

Statistical Methods in Natural Language Processing (NLP)



*Class 8: Introduction to Classification and Machine
Learning*

Charalambos (Haris) Themistocleous

*Department of Philosophy, Linguistics and Theory
of Science, Centre for Linguistic Theory and Studies
in Probability*

Introduction

- ▶ Introduction to Machine Learning
- ▶ Introduction to Classification

Classification

- ▶ Given a speech sound is the speaker happy or unhappy?
- ▶ Given a sound, what are the consonants?
- ▶ Which genre does a text belongs to?
- ▶ Is the “Batrachomyomachia” a text written by Homer?



Types of Learning

1. Supervised
2. Unsupervised

Classification

- ▶ Binary classification (setting the boundaries between two categories: Decision Boundaries)
- ▶ Multiclass Classification

Data Splitting

- ▶ Train - Model Building: Teaching the algorithm
- ▶ Test - Evaluation:
 - ▶ Testing if the algorithm has learned.
 - ▶ How? By providing new data - unknown to the algorithm and asking the algorithm to evaluate the data.
 - ▶ How much data should be allocated to the training and test sets?

Pre-Processing

- ▶ feature selection
- ▶ Predictors: Are the variables that we use to train the algorithm.

Understanding-Observing the data

- ▶ Descriptive statistics
- ▶ Visualizations

Transformations of the data

- ▶ Transformations can improve the performance of the classifier.
- ▶ Skewness
- ▶ Outliers
- ▶ Missing Values

Transformations of the data: Centering

- ▶ The mean of a predictor is subtracted from all the values.
- ▶ The predictor will have a zero mean.

Transformations of the data: Scaling

- ▶ Each value of the predictor is divided by its standard deviation.
- ▶ Scaling results in data values with standard deviation of one.

Outliers

- ▶ plotting the data
- ▶ we need to consider all the options before we attempt to remove any data from the dataset.

Missing Values

- ▶ Missing values are informative by themselves if we understand why they are missing.
- ▶ Evaluations in sites like TripAdvisor are most probably done by people who have a strong opinion about it. So, most people who visit a place do not evaluate it.
- ▶ Python and R provide many ways to deal with missing data.
- ▶ Some algorithms do not have a problem with the missing data like the C5.0

Removing Predictors

- ▶ Having too many predictors is not always good.
- ▶ If there are highly correlated there is no point of having them.
- ▶ There is less complexity and the analysis takes less time.
- ▶ Some predictors are not always good, which can reduce the performance of the model.
- ▶ A near zero variance predictor, i.e., a predictor with that has a single value can create problems for regressions but not for C5.0.

Collinearity

- ▶ Collinearity: a pair of predictors has substantial correlation with each other.
- ▶ Multicollinearity: correlations between multiple predictors.

Dummy variables

- ▶ Creating contrasts with predictors that have many levels

Age	n	< 20	21-25	26-30	31-40
< 20	30	1	0	0	0
21-25	56	0	1	0	0
26-30	23	0	0	1	0
31-40	255	0	0	0	1

Table: Age of children in months.

Tuning the model

- ▶ resampling
- ▶ variable importance estimation

Model Selection

- ▶ Comparing different algorithms
- ▶ Comparing the same algorithm but with different tunings.

Next Class

- ▶ Assignment 1 with Mehdi.
- ▶ Task to write a Naive Bayes classifier for authorship attribution.