

Vector Space Classification (Chapter 14)

14/1

What is the contiguity hypothesis?

Algorithm 1 (Rocchio classification)

- 1: **function** TRAIN-ROCCHIO(\mathbb{C}, \mathbb{D})
- 2: **for all** $c_j \in \mathbb{C}$ **do**
- 3: $D_j \leftarrow \{d : \langle d, c_j \rangle \in \mathbb{D}\}$
- 4: $\vec{\mu}_j \leftarrow \frac{1}{|D_j|} \sum_{d \in D_j} \vec{v}(d)$
- 5: **end for**
- 6: **return** $\{\vec{\mu}_1, \dots, \vec{\mu}_J\}$
- 7: **end function**
- 8:
- 9: **function** APPLY-ROCCHIO($\{\vec{\mu}_1, \dots, \vec{\mu}_J\}, d$)
- 10: **return** $\arg \min_j |\vec{\mu}_j - \vec{v}(d)|$
- 11: **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 $\mathbf{q} = [1, 2]$. Do the classifiers agree?

Handwritten: $k=1$

14/3

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

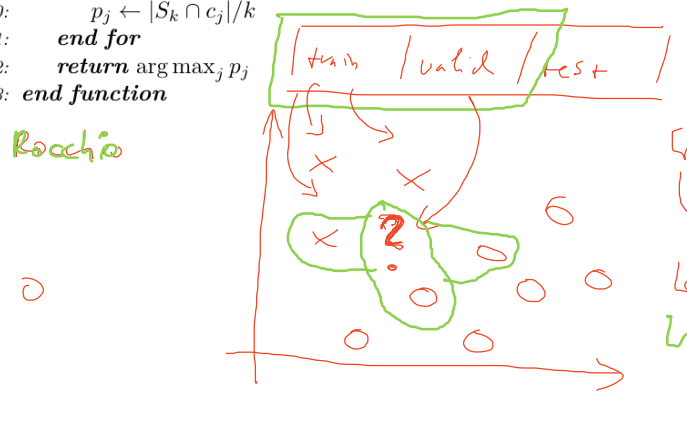
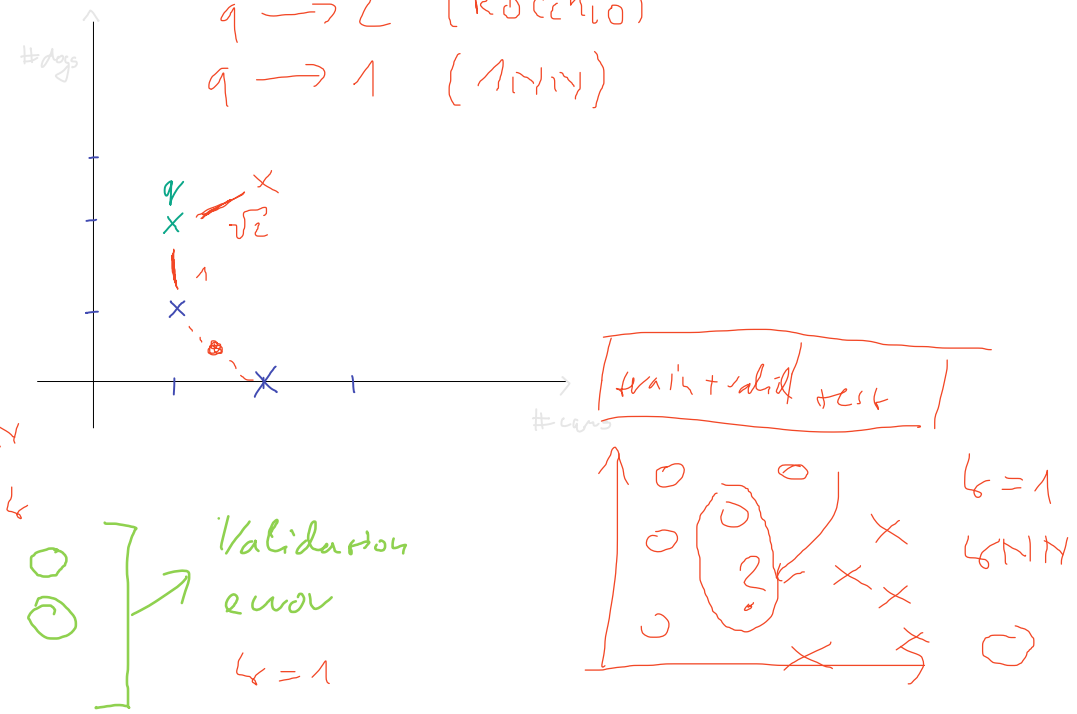
Algorithm 2 (k nearest neighbor classification)

- 1: **function** TRAIN-KNN(\mathbb{C}, \mathbb{D})
- 2: $\mathbb{D}' \leftarrow \text{PREPROCESS}(\mathbb{D})$
- 3: $k \leftarrow \text{SELECT-K}(\mathbb{C}, \mathbb{D}')$
- 4: **return** \mathbb{D}', k
- 5: **end function**
- 6:
- 7: **function** APPLY-KNN($\mathbb{C}, \mathbb{D}', k, d$)
- 8: $S_k \leftarrow \text{COMPUTENEARESTNEIGHBORS}(\mathbb{D}', k, d)$
- 9: **for all** $c_j \in \mathbb{C}$ **do**
- 10: $p_j \leftarrow |S_k \cap c_j| / k$
- 11: **end for**
- 12: **return** $\arg \max_j p_j$
- 13: **end function**

Handwritten:

$q \rightarrow 2$ (Rocchio)

$q \rightarrow 1$ (1NN)



Handwritten:

kNN

$k=1$

$k=3$

\vdots

Validation error

$k=1$

