VISA

PREDICTION OF SERVER/IT OUTAGE

PROBLEM STATEMENT: Methodologies to forecast server outage.

TECHNIQUES: Machine Learning Algorithms, Reliability Analysis

PROCEDURE:

① Identify METRICS and Events

- Metrics and Events are data from IT elements that we monito
- Metrics are numerical data, we can measure on the IT element directly like -
  - No. of SQL query requests on database applicatn
  - CPU & memory utilizatn on server.
- Events are fragments of information for indicating element's status. like :-
  - An event from Microsoft Windows server about an applicatn that has stopped
  - A log report indicating critical element has failed
- Capacity metrics would indicate that IT element is running low on capacity

(2) FEATURE ENGINEERING :

Feature engineering is extracting variables from domain knowledge that might be useful for prediction.

For example, if server log gives date 11/1/2016 we can extract that the day is a working day: Tuesday

Using available data, we extract further information

(3) TRAINING & TESTING DATA :

- Using the feature engineering, we expand our variables and create a dataset.
- Data set is further divided into training data & testing data
- Best known technique for defining training & testing data is CROSS VALIDATION
- Use 10 fold cross validation to define training & test data randomly from sample & repeat 10 times.
- We train our model & then test it on the testing data.

(4.) MACHINE LEARNING ALGORITHMS

(a) Logistic Regression

- Let p be the probability of failure

  Odds ratio $= \frac{p}{1-p}$

- We fit

$$\log\left(\frac{p}{1-p}\right) \sim \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots\ldots + \beta_p X_p$$

$X_i \rightarrow$ features

- Using training data, we fit the logistic model & find cross validation error by 10-fold cross validation technique

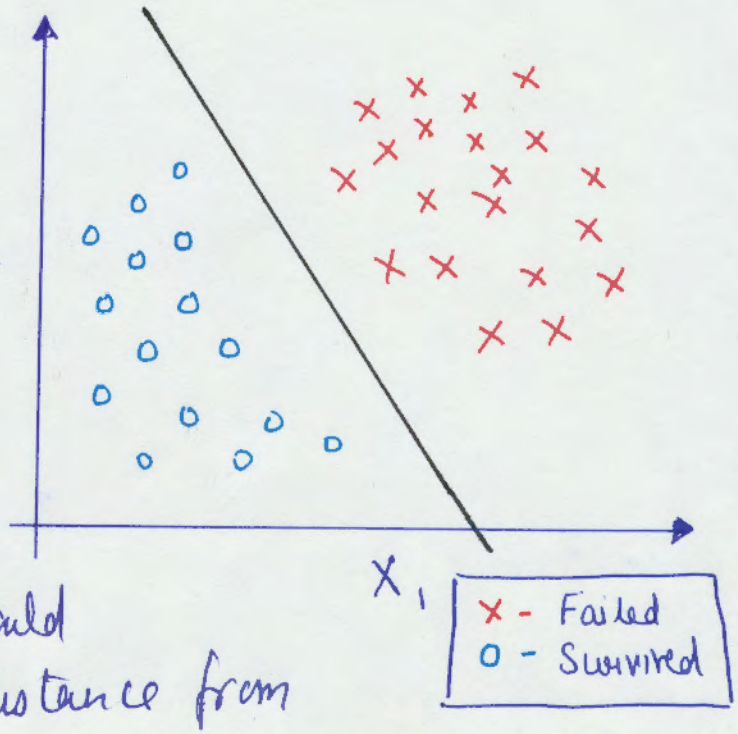
- DEFINE THRESHOLD : We have to define c as a threshold for prediction such that if $p \geq c$, the component is considered as failed and send alerts. Prediction threshold will have margin.

(b) Support Vector Machine :

Instead of finding the probability of failure, we classify the component as failed or not based on the features using Support Vector Machine algorithm.

SVM Example

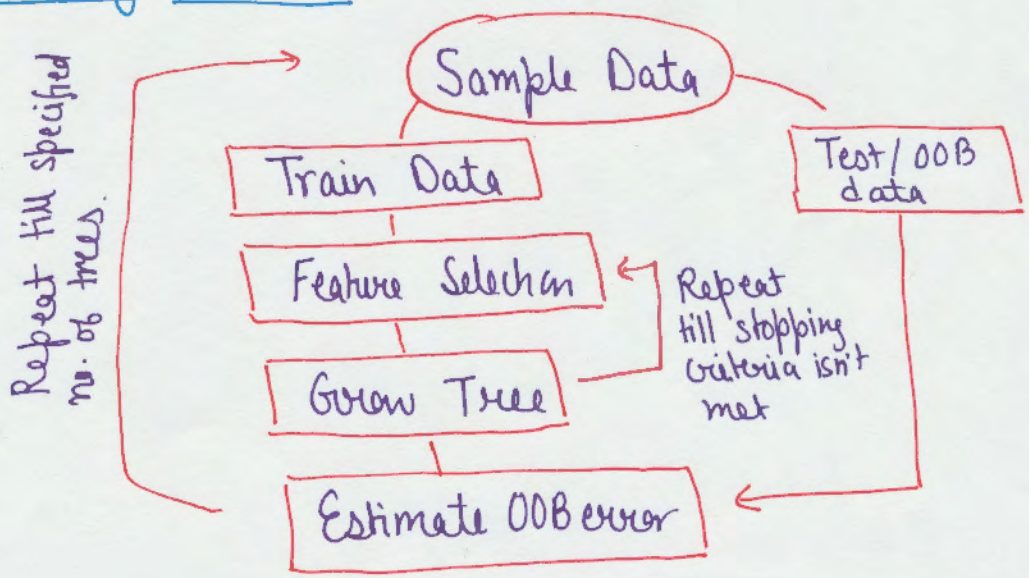
- If X_1 & X_2 are features & the data points are similar to graph on right side, SVM algorithm will define a line which separates the two. The slope & intercept of the line would be such it is at max. distance from point.



- In practice, number of features would be much greater than 2 & points won't be linearly separable, but using Kernels in SVM, we can find the optimum solution.

(C) Random Forest Algorithm

- Using Random forest Algorithm, we can classify the elements as reliable or failed.

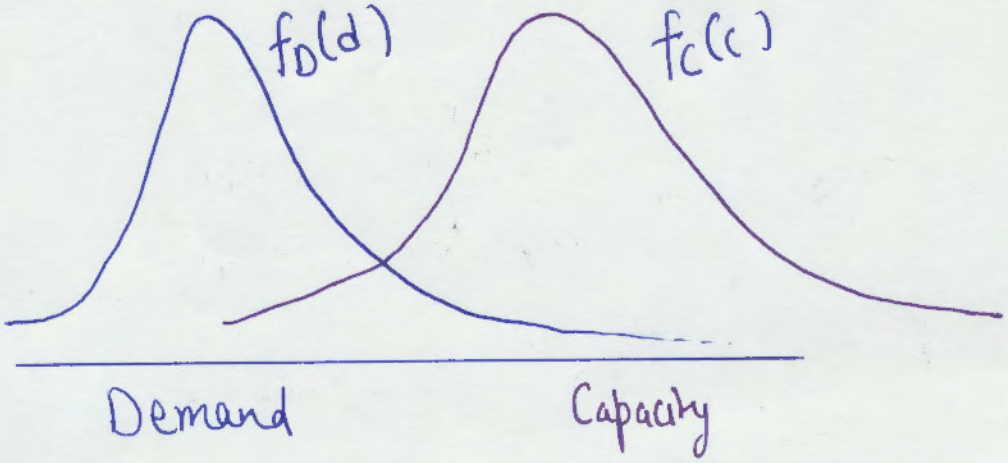


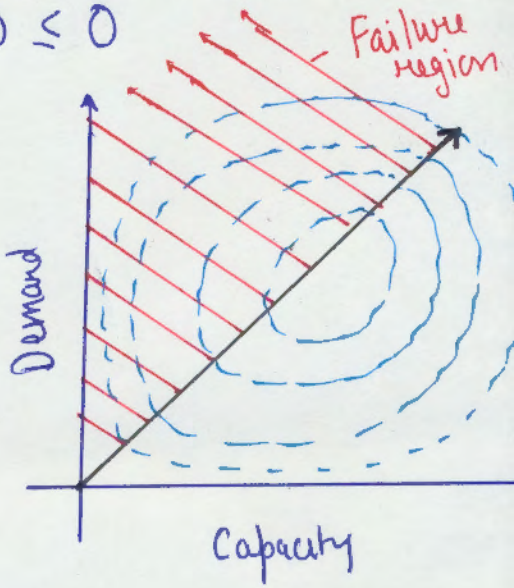
STOCHASTIC MODEL:

If features $X_1 \ldots X_p$ are not deterministic & follow a distribution, we develop a stochastic model to predict probability of failure using **Reliability Analysis**

Idea:

- Let Capacity $C \sim N(\mu_1, \sigma_1^2)$ & demand $D \sim N(\mu_2, \sigma_2^2)$ then failure occurs when $g(x) = C - D \leq 0$





$$P_f = \iint_{g(x)\leq 0} f_{CD}(c,d)\, dc\, dd$$

- Now if Capacity of element $C \sim f_1(X_1, X_2 \ldots X_p)$ & demand for element $D \sim f_2(X_1 \ldots X_p)$ we define $g(X_1 \ldots X_p)$ such that failure region is $g(X \ldots X_p) \leq$

- Therefore in p dimensions we find

$$P_f = \iint_{g(X_1 \ldots X_p)\leq 0} f_{CD}(\underline{X})\, d\underline{X} = P[g(\underline{X}) \leq 0]$$

$$f_{CD} = C(\underline{X}) - D(\underline{X})$$