Link Failure prediction and Localization in Cloud Scale Networks using Supervised Learning

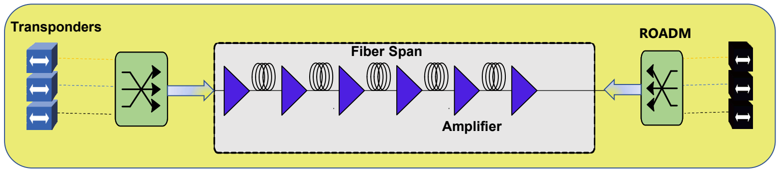
Zahra Bakhtiari
zahra22@stanford.edu
zahrabakhtiari807@gmail.com

Introduction

- Cloud service providers attracts customers all over the world by providing diverse types of services via millions of servers. The key to CSPs' success is the ability to provide services to customers with fast response and all-time availability. The physical infrastructure which supports cloud services is based on IP and optical networks which can fail and go out of service for variety of reasons.
- Proactive link health diagnosis is a game changer for cloud scale service providers. When a link failure happens, there are many risk scenarios including customer outage, long maintenance windows and complexity in terms of failure localization.
- In this project I aim at enabling intelligent, pro-active network management and operation using supervised learning classification methods to pro-actively detect links failures.
- By implementing this operational tool service providers can proactively reroute traffic when link failure is predicted and avoid customer outage.
- 6000 data points of two features each were used for classifiers training and model performance evaluation.
- SVM, GDA and logistic regression classifiers have been trained with 75% of dataset. models were all able to perform with high accuracy but logistic regression and linear SVM classifiers perform slightly better.

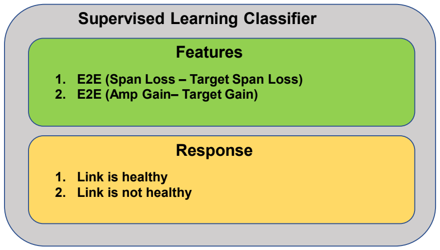
Problem Statement

As can be seen in the conceptual diagram of "Cloud Optical Infrastructure", an end to end (E2E) link is composed of multiple spans. Each span includes fiber links, ROADM and amplifiers. Each fiber span has a baseline/target span loss. The diff between this target value and real time span loss is an important feature in link health diagnosis. Each amplifier is also associated to a baseline/target gain value. The diff between this target value and real time gain value is another important feature in link health diagnosis. Note that the end-to-end span loss and amplifier gain diffs are used here to develop supervised learning models to proactively diagnose marginal or failing links from healthy ones and react to it in order to avoid customers outages.



Methodology Statement

- **Preprocessing:** This step includes data labeling, preparing the features in the diff format for an end-to-end link. It also includes managing empty data cells as well as differentiating data per traffic direction.
- **SVM Model** with linear kernel was used due to the linear nature of the decision boundary
- **Logistic Regression Model** was used as a linear classifier as well as a base line model to be compared against GDA and SVM performances



- **GDA Model** was used with the following joint distribution as a classifier due to its ability to classify data points with linear decision boundary and gaussian distribution:

$$p(y) = \begin{cases} \phi & \text{if } y = 1 \\ 1-\phi & \text{if } y = 0 \end{cases}$$

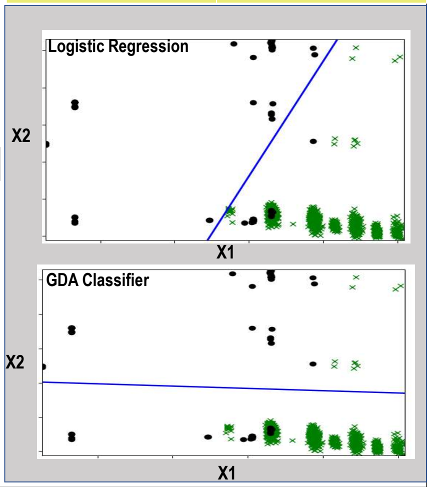
$$p(x|y=0) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)\right)$$

$$p(x|y=1) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right)$$

Key Results

- Logistic regression, GDA and SVM with linear kernel trained via 75% of data and their performances were measured on the test dataset (25% of data)

Classifier	Test Accuracy
Logistic Regression	99%
GDA	97%
Linear SVM	99%



Summary

- We found that logistic regression and SVM with linear kernel perform with high accuracy due to the linear nature of the decision boundary. GDA performance comes second due to the nature of the dataset.
- Due to the linear nature of the decision boundary and high accuracy of linear classifiers there was no need to implement SVM with nonlinear kernel.

- 3-minutes video Link:
- **Please go to this link:**

https://youtu.be/_pooOgddT5k