Anomaly Detection using the Multivariate Gaussian Distribution ¶

To recap the Multivariate Gaussian Distribution or the Multivariate Normal Distribution has two parameters, μ and Σ.

- Parameters: $\mu,\ \Sigma$

Where μ is an n-dimensional ($\mu \in \mathbb{R}^n$) vector and Σ the covariance matrix is an n x n matrix ($\Sigma \in \mathbb{R}^{nxn}$).

As mentioned, multivariate Gaussian modeling uses the following equation;

$$p(x;\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

The probability of x, as parameterized by μ and Σ, and as we vary μ and Σ, we can get a range of different distributions.

Parameter fitting:

Given training set: $\{x^{(1)}, x^{(2)}, \dots, x^{(m)}\}$, each example is an n-dimensional vector $x \in \mathbb{R}^n$

The formula for estimating the parameters is:

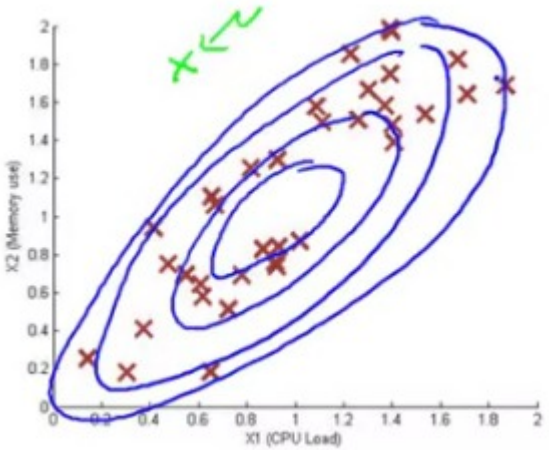
$$\mu = \frac{1}{m}\sum_{i=1}^{m} x^{(i)} \qquad \Sigma = \frac{1}{m}\sum_{i=1}^{m}(x^{(i)}-\mu)(x^{(i)}-\mu)^T$$

So given the data set here is how we estimate μ and Σ.

Let's take this method and just plug it into an anomaly detection algorithm. So how do we put all of this together to develop an anomaly detection algorithm?

Anomaly detection algorithm with multivariate Gaussian distribution

1. Fit model $p(x)$ by setting the parameters μ and Σ using the given previous formulas

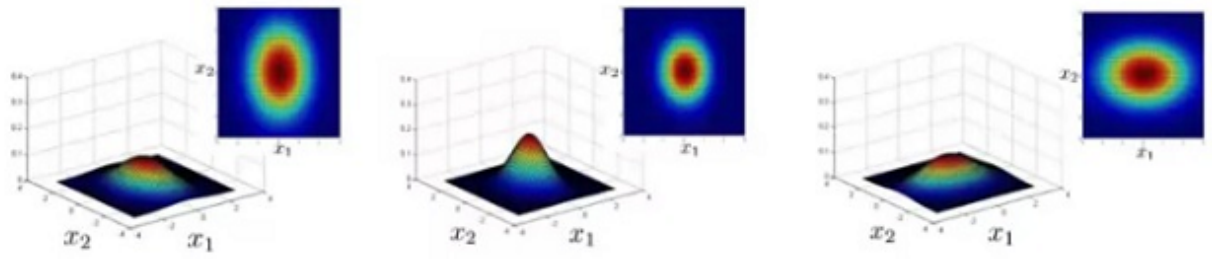


2. We then compute $p(x)$ using the new formula in the previous example and flag an anomaly if $p(x) < \epsilon$.

- If we apply the multivariate Gaussian distribution to a new example (which is an anomaly), it will actually correctly flag that example. as an anomaly.
  - If $p(x_{test}) < \epsilon$ --> flag this as an anomaly
  - If $p(x_{test}) \geq \epsilon$ --> this is OK

Relationship to original model

Original model: $p(x) = p(x_1; \mu_1, \sigma_1^2) \times p(x_2; \mu_2, \sigma_2^2) \times \cdots \times p(x_n; \mu_n, \sigma_n^2)$



Finally, we should mention how multivariate Gaussian relates to our original simple Gaussian model (where each feature is looked at individually)

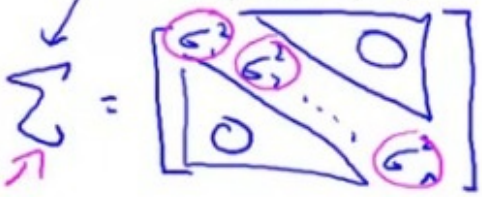
- The original model for $p(x)$ corresponds to a multivariate Gaussian where the contours of $p(x; \mu, \Sigma)$ are axis-aligned.
  - i.e. the normal Gaussian model is a special case of multivariate Gaussian distribution

So, it turns out that it's possible to show mathematically that **the original model** (the ecuation of $p(x)$) is the **same** as a **multivariate Gaussian distribution** but with a constraint.

- Has this constraint that the covariance matrix sigma as zeros on the non-diagonal values

$$\rightarrow p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

where $\Sigma =$



If we plug our variance values into the covariance matrix the models are actually identical

- The multivariate Gaussian model can automatically capture correlations between different features of x.

Original model vs. Multivariate Gaussian

Original model

$$p(x_1; \mu_1, \sigma_1^2) \times \cdots \times p(x_n; \mu_n, \sigma_n^2)$$

- **Probably used more often**
- **Manually create features to capture anomalies where x_1, x_2 take unusual combinations values**
  - So need to make extra features
- **Computationally cheaper (alternatively, scales better to large n)**
  - Even if n = 10,000 or n = 100,000 the original model works fine
- **Ok even if m (training set size) is small**

Multivariate Gaussian

$$p(x;\mu,\Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} exp(-1/2(x-\mu)^T\Sigma^{-1}(x-\mu))$$

- **Used less frequently**
- **Automatically captures correlations between features**
  - So no need to create extra values
- **Computationally more expensive**
  - Must compute inverse of matrix Σ which is [n x n] --> ($\Sigma \in \mathbb{R}^{nxn}$)
  - So lots of features is bad - makes this calculation very expensive
  - So if n = 100,000 not very good
- **Must have $m > n$, or else Σ is non-invertible**
  - If this is not true then we have a singular matrix (non-invertible)
  - So should be used only in $m >> n$
  - **If you find the matrix is non-invertible, could be for one of two main reasons**
    - $m < n$
      - So use original simple model
    - **Redundant features (i.e. linearly dependent)**
      - i.e. two features that are the same
      - If this is the case WE could use PCA or sanity check your data

Video Question: Consider applying anomaly detection using a training set $\{x^{(1)}, \ldots, x^{(m)}\}$ where $x^{(i)} \in \mathbb{R}^n$. Which of the following statements are true? Check all that apply.

> The original model $p(x_1;\mu_1,\sigma_1^2) \times \cdots \times p(x_n;\mu_n,\sigma_n^2)$ corresponds to a multivariate Gaussian where the contours of $p(x;\mu,\Sigma)$ are axis-aligned.

- Using the multivariate Gaussian model is advantageous when m (the training set size) is very small ($m < n$).

> The multivariate Gaussian model can automatically capture correlations between different features in x.

> The original model can be more computationally efficient than the multivariate Gaussian model, and thus might scale better to very large values of n (number of features).