# ML INFOSEC
# 8: Cross Validation, Bag of words, Feature hashing

July 2, 2019

# Cross validation

to be completed

# Bag of words for feature generation I

Let $s = [w1, ...., w_N]$ be list of words from a set $W$. Let $W_s = \{w_1, ..., w_N\}$ be the set of all words appearing in $s$. Then $BoW(s) : W_s \to \mathbb{N}_0$ with

$$BoW(s)(w) = \#\{j \mid w_j = w\},$$

which can be extended to a map $BoW(s) : W \to \mathbb{N}_0$ by setting

$$BoW(s)(w) = 0 \quad if \quad w \notin W_s,$$

is called the **bag of words** of $s$.

# Bag of words for feature generation II

BOW is often encoded as dictionary:

**Example**

- Text = John likes to watch movies. Mary likes movies too.
- s = ["John","likes","to","watch","movies", "Mary","likes","movies","too"]
- BoW = {"John":1,"likes":2,"to":1,"watch":1,"movies":2, "Mary":1,"too":1}

# Feature hashing

Let $S$ a set of samples (e.g. documents, emails, websites, executable files), $A$ a set of attributes (e.g. words), $v_a(s)$ be the (real, non-negative) value of the attribute $a$ for the sample $s$. If $A$ is very large, one can use the **hashing trick**. Let

$$H : A \to Y$$

be a hash function with $\#Y$ small. Then we use $Y$ as set of attributes where we define the value of an attribute $y$ for a sample $s$ by

$$\hat{v}_y(s) = \sum_{a:H(a)=y} v_a(s).$$