Anomaly Detection Algorithm ¶

Let's say that we have an unlabeled training set of m examples: $\{x^{(1)}, \ldots, x^{(m)}\}$ where each example is a vector, $x \in \mathbb{R}^n$ (we have n features). So our training set could be, feature vectors from the last m aircraft engines being manufactured. Or it could be features from m users or something else.

- We're going to Model $P(x)$ from the data set
  - What are high probability features and low probability features
- Each of x is a vector
- So model $p(x)$ as:
  - $p(x) = p(x_1; \mu_1, \sigma_1^2)\dot{p}(x_2; \mu_2, \sigma_2^2) \ldots p(x_n; \mu_n, \sigma_n^2)$

What we're going to do, is assume that each feature, is distributed according to a Gaussian probability distribution $x \sim N(\mu, \sigma^2)$.

- $x_1 \sim N(\mu_1, \sigma_1^2)$
- $x_2 \sim N(\mu_2, \sigma_2^2)$
- $\ldots$
- $x_n \sim N(\mu_n, \sigma_n^2)$

In statistics, this is called an **"independence assumption"** on the values of the features inside training example x. Turns out this equation makes an independence assumption for the features, although algorithm works if features are independent or not.

More compactly, the above expression can be written as follows:

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

The problem of estimation this distribution is sometimes call the problem of **density estimation**.

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

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

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

How do we implement anomaly detection algorithm?

1.- Choose features x_i that we think might be indicative of anomalous examples.

- Try to come up with features which might help identify something anomalous - may be unusually large or small values

- More generally, chose features which describe the general properties
- This is nothing unique to anomaly detection - it's just the idea of building a sensible feature vector

2.- Fit parameters python $\mu_1, \dots, \mu_n; \sigma_1^2, \dots, \sigma_n^2$

- Calculate the mean and variance of each j feature
- Calculate $\mu_j = \frac{1}{m} \sum_{i=1}^{m} x_j^{(i)}$
- Calculate $\sigma_j^2 = \frac{1}{m} \sum_{i=1}^{m} (x_j^{(i)} - \mu_j)^2$
- Fit is a bit misleading, really should just be "Calculate parameters for 1 to n"
- So we're calculating standard deviation σ_i^2 and mean μ_i for each feature
- We should used some vectorized implementation rather than a loop

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

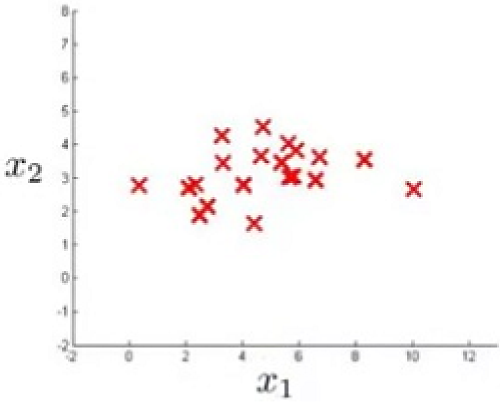
$$p(x) = \prod_{j=1}^{n} p(x_j; \mu_j, \sigma_j^2) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_j} exp(-\frac{(x_j - \mu_j)^2}{2\sigma_j^2})$$

- We compute the formula shown (i.e. the formula for the Gaussian probability)
- If the number is very small, very low chance of it being "normal" Anomaly if $p(x) < \epsilon$

A vectorized version of the calculation for μ is $\mu = \frac{1}{m} \sum_{i=1}^{m} x^{(i)}$. We can vectorize σ^2 similarly.

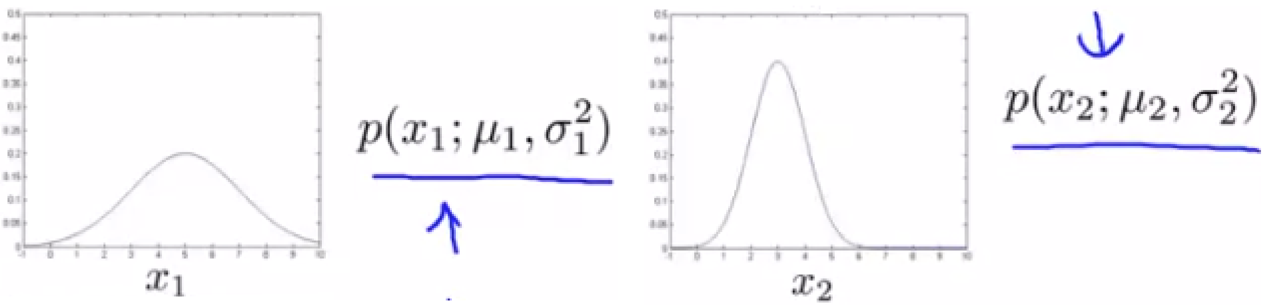
Anomaly Detection example

Let's say we have a data set shown below:

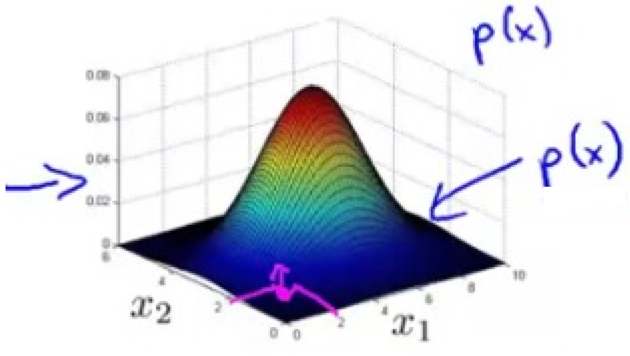


- $x_1 \in \mathbb{R}^n$; μ_1 = 5, σ_1 = 2
- $x_2 \in \mathbb{R}^n$; μ_2 = 3, σ_2 = 1

If we plot the Gaussian probability distributin for x_1 and x_2 we get something like this:



If you plot the product of these things you get a surface plot like this:



- With this surface plot, the height of the surface is the probability - $p(x)$
- We can't always do surface plots, but for this example it's quite a nice way to show the probability of a 2D feature vector
- **Check if a value is anomalous**:
  - Set ϵ as some value, say ϵ = 0.02
  - Say we have two new data points new data-point has the values
    - $x_{\text{test}}^{(1)}$
    - $x_{\text{test}}^{(2)}$
  - We compute the probability (with ϵ = 0.02):
    - p($x_{\text{test}}^{(1)}$) = 0.0426 $\geq \epsilon$ ($x_{\text{test}}^{(1)}$ it's not an anomaly)
    - p($x_{\text{test}}^{(2)}$) = 0.0021 $< \epsilon$ ($x_{\text{test}}^{(2)}$ it's an anomaly)

What this is saying is if you look at the surface plot, all values above a certain height are normal, all the values below that threshold are probably anomalous.