Notes on Natural Language Processing by Eisenstein

compiled by D.Gueorguiev 9/21/2024

# Linear Text Classification

**Definition**: *Text Classification Problem*

Given a text document, assign a label from a set of discrete labels.

## The Bag of Words

Question: how do we represent a document with text?

Use a column vector of word counts:

where is count for the -th word in a vocabulary . Here the size of the vocabulary is denoted with .

In linear classification the decision is based on a weighted sum of individual word counts where the word set is the feature set of the classification problem. The classification object is the vector ; this object is often called *a bag of words*. With a bag of words representation we are ignoring everything else but the frequency count of each word – we are not accounting for grammatical and syntactic constructs, sentence boundaries, paragraphs.

To predict a label from a bag of words we assign a score to each word in the vocabulary measuring fitness of the word with this label. These word scores are known as weights and are stored in column vector .

Let us consider multi-class classifier where The goal is to predict a label , given the bag of words , using weights . For each label , we compute a score , which is a scalar measure of the compatibility between the bag-of-words and the label . In a linear bag-of-words classifier, this score is the vector inner product between the weights and the output of a *feature function*

(2)

For example, given arguments x and y, element j