WASHINGTON STATE UNIVERSITY

RNA Repeat Data SSR of AM

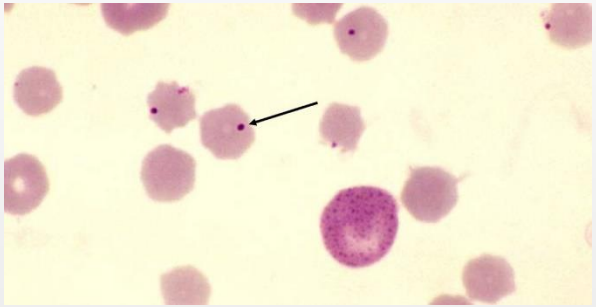
Aditya Jasuja
Minjian Li
CPTS_575 Data Science

Introduction

Anaplasma Marginale(AM)

- Blood parasite in cattle
  - Losing weight
  - Jaundice
  - Pyrexia
  - Anemia

Blood from a cow with anaplasmosis.



(From https://news.okstate.edu/articles/veterinary-medicine/2018/five-things-you-should-know-about-anaplasmosis-this-fall.html)

- Causing great impact of cattle production

Feature Extraction

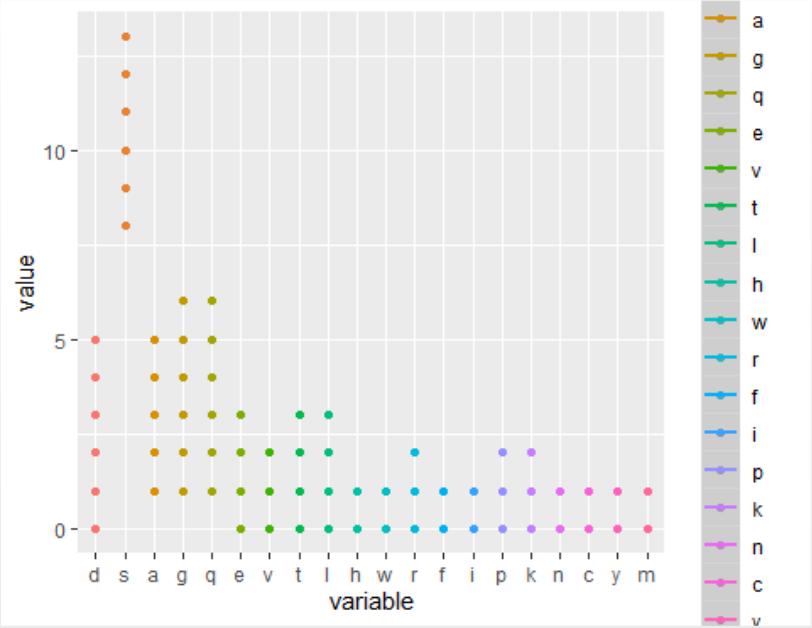
Approaches

- Bag-of-Words
- N-gram Model
- Longest Common Subsequence
- Alphabetical Counter
- Length-based
- K-mers

Feature Extraction

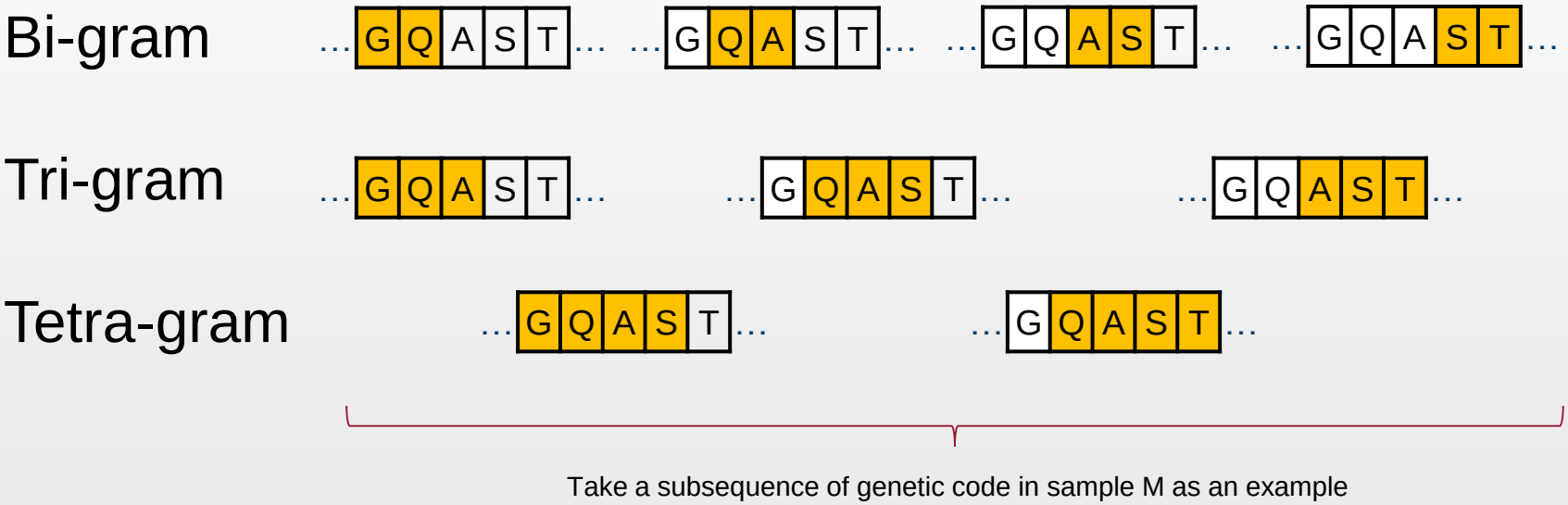
- **Bag-of-Words**

  Bag of words simply breaks apart the words into individual words count statistics



Feature Extraction

- **N-gram Model**



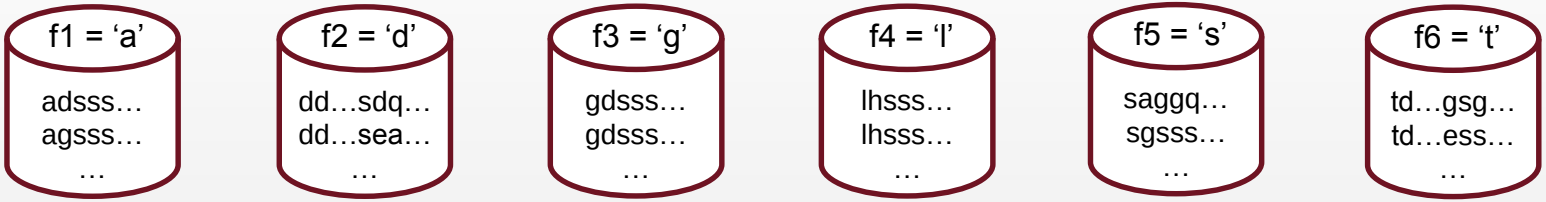
Feature Extraction

- **Longest Common Subsequence**

  It is a sequence that can be derived from another sequence by deleting some elements without changing the order of the remaining elements.

Feature Extraction

- **Alphabetical Counter**

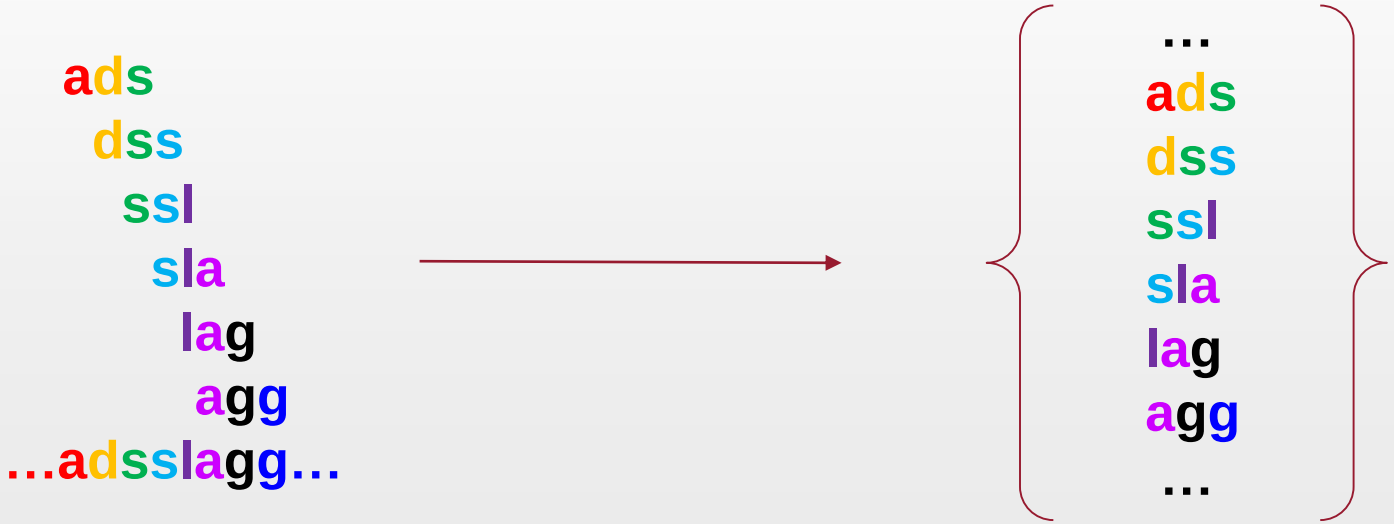


- **Length-based**

Counting the length of each sequence, and use it as features

Feature Extraction

- K-mers (K = 3)



Feature Selection

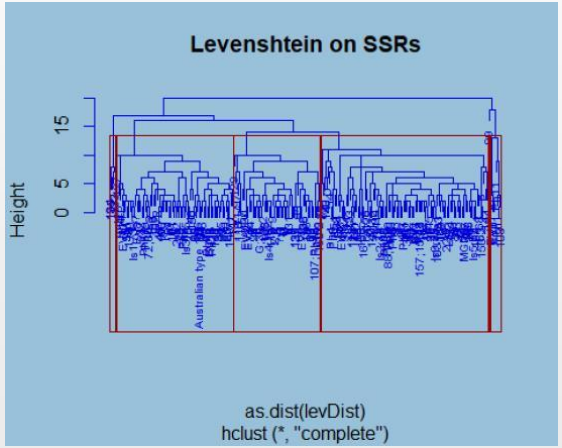
Approaches

- Zero- and Near Zero-Variance Predictors
- Removing Highly Correlated Predictors
- Boruta

Classification

Labels

- To define labels via hierarchical clustering.



Statistics by Class:

	Class: 1	Class: 2	Class: 3	Class: 4	Class: 5	Class: 6
Sensitivity	0.8333	1.0000	0.9333	NA	NA	NA
Specificity	0.9787	0.9524	0.9655	1	1	1
Pos Pred Value	0.9091	0.8947	0.9655	NA	NA	NA
Neg Pred Value	0.9583	1.0000	0.9333	NA	NA	NA
Prevalence	0.2034	0.2881	0.5085	0	0	0
Detection Rate	0.1695	0.2881	0.4746	0	0	0

(From https://bigishere.wordpress.com/2018/09/22/random-forest-introduction/)

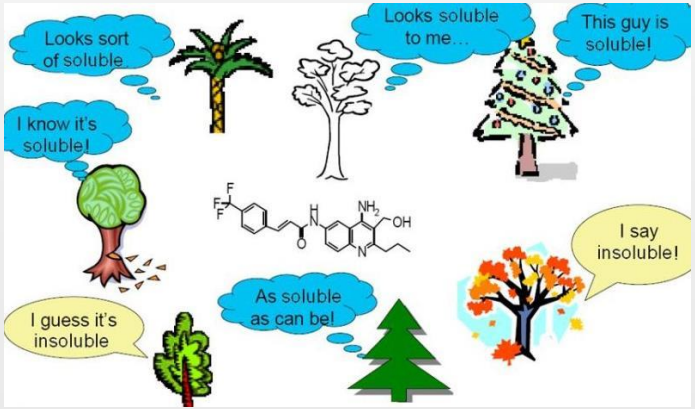
Classification

Classifiers

- Naïve Bayes
  - Stability
  - Conditional Independent

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

- Random Forest
  - Reduce Overfitting
  - Balance Error for Imbalanced Datasets
  - Fast-Learn



(From https://bigishere.wordpress.com/2018/09/22/random-forest-introduction/)

Classification

Results

Naïve Bayes

```
Overall Statistics

               Accuracy : 0.9322
                 95% CI : (0.8354, 0.9812)
    No Information Rate : 0.5085
    P-Value [Acc > NIR] : 2.003e-12

                  Kappa : 0.8905

 Mcnemar's Test P-Value : NA
```

Random Forest

```
Overall Statistics

               Accuracy : 0.8644
                 95% CI : (0.7502, 0.9396)
    No Information Rate : 0.5085
    P-Value [Acc > NIR] : 9.358e-09

                  Kappa : 0.7763

 Mcnemar's Test P-Value : NA
```

Thank you!