Class 8 Review Notes

AI & Machine Learning
Fall 2023 - Laurie Ye

8 Introduction to the UMI "Spambase" Dataset and the Idea of Features

8.1 Spambase Datasets Introduction

- The dataset is derived from 4601 e-mail messages that have been classified (by human inspection) as spam or not spam ("ham").
- Spam is defined as "unsolicited commercial e-mail."
- These e-mails are from Hewlett-Packard and collected around 1999.

8.2 Feature

- For these words, the value of a "feature" based on a target word (e.g., "free") is the percent of the words in the e-mail that match the target word:

$$\text{Feature Value} = \frac{\text{\# of target words in the e-mail}}{\text{total \# of words in the e-mail}} \times 100$$

- Some features are based on the percentage of the characters in the e-mail that match a specific character like ";".

$$\text{Feature Value} = \frac{\text{\# of target characters in the e-mail}}{\text{total \# of characters in the e-mail}} \times 100$$

- A few features are based on the run lengths of consecutive capital letters.
  - Average run length
  - Maximum run length
  - Total run length = sum of the run lengths of all capital letter runs.
- So, all of the variables (features) are continuous except for 2 which are counts.

Note 1. Selecting the right features can be critical for the success of many machine learning approaches. Feature development generally requires subject-matter knowledge.

Note 2. Developing the right features from unstructured data is so important that it has its own name:"feature engineering".

8.3 Reference

Goizueta Business School-Emory University: Professor George S. Easton