Data Science Advanced

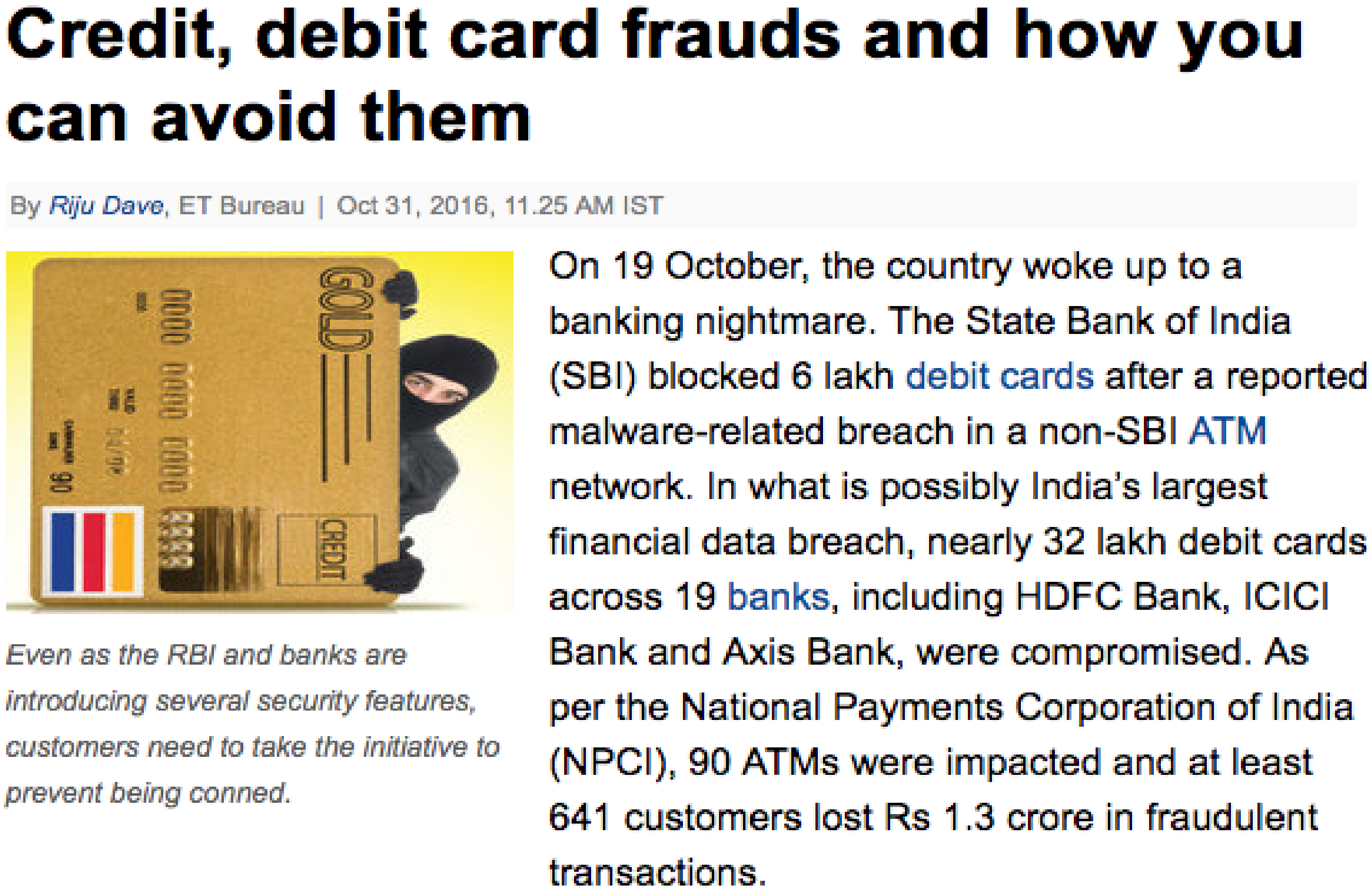
Lesson05-Gaussian Distribution in Anomaly Detection

Objective

After completing this lesson you will be able to:

Anomaly Detection

- Understand the issue with bias in data
- Understand Gaussian distribution fitting for detecting anomaly
- Define the steps in building anomaly detection model when features are modelled independently
- Differentiate anomaly detection from supervised learning
- Multivariate Gaussian distribution and anomaly detection when features are modelled together

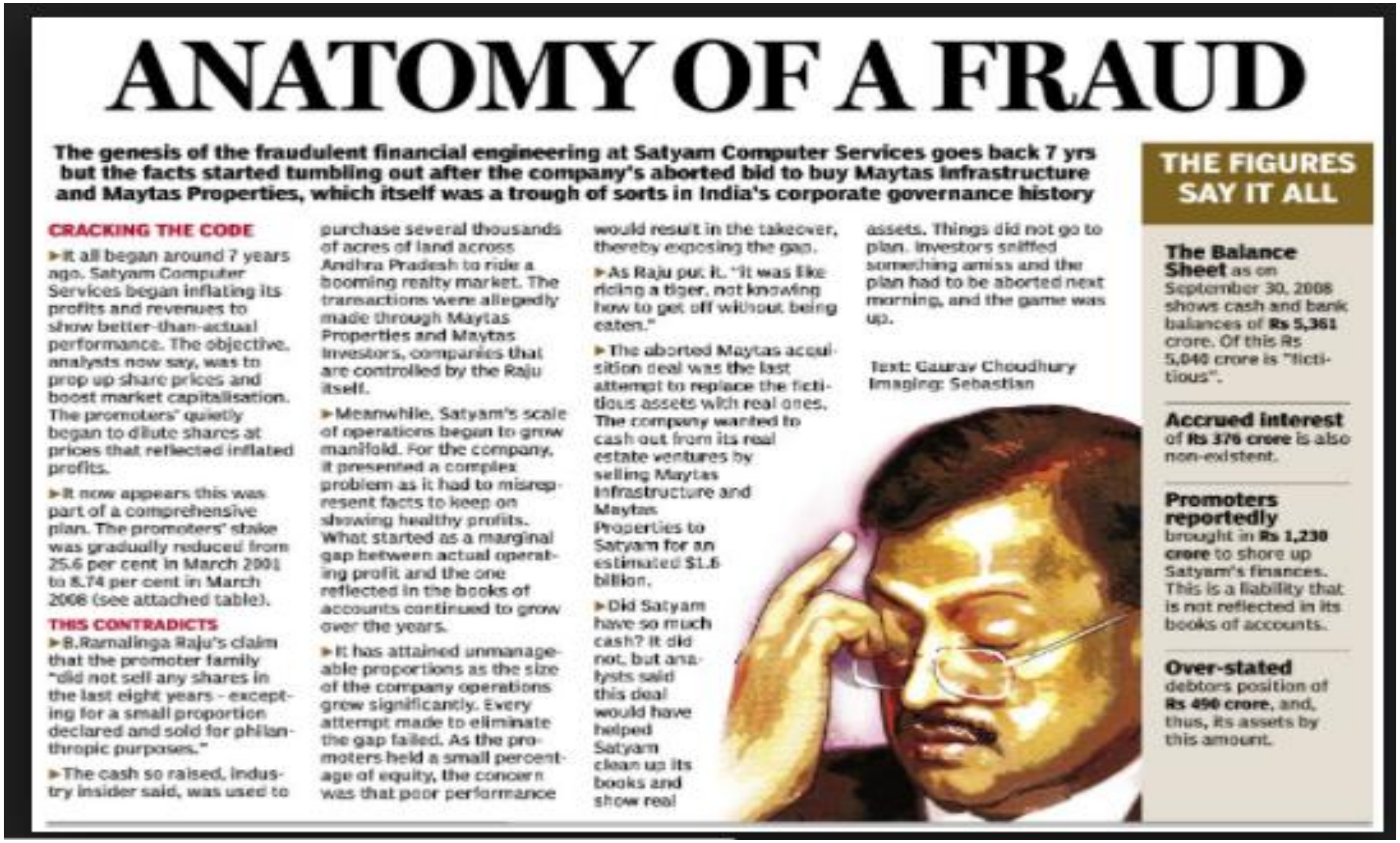


Credit, debit card frauds and how you can avoid them

By *Riju Dave*, ET Bureau | Oct 31, 2016, 11.25 AM IST

Even as the RBI and banks are introducing several security features, customers need to take the initiative to prevent being conned.

On 19 October, the country woke up to a banking nightmare. The State Bank of India (SBI) blocked 6 lakh debit cards after a reported malware-related breach in a non-SBI ATM network. In what is possibly India's largest financial data breach, nearly 32 lakh debit cards across 19 banks, including HDFC Bank, ICICI Bank and Axis Bank, were compromised. As per the National Payments Corporation of India (NPCI), 90 ATMs were impacted and at least 641 customers lost Rs 1.3 crore in fraudulent transactions.



ANATOMY OF A FRAUD

The genesis of the fraudulent financial engineering at Satyam Computer Services goes back 7 yrs but the facts started tumbling out after the company's aborted bid to buy Maytas Infrastructure and Maytas Properties, which itself was a trough of sorts in India's corporate governance history

CRACKING THE CODE

►It all began around 7 years ago. Satyam Computer Services began inflating its profits and revenues to show better-than-actual performance. The objective, analysts now say, was to prop up share prices and boost market capitalisation. The promoters' quietly began to dilute shares at prices that reflected inflated profits.

►It now appears this was part of a comprehensive plan. The promoters' stake was gradually reduced from 25.6 per cent in March 2001 to 8.74 per cent in March 2008 (see attached table).

THIS CONTRADICTS

►B.Ramalinga Raju's claim that the promoter family "did not sell any shares in the last eight years - excepting for a small proportion declared and sold for philanthropic purposes."

►The cash so raised, industry insider said, was used to purchase several thousands of acres of land across Andhra Pradesh to ride a booming realty market. The transactions were allegedly made through Maytas Properties and Maytas Investors, companies that are controlled by the Raju itself.

►Meanwhile, Satyam's scale of operations began to grow manifold. For the company, it presented a complex problem as it had to misrepresent facts to keep on showing healthy profits. What started as a marginal gap between actual operating profit and the one reflected in the books of accounts continued to grow over the years.

►It has attained unmanageable proportions as the size of the company operations grew significantly. Every attempt made to eliminate the gap failed. As the promoters held a small percentage of equity, the concern was that poor performance would result in the takeover, thereby exposing the gap.

►As Raju put it. "It was like riding a tiger, not knowing how to get off without being eaten."

►The aborted Maytas acquisition deal was the last attempt to replace the fictitious assets with real ones. The company wanted to cash out from its real estate ventures by selling Maytas Infrastructure and Maytas Properties to Satyam for an estimated $1.6 billion.

►Did Satyam have so much cash? It did not, but analysts said this deal would have helped Satyam clean up its books and show real assets. Things did not go to plan. Investors sniffed something amiss and the plan had to be aborted next morning, and the game was up.

Text: Gaurav Choudhury
Imaging: Sebastian

THE FIGURES SAY IT ALL

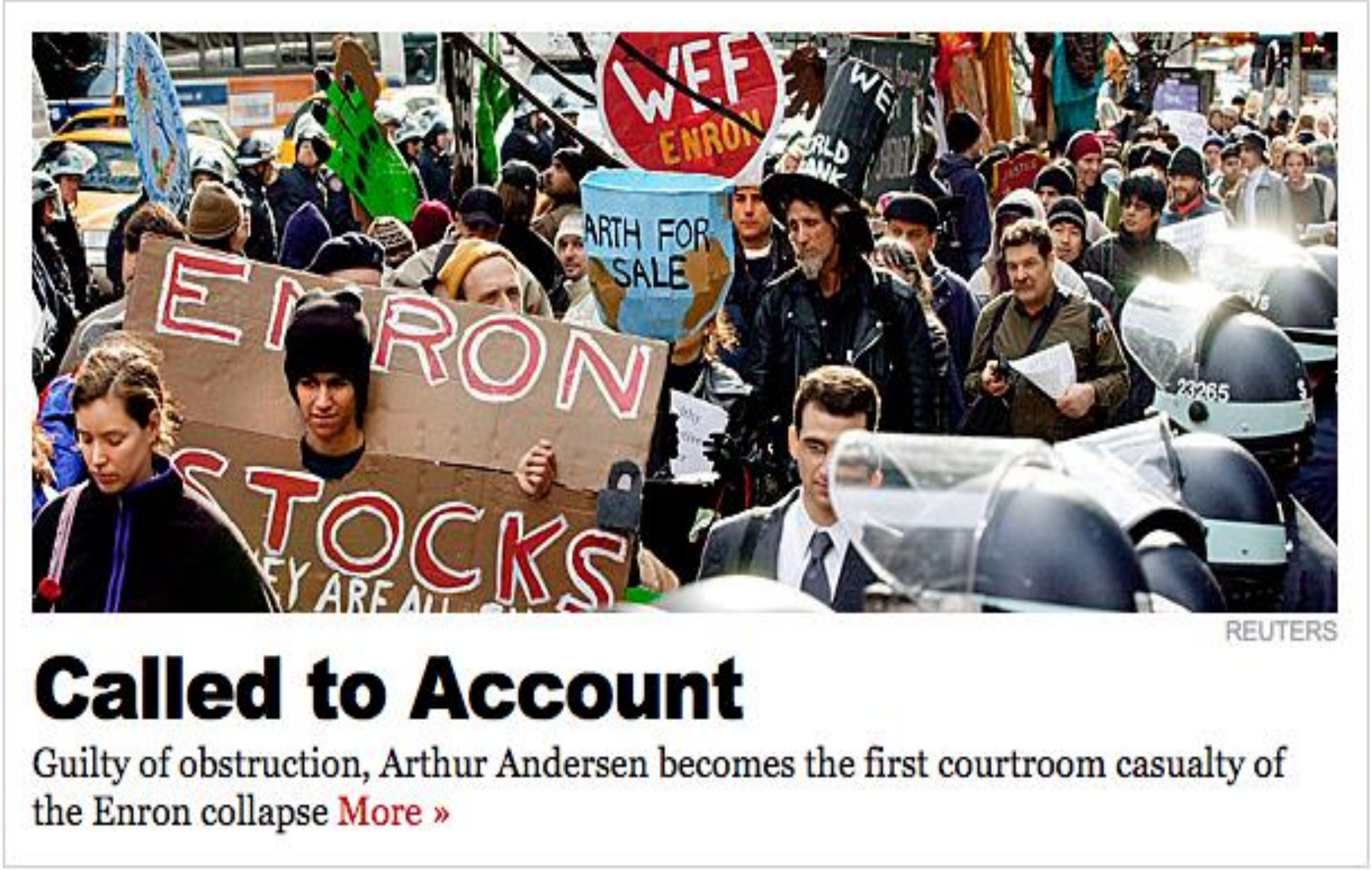
The Balance Sheet as on September 30, 2008 shows cash and bank balances of **Rs 5,361** crore. Of this Rs 5,040 crore is "fictitious".

Accrued interest of **Rs 376 crore** is also non-existent.

Promoters reportedly brought in **Rs 1,230 crore** to shore up Satyam's finances. This is a liability that is not reflected in its books of accounts.

Over-stated debtors position of **Rs 490 crore**, and, thus, its assets by this amount.

Behind the Enron Scandal



REUTERS

Called to Account

Guilty of obstruction, Arthur Andersen becomes the first courtroom casualty of the Enron collapse More »

Earnings Manipulations to detect company fraud

Approx. 6000 public listed companies in India and are obliged by the regulatory bodies to report financial health Quarter on Quarter.

Very less number of these companies manipulate financial statement.

Anomaly Detection

Eight financial ratios reported by the companies are:

- Days Sales to Receivables Index (DSRI)
- Gross Margin Index (GMI)
- Asset Quality Index (AQI)
- Sales Growth Index (SGI)
- Depreciation Index (DEPI)
- Sales General and Administrative Index (SGAI)
- Accruals to Total Assets (ACCR)
- Leverage Index (LEVI)

Can someone be identified as potential manipulator based on the financial ratios reported

Anomaly Detection–Intuition

A sample of 1239 companies with 1200 non manipulators and 39 manipulators with the following attributes/features in the data

- $x_1 = DSRI$
- $x_2 = GMI$
- $x_3 = AQI$
- $x_4 = SGI$
- $x_5 = DEPI$
- $x_6 = SGAI$
- $x_7 = ACCR$
- $x_8 = LEVI$

$$Dataset(x_1) = \{x^{(1)}, x^{(2)} \ldots . x^{(m)}\}$$

Build a model which gives the probability of observing the train data.

$$P(x) = ?$$

$Is\ x_{test}\ anomalous?$

$If\ P(x_{test}) < \in\ then\ flag\ anomaly$
$If\ P(x_{test}) \geq \in\ then\ flag\ ok$

Gaussian distribution fitting for detecting anomaly

Probability of observing the features are given by

$$P(X) = P(x_1; \mu_1, \sigma^2{}_1) * P(x_2; \mu_2, \sigma^2{}_2) \ldots . P(x_8; \mu_8, \sigma^2{}_8)$$

In General:

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

Steps in model

- Identify the features
  - company fraud there are 1239 observations across eight financial ratios
  - $1200\ (Y = 0;\ non\ manipulators)\ and\ 39\ (Y = 1;\ manipulators)$
  - $training\ set = \{x^{(1)}, x^{(2)}, \ldots .\, x^{(m)}\}\ where\ m = 900\ \text{obs. across eight ratios}$
  - $test\ set = \{(x^1{}_{test}, y^1{}_{test}), (x^2{}_{test}, y^2{}_{test}), \ldots .\,(x^m{}_{test}, y^m{}_{test})\}\ where\ m = 339\ \text{obs. across eight ratios (300 non manipulators and 39 manipulators)}$
- Compute mean and standard deviation for all the features in train set
- Given a test set compute

$$P(x) = \prod_{j=1}^{n} (1/\sqrt{2\pi}\,\sigma_j) * \exp\left(-\frac{(x_j - \mu_j)^2}{2\sigma_j{}^2}\right)\ where\ n\ is\ the\ number\ of\ feature$$

$$Y = \begin{cases} 1 & if\ P(x) < \in\ then\ anomaly \\ 0 & If\ P(x) \geq \in\ then\ ok \end{cases}$$

- Fine tune ∈ to catch anomaly

Error analysis and Non Gaussian features

Want P(x) large for normal examples and P(x) small for anomalous examples. If not,

Fine tune ∈ to improve the model in demarcating

Transform features to improve the model

What is the feature is not normally distributed?

Model may still work even if the feature is not normal

Apply transformations to make the feature normal

Anomaly detection vs. Supervised learning

Anomaly Detection

- Very small number of anomalous example (Y=1) and large number of non-anomalous example
- Hard for the algorithm to learn from restricted anomalous example.

Supervised learning

- Large number of positive and negative examples
- Enough learning opportunity for the model to learn from positive cases

Some other use cases on anomaly detection

Unusual behavior on the website based on the browsing behavior

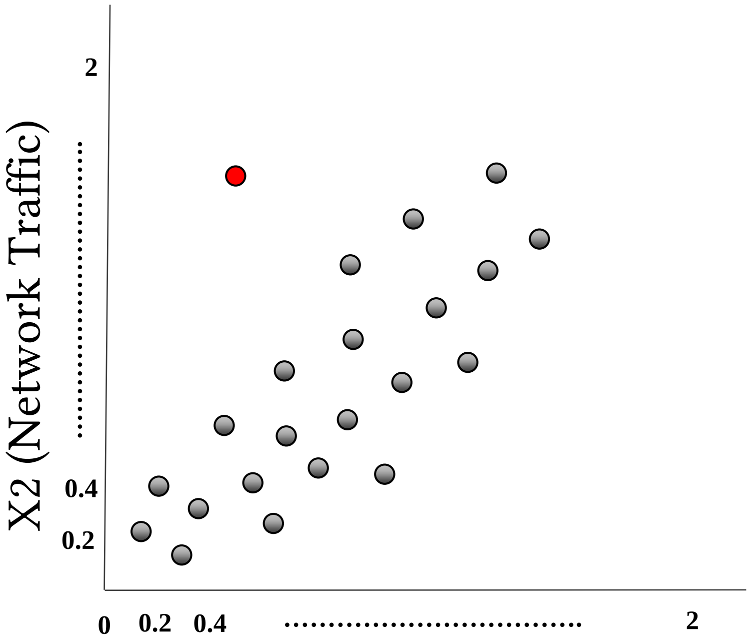
- $x_1 = No\ of\ logins$
- $x_2 = Typing\ speed$
- $x_3 = No\ of\ webpage\ visits$
- $x_4 = No\ of\ transactions$

Monitoring computers in a data center for suspicious usage

- $x_1 = Memory\ use$
- $x_2 = No\ of\ disc\ access\ per\ sec$
- $x_3 = CPU\ load$
- $x_4 = Network\ traffic$

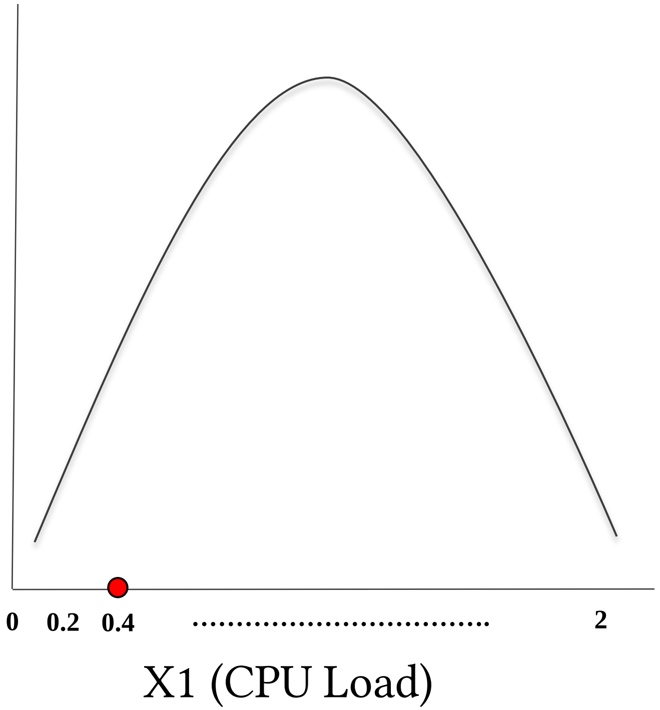
Multivariate Gaussian Distribution and Anomaly Detection

Multivariate Gaussian Intuition



Modelling P(X1) and P(X2) separately:

may show the point (0.4, 1.4) to have a significantly high probability of being not anomalous. Example with X1.



Multivariate Gaussian will model all the P(x) at one go.

Multivariate Gaussian

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

- $\mu = mean = \frac{1}{m} * \sum_{i=1}^{m} x^{(i)}\,;$
- $\Sigma = covariance\ matrix = \frac{1}{m} * \sum_{i=1}^{m}(x^{(i)} - \mu)(x^{(i)} - \mu)^T;$
- $|\Sigma| = determinant\ of\ the\ matrix;$
- $n\ is\ the\ number\ of\ features$
- $\mathrm{m}\ is\ the\ number\ of\ observations$

Steps in model building

- Fit model by setting

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

$$\Sigma = covariance\ matrix = \frac{1}{m} * \sum_{i=1}^{m} (x^{(i)} - \mu)(x^{(i)} - \mu)^T;$$

- Given a new test set compute

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

$$Y = \begin{Bmatrix} 1 & if\ P(x) < \in\ then\ anomaly \\ 0 & If\ P(x) \geq \in\ then\ ok \end{Bmatrix}$$

- Fine tune ∈ to catch anomaly

Gaussian vs. Multivariate Gaussian

Gaussian

- Manually create features which can capture anomaly
- Computationally cheaper
- Works fines with small number of observations as well (m)

Special case of multivariate Gaussian model where covariance matrix Σ has value of zero on the off diagonal element

Multivariate Gaussian

- Correlation between features is captured by the model automatically
- Computationally expensive
- Must have m>n

End of Lesson05–Gaussian Distribution in Anomaly Detection