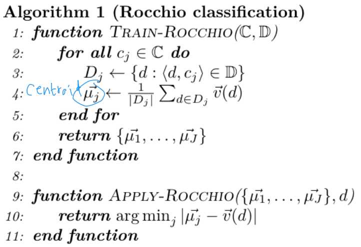
https://miro.com/app/board/o9J_lGkE_Tc=/

Vector Space Classification (Chapter 14)

14/1

What is the contiguity hypothesis?



Algorithm 1 (Rocchio classification)

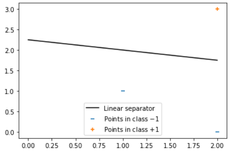
```
function TRAIN-ROCCHIO($\mathbb{C}, \mathbb{D}$)
    for all $c_j \in \mathbb{C}$ do
        $D_j \leftarrow \{d : \langle d, c_j \rangle \in \mathbb{D}\}$
        Centroid $\vec{\mu}_j \leftarrow \frac{1}{|D_j|} \sum_{d \in D_j} \vec{v}(d)$
    end for
    return $\{\vec{\mu}_1, \ldots, \vec{\mu}_J\}$
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
```

14/2

Discuss the main idea behind the Rocchio classification.

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?



$\vec{\mu}_1 = \left[\frac{3}{2} ; \frac{1}{2}\right] = [1,5 ; 0,5]$

Rocchio: $q \rightarrow$ class 2

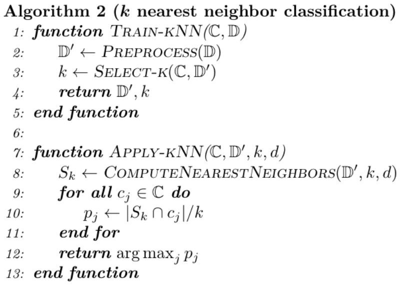
$|\vec{\mu}_1 - \vec{q}| = \sqrt{(1,5-1)^2 + (0,5-2)^2} = \sqrt{2,5}$

$|\vec{\mu}_2 - \vec{q}| = \ldots = \sqrt{2}$

14/3

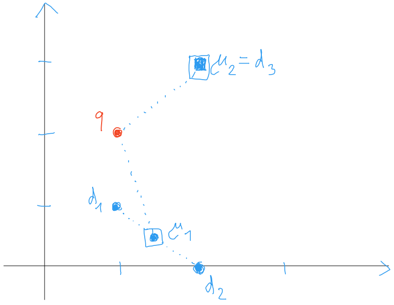
Discuss the main idea behind the k-Nearest Neighbor (kNN) classification. How large k (how many neighbors) should we use?

k .. odd



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
```

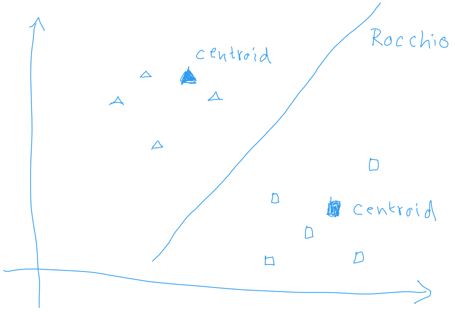


1NN: $k=1$, $\rightarrow$ class 1

$|\vec{d}_1 - \vec{q}| = \ldots = \sqrt{1}$

$|\vec{d}_2 - \vec{q}| = \ldots = \sqrt{5}$

$|\vec{d}_3 - \vec{q}| = \ldots = \sqrt{2}$



centroid
Rocchio
centroid