PHISHING WEBSITE DETECTION

Syed Hasan Raza and Burak Mandira
Department of Electrical and Electronics Engineering, Bilkent University
Term Project for EEE 485/585, Statistical Learning & Data Analytics

Project Description

- What is a phishing Website?
- Important aspect of Cyber Security.
- Binary Classification Problem.

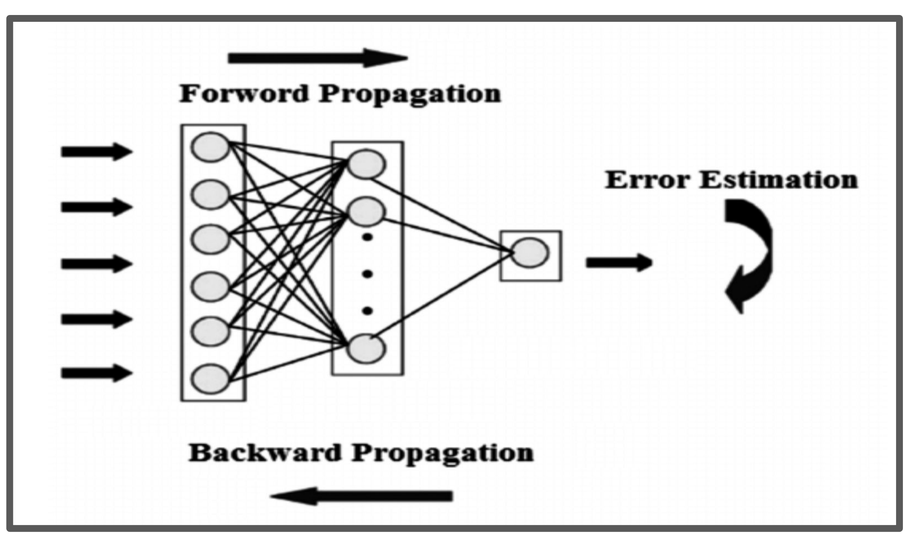
Dataset Description

- 11055 instances with 30 features each.
- Almost Balanced. (56% phishing, 44% not phishing)
- Features categorized into Address bar based, Domain based.

Methods

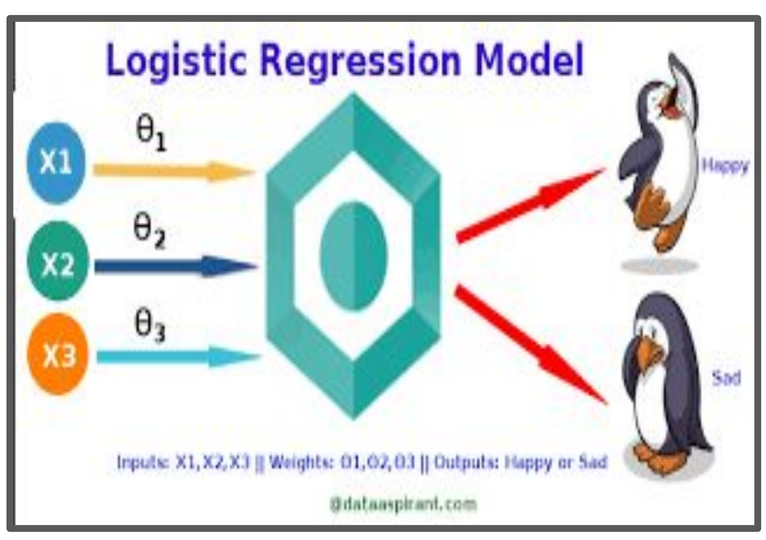
Neural Network

- Define the architecture and initialise weights.
- Propagate forwards to get the output.
- Cross entropy cost function to calculate error.
- Propagate error backwards.
- Use Gradient Descent to update weights.
- Iterate until convergence.



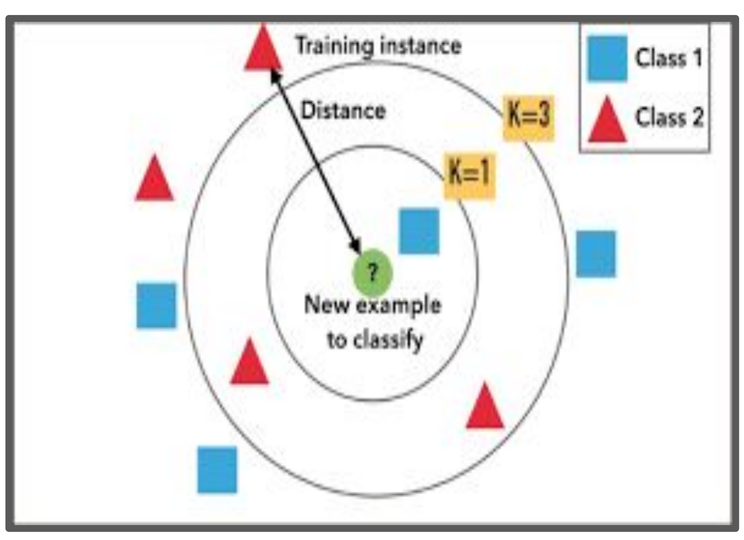
Logistic Regression

- Find weights which maximise the MLE equation.
- No closed form solution for transcendental equation.
- Use Gradient Ascent iteratively to learn weights.
- Step function at output of Logistic function to classify.



K-Nearest Neighbors

- If it looks like a duck, swims like a duck, and quacks like a duck, then it probably is a duck.
- No separate learning for algorithm.
- Distance measure to determine k nearest neighbors.
- Find k using Cross validation.



Results

Model	*Training Accuracy*	*Validation Accuracy*	*Test Accuracy*	*Training Time (seconds)*
k-NN	NA	0.9559	0.9553	330
Logistic	0.9267	0.9686	0.9675	145
Neural Nets	0.9862	0.9264	0.9107	22

References

[1] D. Dua and E. Karra Taniskidou, "UCI Machine Learning Repository", *University of California, School of Information and Computer Science*, 2017. [Online]. Available: http://archive.ics.uci.edu/ml. [Accessed: 19- Feb- 2019].

[2] R. Mohammad, L. McCluskey and F. Thabtah, "Phishing Websites Data Set", *UCI Machine Learning Repository*, 2015. [Online].Available:https://archive.ics.uci.edu/ml/datasets/Phishing+Websites. [Accessed: 19- Feb- 2019].