

Anomaly detection

2023年6月2日 16:23

1.原理

Anomaly detection example

Aircraft engine features:

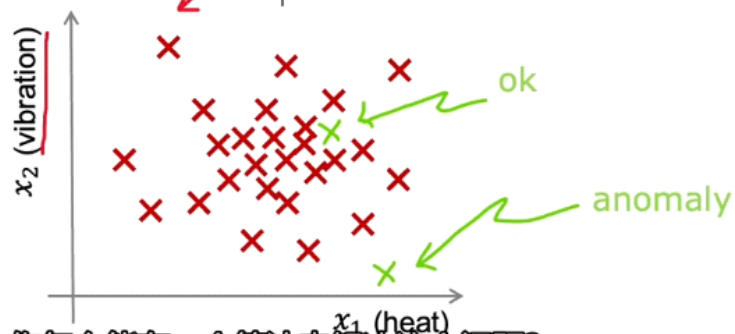
x_1 = heat generated

x_2 = vibration intensity

...

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

New engine: x_{test}



DeepLearning.AI How can you have an algorithm address this problem? Andrew Ng

屏幕剪辑的捕获时间: 2023/6/4 21:28

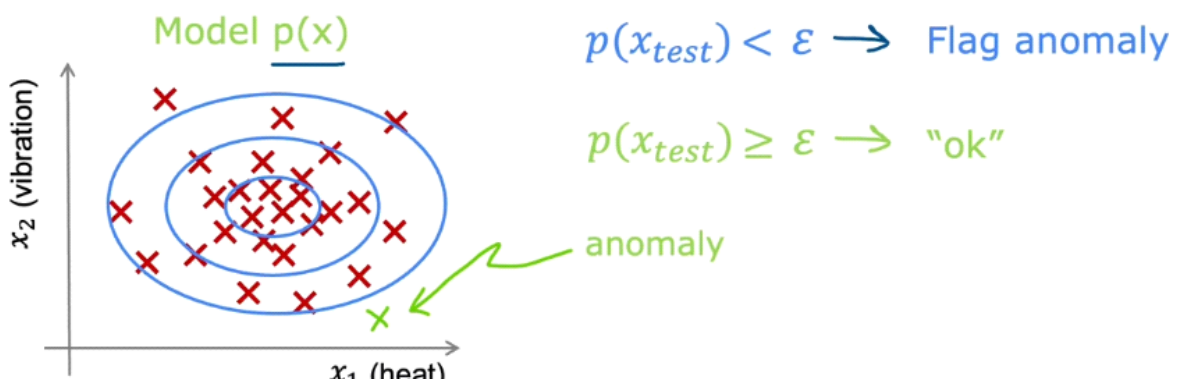
密度估计(Density Estimation)

评估数据各个特征取特定值的概率, 设定阈值以评估test data是否异常

Density estimation

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

Is x_{test} anomalous?



新飞机发动机很有可能具有接近这些内

There's a very high chance that the new airplane engine will have features

屏幕剪辑的捕获时间: 2023/6/4 21:32

异常监测在各个领域都有很多应用, 主要作用是对于有轻微异常嫌疑的案例, 可以指导安全团队对其进行进一步监测

2. 高斯分布 (正态分布)

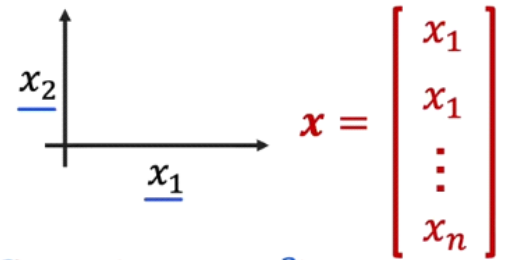
如何应用于Anomaly Detection?

- ①. 单个数据情况下, 利用Training set拟合出正态分布, 利用“ 3σ ”原则即可;
- ②. 对于 n 个特征, 假设各个特征相互独立, 拟合出 n 个正态分布模型, 计算概率

Density estimation

Training set: $\{x^{(1)}, x^{(2)}, \dots, x^{(m)}\}$

Each example x_i has n features



$$p(x) = p(x_1; \mu_1, \sigma_1^2) * p(x_2; \mu_2, \sigma_2^2) * p(x_3; \mu_3, \sigma_3^2) *$$

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

$$p(x_n; \mu_n, \sigma_n^2)$$

$$\begin{aligned} p(x_1 = \text{high temp}) &= 1/10 \\ p(x_2 = \text{high vibra}) &= 1/20 \\ p(x_1, x_1) &= p(x_1) * p(x_2) \\ &= \frac{1}{10} * \frac{1}{20} = \frac{1}{200} \end{aligned}$$

3. 选择参数 ϵ

引入交叉验证数据集, 并且在检验数据中加入一些带有标签的数据 (异常/非异常), 会很有用

The importance of real-number evaluation

When developing a learning algorithm (choosing features, etc.), making decisions is much easier if we have a way of evaluating our learning algorithm.

Assume we have some labeled data, of anomalous and non-anomalous examples. ($y = 0$ if normal, $y = 1$ if anomalous).

Training set: $x^{(1)}, x^{(2)}, \dots, x^{(m)}$ (assume normal examples/not anomalous)

Cross validation set: $(x_{cv}^{(1)}, y_{cv}^{(1)}), \dots, (x_{cv}^{(m_{cv})}, y_{cv}^{(m_{cv})})$ } Include a few anomalous examples

Test set: $(x_{test}^{(1)}, y_{test}^{(1)}), \dots, (x_{test}^{(m_{test})}, y_{test}^{(m_{test})})$ } $y=1$

让我们用飞机发动机的例子来说明这一点。

DeepLearning.AI Let's illustrate this with the aircraft engine example. Andrew Ng

使用交叉验证数据集的目的是确定合适的 ϵ 值, 虽然这是无监督学习, 但是引入带标签的数据对于训练算法十分有效

在CV数据集上确定了合适的 ϵ , 再用test set检验算法的准确度

还可以只用Training set与CV set, 这种情况主要是异常案例很少时, 没有足够数据创建一个与CV集完全不同的test set, 但是无法评估算法在新数据上的工作性能, 可能对 ϵ 与特征选择过拟合

Aircraft engines monitoring example

10000 good (normal) engines $y=0$ \leftarrow 2-50
→ 20 flawed engines (anomalous) $y=1$
Train algorithm
(Training set: 6000 good engines $y=0$, 10 anomalous $y=1$) \leftarrow ϵ, x_j
(CV: 2000 good engines $y=0$, 10 anomalous $y=1$) \leftarrow
(Test: 2000 good engines $y=0$, 10 anomalous $y=1$) \leftarrow

Alternative:

→ Training set: 6000 good engines $y=0$ \leftarrow 2
→ CV: 4000 good engines $y=0$, 20 anomalous $y=1$ \leftarrow
No test set \leftarrow ϵ, x_j

况, 请注意有

DeepLearning.AI

Stanford ONL

just be aware that there's

Andrew Ng

屏幕剪辑的捕获时间: 2023/6/4 22:09

既然用了Labeled data, 为何不用监督学习, 而是用异常监测?

1. 当“异常”数据的数量很小时, 倾向于用异常监测, 而当“异常”数据较多时, 可以使用监督学习
2. 如果“异常”的种类可能有許多, 未来出现的异常情况很可能与现有的异常情况全都不相同, 这种情况下Anomaly Detection会更合适, 如金融诈骗;
如果认为未来会出现的“异常”情况与现有的情况很相似, 就可以考虑采用监督学习, 这种情况下还要有足够的案例让算法学习“异常”的特征, 如垃圾邮件监测。

Anomaly detection vs. Supervised learning

Very small number of positive examples ($y=1$). (0-20 is common).
Large number of negative ($y=0$) examples. $p(x)$ $y=1$

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.

Fraud

Large number of positive and negative examples.

20 positive 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.

Spam

您过去可能在训练集中看到的电子邮件。

emails that you have probably seen in the past in your training set. Andrew Ng

屏幕剪辑的捕获时间: 2023/6/6 15:49

Anomaly detection vs. Supervised learning

- | | |
|--|--|
| → Fraud detection | → Email spam classification |
| → Manufacturing - Finding <u>new previously unseen defects</u> in manufacturing. (e.g. aircraft engines) | → Manufacturing - Finding known, previously <u>seen</u> defects <u>scratches</u> $y=1$ |
| → Monitoring machines in a data center | → Weather prediction (sunny/rainy/etc.) |
| ⋮ | → Diseases classification |
| ⋮ | ⋮ |

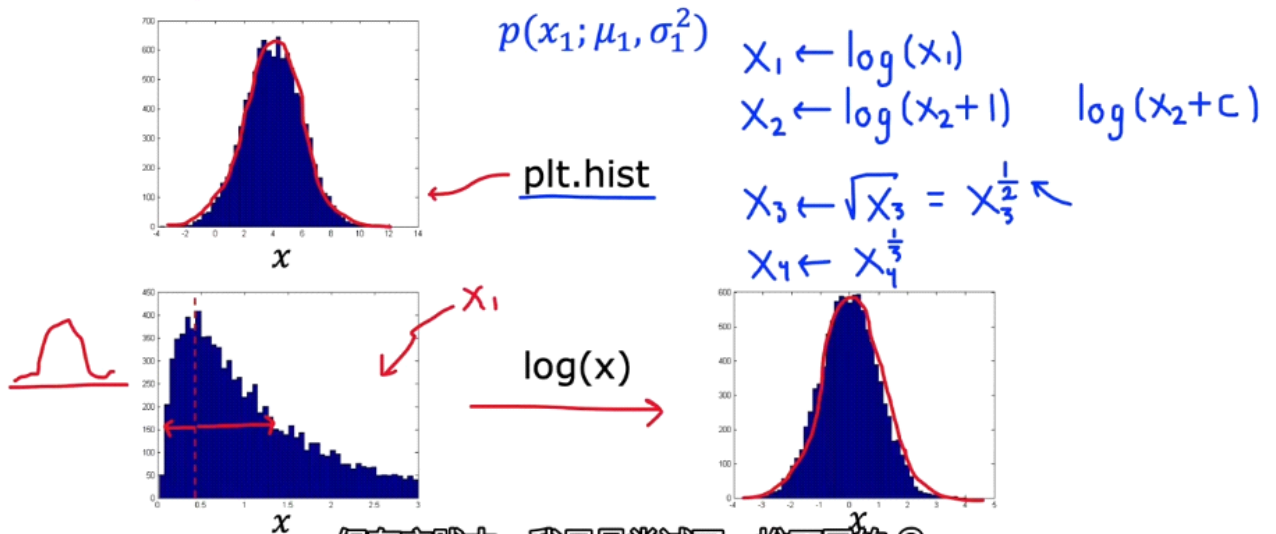
屏幕剪辑的捕获时间: 2023/6/11 9:31

总之，异常监测偏向于找出数据中与平常情况不同的个例，尤其是Training Set中未出现过的情况；而监督学习目的是学会如何分辨不同的异常情况，所以更多时候适用于检查Training Set中出现过的异常情况

4. 如何选择特征

选择特征的重要性比监督学习更重要，尤其是需要符合高斯分布的特征，如果不符合高斯分布，可以通过一些变换使其近似于高斯分布

Non-gaussian features



但在实践中，我只是尝试了一堆不同的 C

DeepLearning.AI But in practice I just try a bunch of different values of C, Andrew Ng

屏幕剪辑的捕获时间: 2023/6/11 9:47

5. 误差分析

试想，如果实际有一些数据是异常的，但是将其放在训练集中来看，异常的概率与正常的概率相近且都很大，那么这可能说明需要一些其他的特征来区分异常与否，即通过观察这个异常的个例，思考是什么让我认为他是异常的，从而加入其他的特征改善算法

Error analysis for anomaly detection

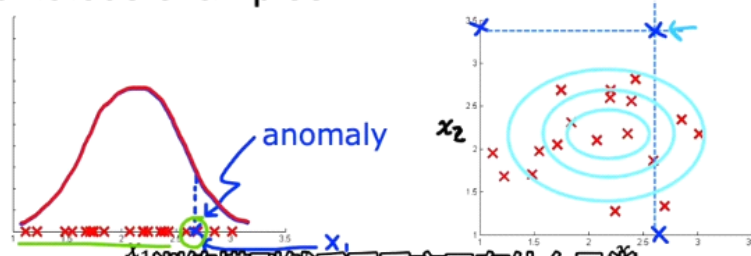
Want $p(x) \gg \epsilon$ large for normal examples x .
 $n(x) \ll \epsilon$ small for anomalous examples x .

Error analysis for anomaly detection

Want $p(x) \gg \epsilon$ large for normal examples x .
 $p(x) \ll \epsilon$ small for anomalous examples x .

Most common problem:

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



to train the model and then to see what anomalies in the cross

屏幕剪辑的捕获时间: 2023/6/11 9:58

此外，有时在实际算法中，可能单独拿出监测指标中的一个来看是正常的，比如特征A很高，特征B很低，单独看都是正常的，但是这两种情况同时出现就是异常的，这时可以尝试组合原有的特征以产生新的特征

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 ←

$x_5 = \frac{\text{CPU load}}{\text{network traffic}}$

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

Deciding feature choice based on $p(x)$

– large for normal examples, and becomes small for anomaly in the cross validation set

对于正常示例仍然很大，但在您的交叉验证集中的异常中它变得很小。
it becomes small in the anomalies in your cross validation set.

屏幕剪辑的捕获时间: 2023/6/11 10:03