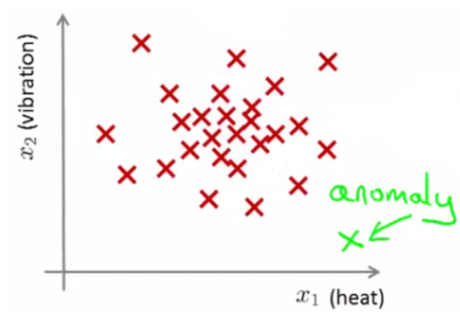
Week 15

August 18, 2021

Anomaly detection

- We can assess whether data points are anomalous by using the dataset as a baseline.
- if $p(x_{test}) < \epsilon$, then flag this as an anomaly
- if $p(x_{test}) \geq \epsilon$, then this is OK
- ϵ is a threshold probability number that we determine based on how certain we need/want to be.



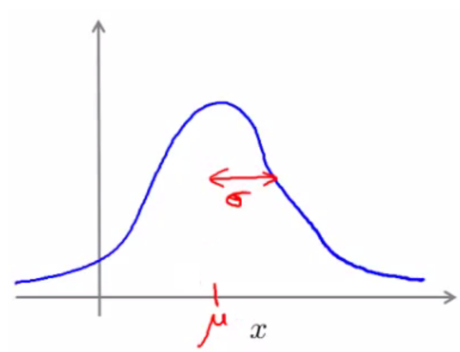
Applications

- Fraud detection
  - Users have activities connected with them, such as the amount of time spent online, the location of login, and the frequency with which they spend money.
  - Using this information, we can create a model of what regular users do.
  - What is the probability of "normal" behavior?
  - Send atypical users' data through the model to identify them. Make a note of everything that appears unusual. Block cards/transactions automatically.
- Manufacturing
  - Aircraft engine example.
- Monitoring computers in data center
  - If you have many machines in a cluster (x1 = memory use, x2 = number of disk accesses/sec, x3 = CPU load).
  - When you notice an anomalous machine, it is likely that it is soon to fail.
  - Consider replacing parts of it.

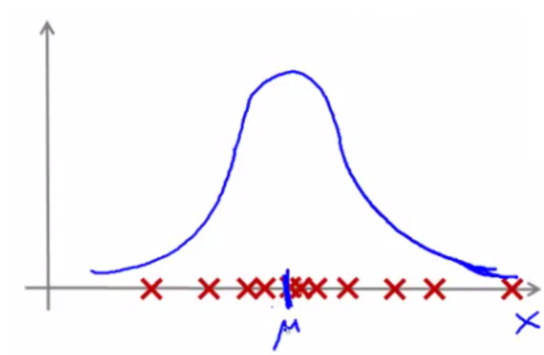
The Gaussian distribution

- μ is mean.
- σ^2 is variance and σ is a standard deviation.
- probability of x, parameterized by the mean and variance:

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



- Assume we have a data collection of m examples.
- Given that each example is a real number, we plot the data on the x axis.
- Given the dataset can you estimate the distribution?



Seems like a good fit - data suggests a higher likelihood of being in the center and a lower likelihood of being further out.

Anomaly detection

Algorithm 1 Anomaly detection

1: Choose features x_i that you think might be indicative of anomalous examples.

2: Fit parameters $\mu_1, ..., \mu_n, \sigma_1^2, ..., \sigma^n$

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

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

3: Given new example x, compute p(x):

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

Developing and evaluating and anomaly detection system

- You have some labeled data.
  - $y = 0$ for engines which were non-anomalous.
  - $y = 1$ for engines which were anomalous.
- Training set is the collection of normal examples.
- Next define:
  - Cross validation set.
  - Test set.
  - For both assume you can include a few examples which have anomalous examples.
- In our example we have:
  - 10000 good engines.
  - 50 flawed engines.
- Split into:

- Training set: 6000 good engines (y = 0).
- CV set: 2000 good engines, 10 anomalous.
- Test set: 2000 good engines, 10 anomalous.

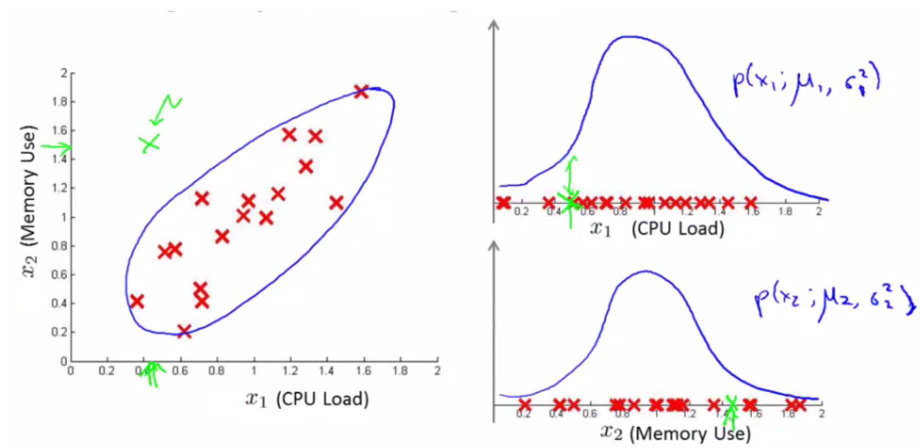
What's a good metric to use for evaluation?

- Compute fraction of true positives/false positive/false negative/true negative.
- Compute precision/recall.
- Compute F1-score.

Multivariate Gaussian distribution

It is a somewhat different approach that can occasionally discover anomalies that normal Gaussian distribution anomaly detection fails to detect.

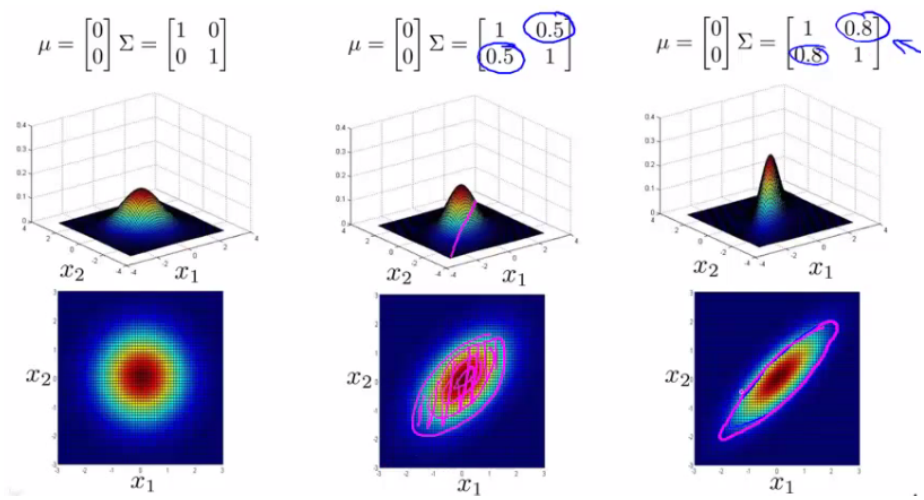
- Assume you can fit a Gaussian distribution to CPU load and memory use.
- Assume we have an example in the test set that appears to be an anomaly (e.g. x1 = 0.4, x2 = 1.5).
- Here memory use is high and CPU load is low (if we plot x1 vs. x2 our green example looks miles away from the others).
- The problem is that if we look at each characteristic individually, they may fall inside acceptable bounds - the difficulty is that we know we shouldn't obtain those types of numbers together, but they're both okay individually.



What are the parameters for this new model?

- μ which is an n dimensional vector (where n is number of features)
- Σ which is an [n x n] matrix - the covariance matrix

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



Very tall thin distribution, shows a strong positive correlation.

Gaussian model - summary

- Probably used more often.
- There is a need to manually create features to capture anomalies where x1 and x2 take unusual combinations of values.
- So need to make extra features and might not be obvious what they should be.
- Much cheaper computationally.
- Scales much better to very large feature vectors.
- Works well even with a small training set e.g. 50, 100.

Multivariate gaussian model - summary

- Used less frequently.
- Can capture feature correlation.

- So no need to create extra values.
- Less computationally efficient.
- Needs for m ¿ n i.e. number of examples must be greater than number of features.