Vector space classification (Chapter 14)

Algorithm 1 (Rocchio classification)

```
function TRAIN-ROCCHIO($\mathbb{C}, \mathbb{D}$)
    for all $c_j \in \mathbb{C}$ do
        $D_j \leftarrow \{d : \langle d, c_j \rangle \in \mathbb{D}\}$
        $\vec{\mu_j} \leftarrow \frac{1}{|D_j|} \sum_{d \in D_j} \vec{v}(d)$
    end for
    return $\{\vec{\mu_1}, \ldots, \vec{\mu_J}\}$
end function

function APPLY-ROCCHIO($\{\vec{\mu_1}, \ldots, \vec{\mu_J}\}, d$)
    return $\arg\min_j |\vec{\mu_j} - \vec{v}(d)|$
end function
```

Algorithm 2 (*k* nearest neighbor classification)

```
function TRAIN-kNN($\mathbb{C}, \mathbb{D}$)
    $\mathbb{D}' \leftarrow$ PREPROCESS($\mathbb{D}$)
    $k \leftarrow$ SELECT-K($\mathbb{C}, \mathbb{D}'$)
    return $\mathbb{D}', k$
end function

function APPLY-kNN($\mathbb{C}, \mathbb{D}', k, d$)
    $S_k \leftarrow$ COMPUTENEARESTNEIGHBORS($\mathbb{D}', k, d$)
    for all $c_j \in \mathbb{C}$ do
        $p_j \leftarrow |S_k \cap c_j|/k$
    end for
    return $\arg\max_j p_j$
end function
```

Exercise 14/1

What is the contiguity hypothesis?

Exercise 14/2

Discuss the main idea behind the Rocchio classification. How is Rocchio classification different to our linear classifier from exercises 13/3 and 13/4 in the previous seminar?

Exercise 14/3

Discuss the main idea behind the k Nearest Neighbor (kNN) classification. How large k (how many neighbors) should we use?

Exercise 14/4

Build Rocchio and 1NN classifiers for the training set $\{([1,1],1), ([2,0],1), ([2,3],2)\}$ and classify the document $q = [1,2]$. Do the classifiers agree?