right)$

○ $p(x) = \frac{1}{\sqrt{2\pi}\times 4} \exp\left(-\frac{(x+3)^2}{2\times 2}\right)$

Correct! ✕

Algorithm

密度函数预计算法：



Density estimation

Training set: $\{x^{(1)},\ldots,x^{(m)}\}$

Each example is $\underline{x\in\mathbb{R}^n}$

$x_1\sim\mathcal{N}(\mu_1,\sigma_1^2)$

$x_2\sim\mathcal{N}(\mu_2,\sigma_2^2)$

$x_3\sim\mathcal{N}(\mu_3,\sigma_3^2)$

$$p(x)$$

$$=\boxed{p(x_1;\mu_1,\sigma_1^2)p(x_2;\mu_2,\sigma_2^2)p(x_3;\mu_3,\sigma_3^2)\cdots p(x_n;\mu_n,\sigma_n^2)}$$

$$=\boxed{\prod_{j=1}^{n}}\underline{p(x_j;\mu_j,\sigma_j^2)}$$

$$\sum_{i=1}^{n} i = 1+2+3+\ldots+n$$

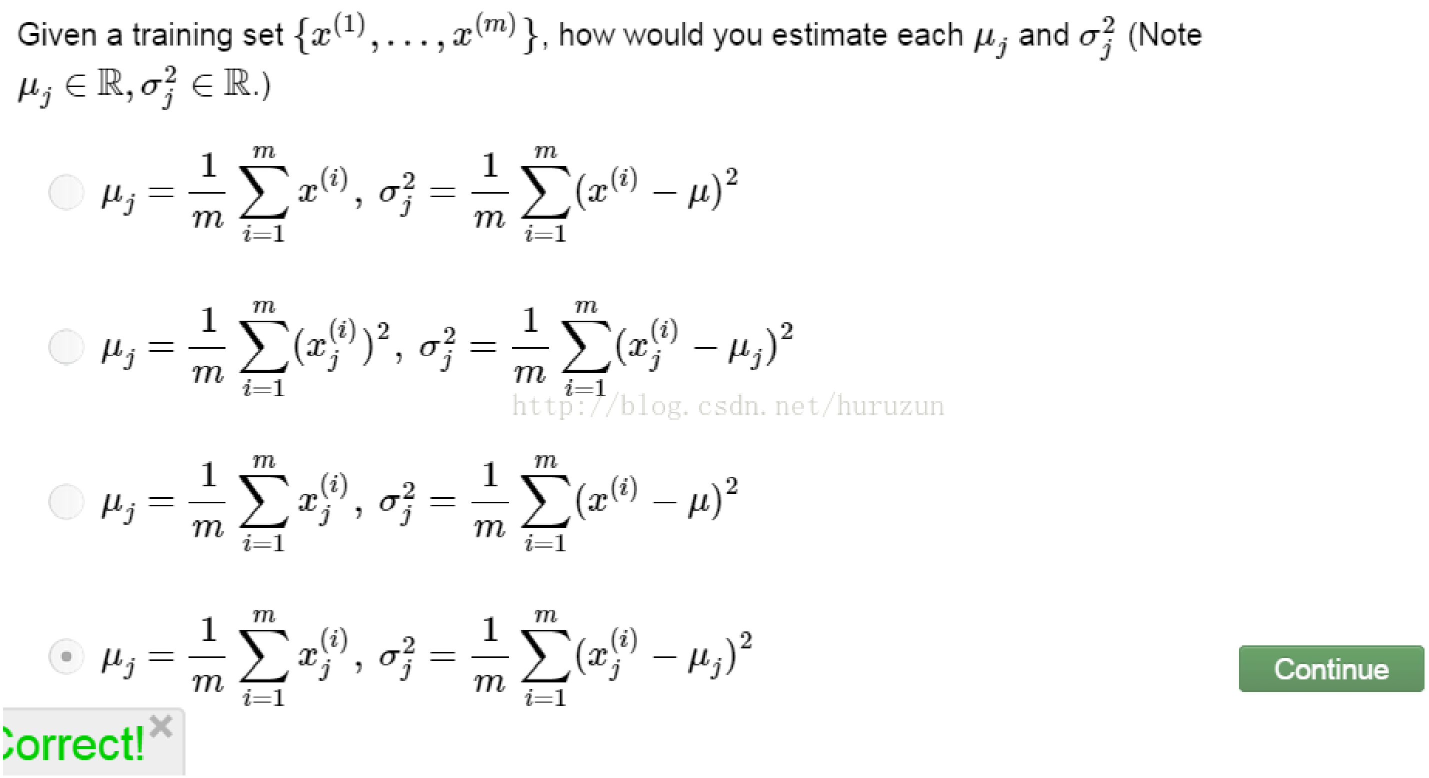
$$\prod_{i=1}^{n} i = 1\times2\times3\times\ldots\times n$$

求P（X）就是密度预计过程。

连乘公式须要每个条件独立，可是假设不是条件独立也能这么计算得到正确结果。

习题：对均值方差预计公式。

J下标表示第J个特征



Given a training set $\{x^{(1)},\ldots,x^{(m)}\}$, how would you estimate each μ_j and σ_j^2 (Note $\mu_j\in\mathbb{R},\sigma_j^2\in\mathbb{R}$.)

○ $\mu_j=\frac{1}{m}\sum_{i=1}^{m}x^{(i)},\ \sigma_j^2=\frac{1}{m}\sum_{i=1}^{m}(x^{(i)}-\mu)^2$

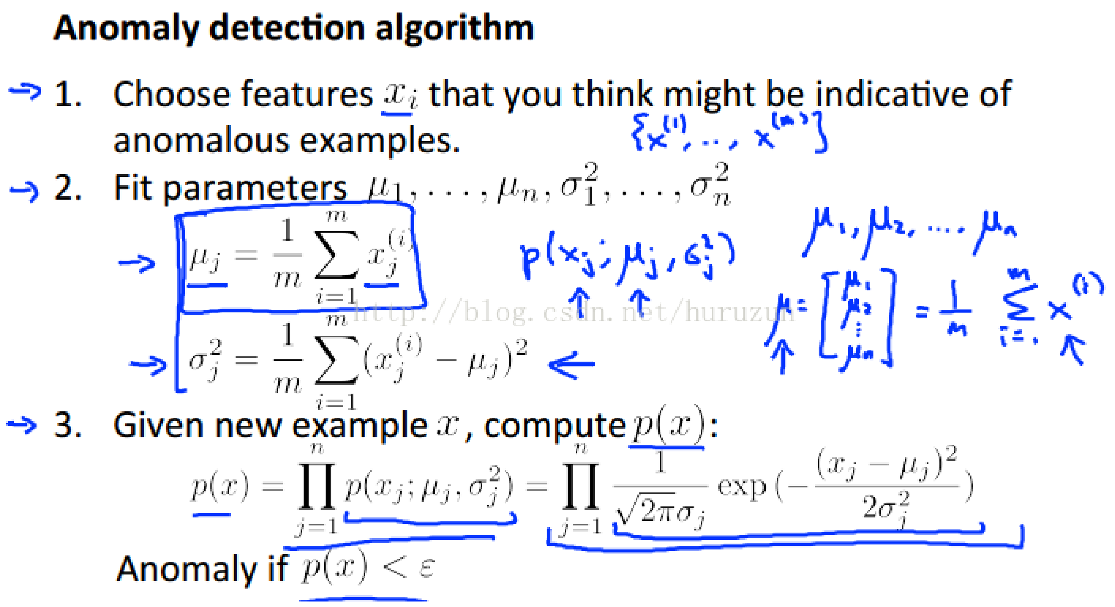
○ $\mu_j=\frac{1}{m}\sum_{i=1}^{m}(x_j^{(i)})^2,\ \sigma_j^2=\frac{1}{m}\sum_{i=1}^{m}(x_j^{(i)}-\mu_j)^2$

○ $\mu_j=\frac{1}{m}\sum_{i=1}^{m}x_j^{(i)},\ \sigma_j^2=\frac{1}{m}\sum_{i=1}^{m}(x^{(i)}-\mu)^2$

● $\mu_j=\frac{1}{m}\sum_{i=1}^{m}x_j^{(i)},\ \sigma_j^2=\frac{1}{m}\sum_{i=1}^{m}(x_j^{(i)}-\mu_j)^2$

Correct! ✕

Anomaly detectionalgorithm



Anomaly detection algorithm

→ 1. Choose features x_i that you think might be indicative of anomalous examples. $\{x^{(1)},\dots,x^{(m)}\}$

→ 2. Fit parameters $\mu_1,\dots,\mu_n,\sigma_1^2,\dots,\sigma_n^2$

$$\to \boxed{\mu_j = \frac{1}{m}\sum_{i=1}^{m} x_j^{(i)}}$$

$$\to \sigma_j^2 = \frac{1}{m}\sum_{i=1}^{m}(x_j^{(i)}-\mu_j)^2 \leftarrow$$

$p(x_j;\mu_j,\sigma_j^2)$

$\mu_1,\mu_2,\dots,\mu_n$

$$\mu = \begin{bmatrix}\mu_1\\ \mu_2\\ \vdots\\ \mu_n\end{bmatrix} = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$$

→ 3. Given new example x, compute $p(x)$:

$$p(x) = \prod_{j=1}^{n} p(x_j;\mu_j,\sigma_j^2) = \prod_{j=1}^{n}\frac{1}{\sqrt{2\pi}\sigma_j}\exp\left(-\frac{(x_j-\mu_j)^2}{2\sigma_j^2}\right)$$

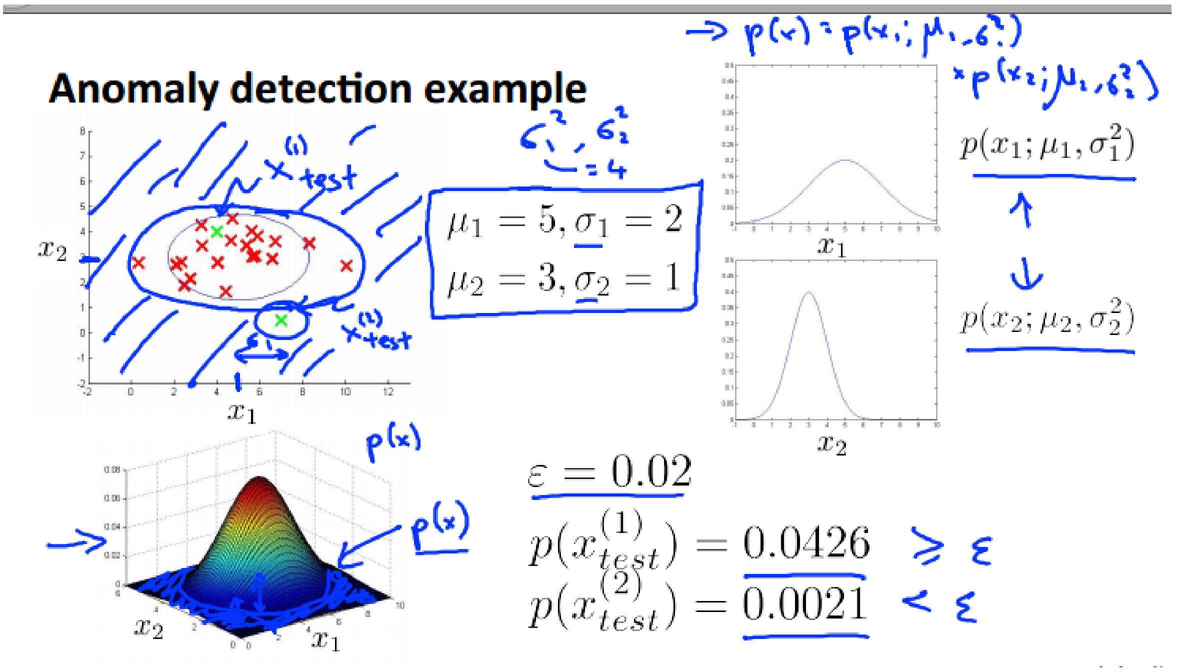
Anomaly if $p(x) < \varepsilon$

1. 选择你觉得可以区分是否为anomalous 的样例特征。

2. 拟合参数即均值和方差。

3. 计算给定数据集上的联合概率密度函数。假设小于设定阈值则判定为异常数据。

进行实例描写叙述这个算法：



Anomaly detection example

$\to p(x) = p(x_1;\mu_1,\sigma_1^2) \times p(x_2;\mu_2,\sigma_2^2)$

σ_1^2, σ_2^2

$\sigma_1^2 = 4$

$$\mu_1 = 5, \sigma_1 = 2$$
$$\mu_2 = 3, \sigma_2 = 1$$

$p(x_1;\mu_1,\sigma_1^2)$

$p(x_2;\mu_2,\sigma_2^2)$

$$\varepsilon = 0.02$$
$$p(x_{test}^{(1)}) = 0.0426 \geq \varepsilon$$
$$p(x_{test}^{(2)}) = 0.0021 < \varepsilon$$

依照上面所写三步流程即可计算：看左下角图。假设我们计算联合概率值较大图形上反映为高度较高，则判定为normal，假设计算得到高度较低，判定为异常。

到这里为止还仅仅是描写叙述了算法运行流程。我们并没有深入描写叙述每一步细节。

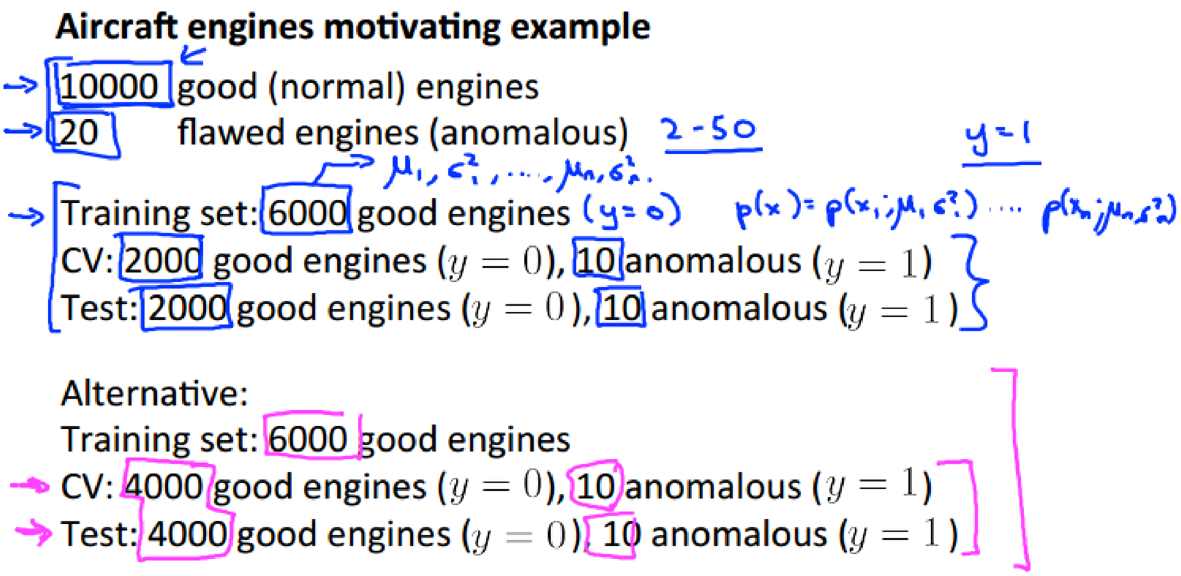
Developing andEvaluating an Anomaly Detection System

我们会发现能用一个数值标准去评价一个学习算法是很重要的，我们能够尝试增加某个feature进行评估，然后去掉该feature再次进行评估。这样得到feature对学习算法的影响。

到如今为止异常检测我们仅仅利用数据并没有数据类标签，是一种无监督学习。

如果我们已经有类标签标记的数据。这样使用异常检测算法就能非常好的进行评估！这是非常重要的一种思维转换。

继续上面提到的飞机发动机样例。



Aircraft engines motivating example

10000 good (normal) engines

20 flawed engines (anomalous) 2-50 y=1

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

Training set: 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: 2000 good engines ($y = 0$), 10 anomalous ($y = 1$)

Test: 2000 good engines ($y = 0$), 10 anomalous ($y = 1$)

Alternative:

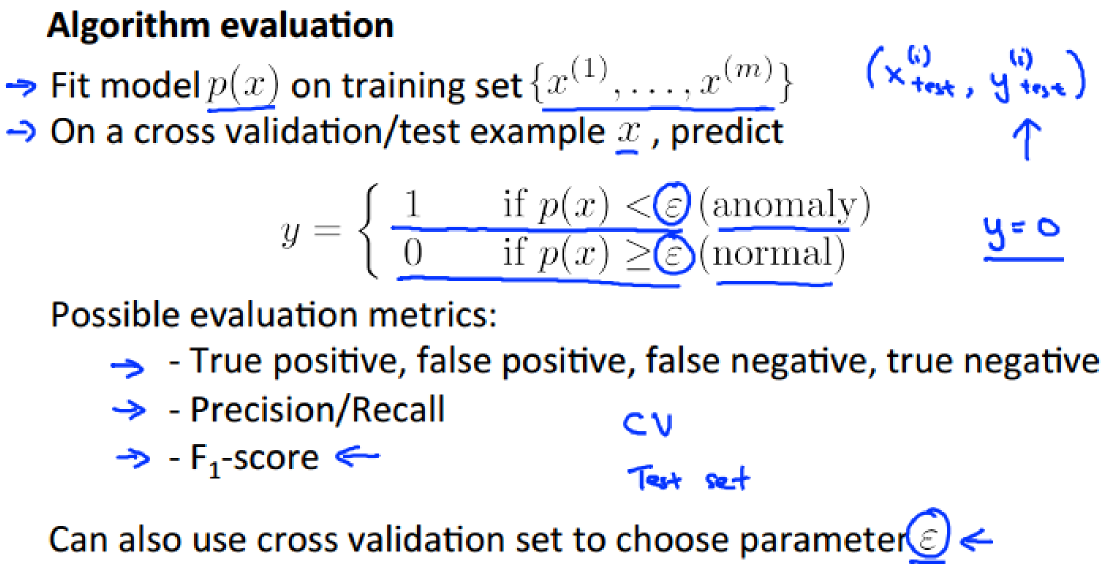
Training set: 6000 good engines

CV: 4000 good engines ($y = 0$), 10 anomalous ($y = 1$)

Test: 4000 good engines ($y = 0$), 10 anomalous ($y = 1$)

我们推荐使用蓝色标记的划分，可是红色标记的划分也有人在这么操作。

算法效果的评估：



Algorithm evaluation

Fit model $p(x)$ on training set $\{x^{(1)}, \ldots, x^{(m)}\}$ $(x_{test}^{(i)}, y_{test}^{(i)})$

On a cross validation/test example x, predict

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

y=1

y=0

Possible evaluation metrics:

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

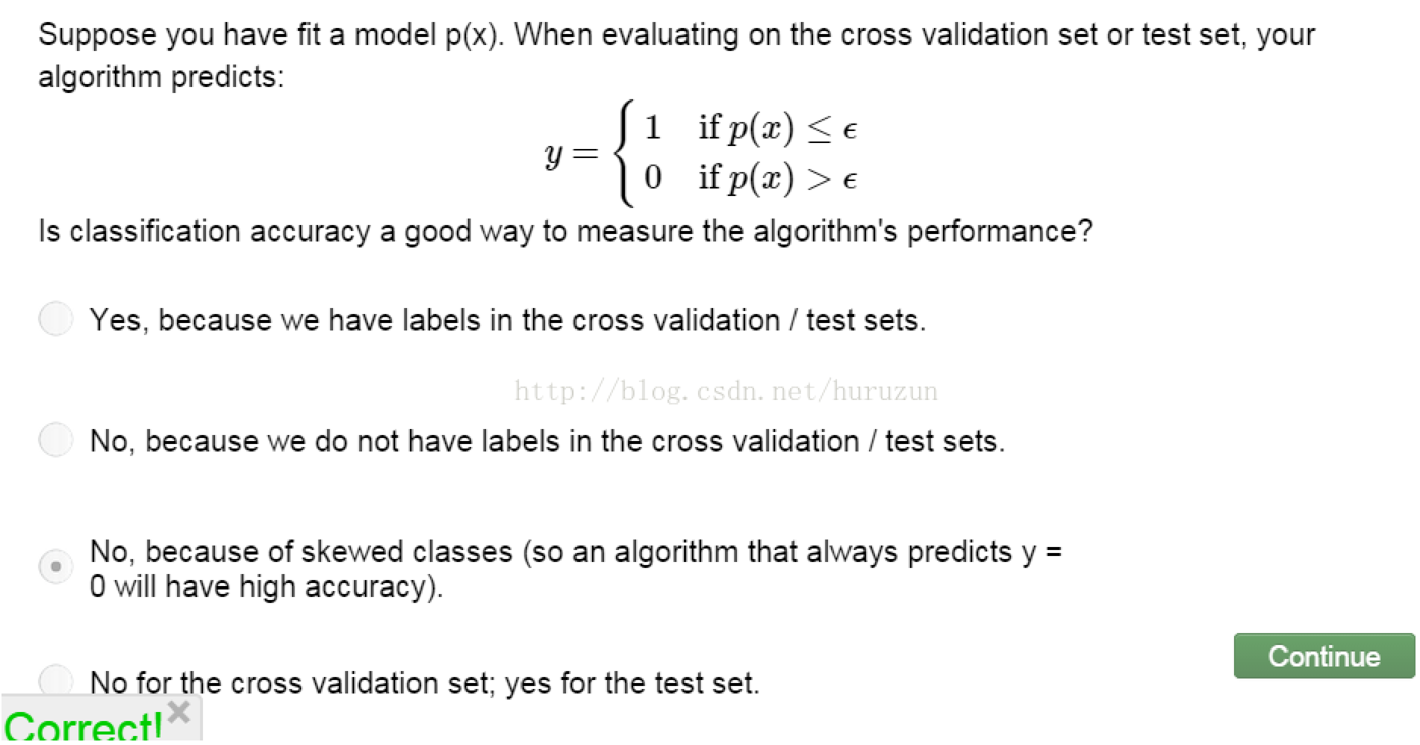
CV

Test set

Can also use cross validation set to choose parameter ε

习题：非常明显在test集上的accuracy不是好的评估标准。由于我们这里是倾斜类！须要用到Precision 和recall F_score来进行评估。

阈值的确定能够通过evaluation metric取值最大确定。当你在设计一个异常检测系统时，关键须要考虑选择何种feature、设定多大的阈值。



Suppose you have fit a model p(x). When evaluating on the cross validation set or test set, your algorithm predicts:

$$y = \begin{cases} 1 & \text{if } p(x) \le \epsilon \\ 0 & \text{if } p(x) > \epsilon \end{cases}$$

Is classification accuracy a good way to measure the algorithm's performance?

○ Yes, because we have labels in the cross validation / test sets.

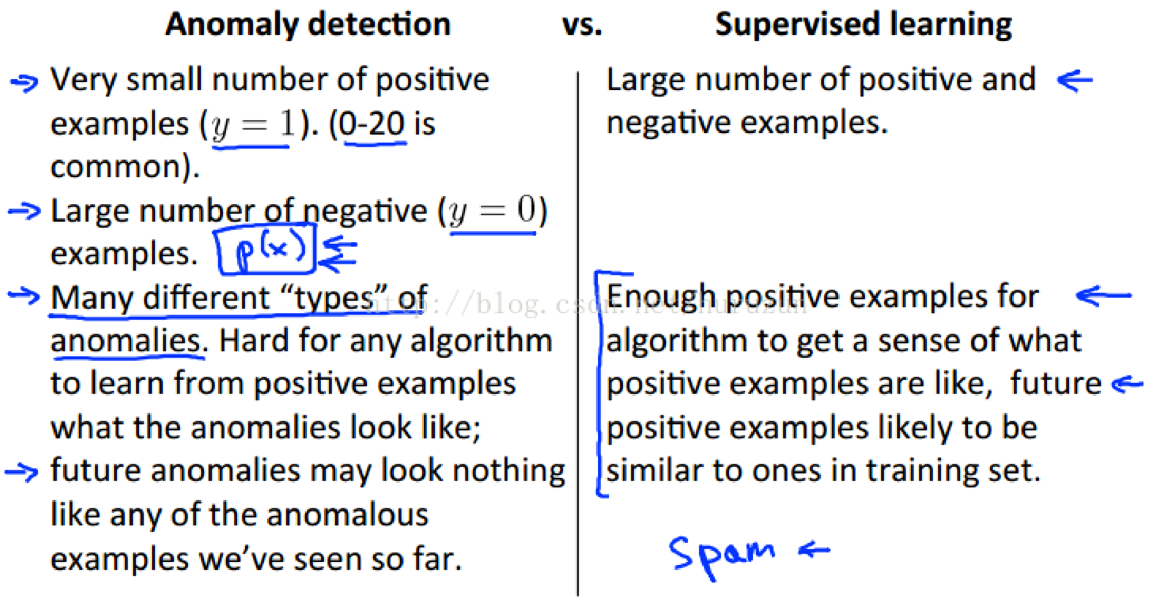
○ No, because we do not have labels in the cross validation / test sets.

● No, because of skewed classes (so an algorithm that always predicts y = 0 will have high accuracy).

○ No for the cross validation set; yes for the test set.

Anomaly DetectionVS Supervised Learning

讲到这里我们肯定都有困惑。当我们有数据类标签，为什么我们不直接使用监督性学习而使用Anomaly detection，接下来就对两者进行对照。



Anomaly detection vs. **Supervised learning**

Anomaly detection:

- Very small number of positive examples ($y = 1$). (0-20 is common).
- Large number of negative ($y = 0$) examples. $\boxed{p(x)}$
- Many different "types" of anomalies. Hard for any algorithm to learn from positive examples what the anomalies look like;
- future anomalies may look nothing like any of the anomalous examples we've seen so far.

Supervised learning:

- Large number of positive and negative examples.
- Enough positive examples for algorithm to get a sense of what positive examples are like, future positive examples likely to be similar to ones in training set.

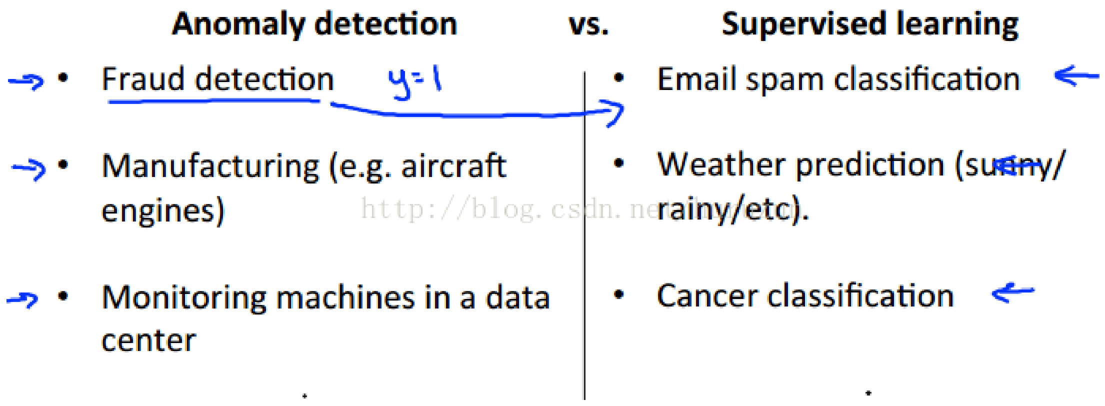
首先Anomaly detection 在数据集上特点是：非常少量的positive 数据，非常大量的negative数据，这样我们使用大量的negative数据可以好的拟合求得联合高斯概率密度函数。而supervised learning中 positive negative数据量都大。

其次我们有不同类型的异常数据，可是异常数据量非常小，不论什么算法都非常难在小的Anomaly数据集上学习得到Anomaly是什么样子。

上面两者对照是你应用Anomaly detection 还是supervise learning 的一些重要区分标准。

Spam是常常提到的一种学习系统。尽管我们有非常多类型的Spam。可是每种类型的Spam我们都有比較多的数据。所以Spam问题我们应用的是supervise learning。

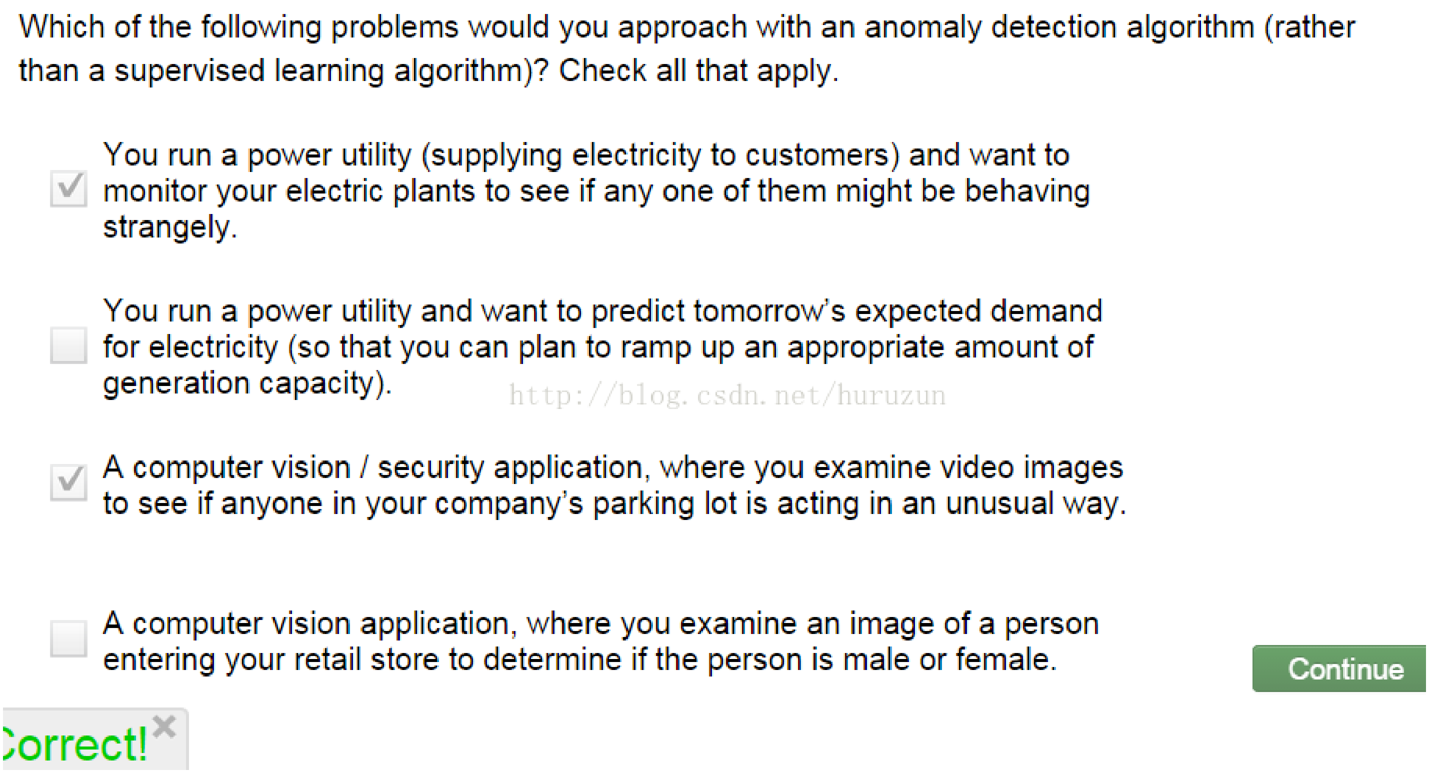
事实上这两种状态并非全然切割的。举例说假设我们在交易时有非常多为Fraud的，则我们学习问题由Anomaly detection 转变为supervise learning。



Anomaly detection vs. **Supervised learning**

Anomaly detection	Supervised learning
• Fraud detection	• Email spam classification
• Manufacturing (e.g. aircraft engines)	• Weather prediction (sunny/ rainy/etc).
• Monitoring machines in a data center	• Cancer classification

习题：直观对两种情况的推断



Which of the following problems would you approach with an anomaly detection algorithm (rather than a supervised learning algorithm)? Check all that apply.

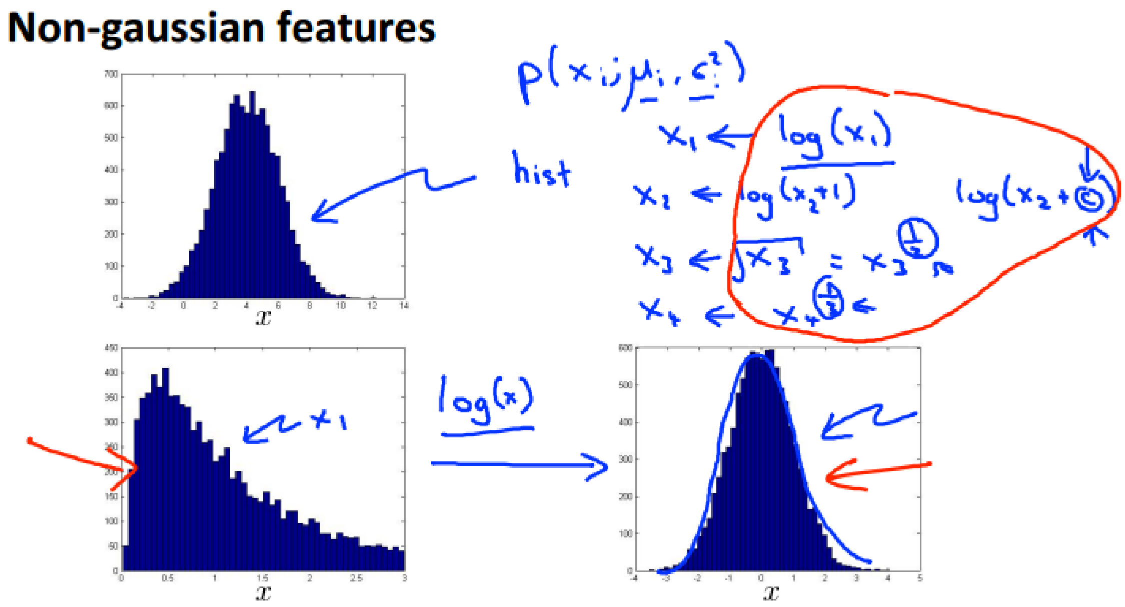
☑ You run a power utility (supplying electricity to customers) and want to monitor your electric plants to see if any one of them might be behaving strangely.

☐ You run a power utility and want to predict tomorrow's expected demand for electricity (so that you can plan to ramp up an appropriate amount of generation capacity).

☑ A computer vision / security application, where you examine video images to see if anyone in your company's parking lot is acting in an unusual way.

☐ A computer vision application, where you examine an image of a person entering your retail store to determine if the person is male or female.

ChoosingWhat Features to Use

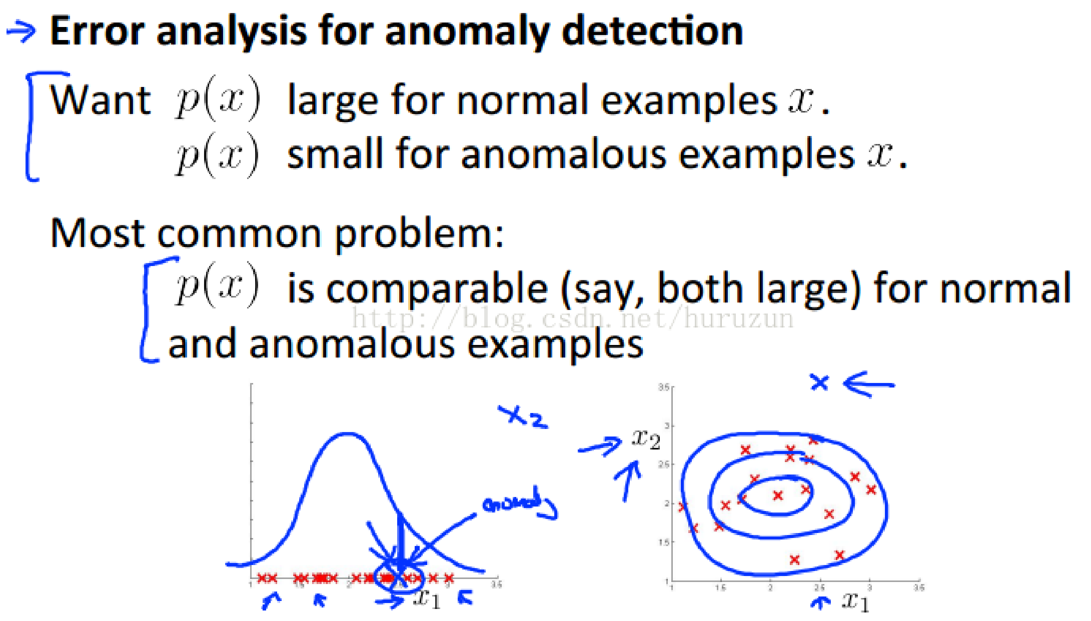


Non-gaussian features

前面说的方法都是假定数据满足高斯分布，也提到了假设分布不是高斯分布，上述方法也是能够使用，仅仅是假设我们对分布进行转换使得近似于高斯分布，那么会得到更好的效果。

上图就举例用log等函数进行转换。实现层面octave以下尝试转换就能够得到非常好近似高斯。

怎样选择feature：



Error analysis for anomaly detection

Want $p(x)$ large for normal examples x.

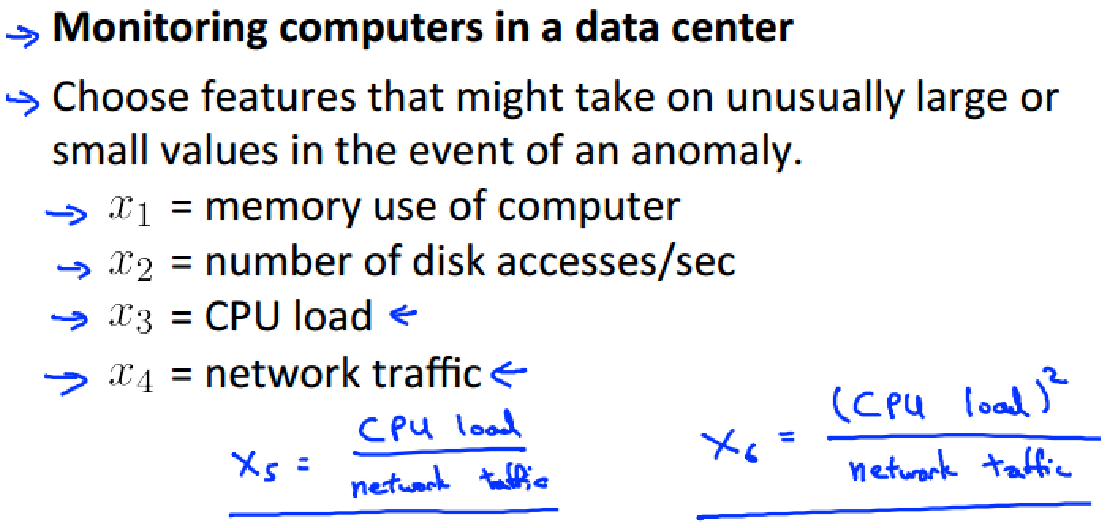
$p(x)$ small for anomalous examples x.

Most common problem:

$p(x)$ is comparable (say, both large) for normal and anomalous examples

思想类似于supervise learning中的error analysis ，当左边图进行学习时我们得到了错误结果，这时我们须要增加新的特征X2使得那个点与正常数据得到区分！如上图从左到右所看到的意。

还是回到前面提到的监视数据中心电脑，我们为了分析究竟是哪个特征引起Anomaly，通过构造新的feature来进行推断。



Monitoring computers in a data center

Choose features that might take on unusually large or small values in the event of an anomaly.

- x_1 = memory use of computer
- x_2 = number of disk accesses/sec
- x_3 = CPU load
- x_4 = network traffic

$x_5 = \frac{\text{CPU load}}{\text{network traffic}}$ $\qquad$ $x_6 = \frac{(\text{CPU load})^2}{\text{network traffic}}$

习题：Anomaly detection 算法不能非常好区分开normal和Anomaly时。我们通常须要添加特征来使得它们得以区分。

Suppose your anomaly detection algorithm is performing poorly and outputs a large value of p(x) for many normal examples and for many anomalous examples in your cross validation dataset. Which of the following changes to your algorithm is most likely to help?

○ Try using fewer features.

● Try coming up with more features to distinguish between the normal and the anomalous examples.

○ Get a larger training set (of normal examples) with which to fit p(x).

○ Try changing ϵ.

Continue