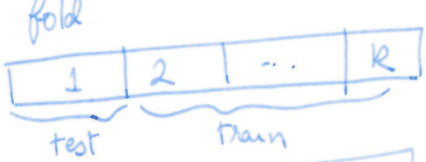
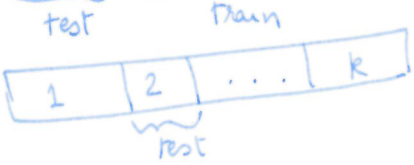
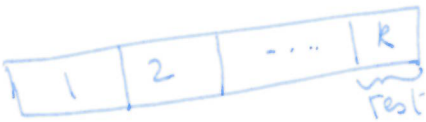
L 16: 15.07.2022

Experiment	fold	error rate
1		ER_1
2		ER_2
k		ER_k

5 fold or 10 fold is very common.

- Use k-1 fold for training, 1 fold for testing
- $ER = \frac{1}{k}\sum_{i=1}^{k} ER_i$

Important: all classes $\omega_1 \dots \omega_c$ well represented in all folds.

(B) Training, Validation, Test set

↓

hyperparameter optimization

4-110

4-111

5 Unsupervised Learning

5 - 1

5 - 2

5 - 3

5 - 4

5 - 5

5.1 Clustering for a known number of clusters

5.1.1 Problem formulations

Given:

- N samples $S = \{ \underline{x}_n \in \mathbb{R}^d, \; 1 \leq n \leq N \}$
- number of clusters C

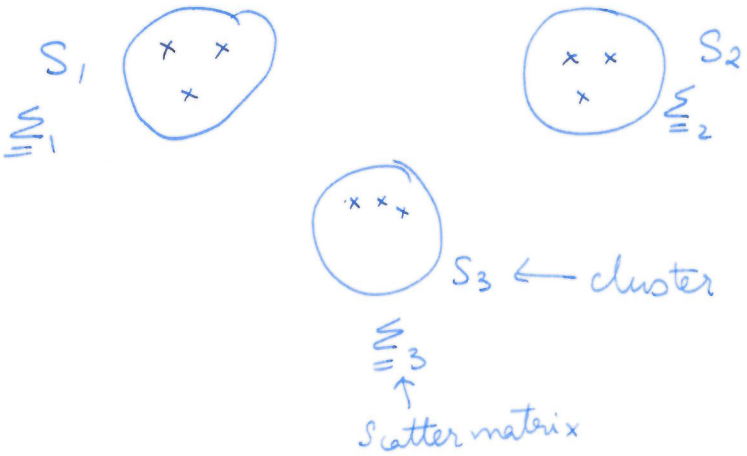
Wanted:

- C clusters $S_i \subset S$; $1 \leq i \leq C$ $N_i = |S_i|$ samples in S_i
- each sample exactly in one cluster
  i.e. $\sum N_i = N$

Wishes:

- Samples of one cluster stay "together"
- Samples of different clusters stay "far away"

Measure for "closeness": "scatter matrix ~ covariance matrix



(A) For each cluster S_i:

center $\underline{\mu}_i = \frac{1}{N_i} \sum_{\underline{x}_i \in S_i} \underline{x}_i$

scatter matrix $\underline{\underline{\Sigma}}_i = N_i \underbrace{\frac{1}{N_i} \sum_{\underline{x}_n \in S_i} (\underline{x}_n - \underline{\mu}_i)(\underline{x}_n - \underline{\mu}_i)^T}_{\text{sample covariance matrix}}$

within-cluster scatter matrix

$$\underline{\underline{\Sigma}}_W = \sum_{i=1}^{C} \underline{\underline{\Sigma}}_i$$

(B) Between clusters:

total ~~clus~~ center: $\underline{\mu} = \frac{1}{N} \sum_{n=1}^{N} \underline{x}_n$

between-cluster scatter matrix $\underline{\underline{\Sigma}}_B = \sum_{i=1}^{C} N_i (\underline{\mu}_i - \underline{\mu})(\underline{\mu}_i - \underline{\mu})^T$

(C) For all samples:

total scatter matrix $\underline{\underline{\Sigma}}_T = \sum_{n=1}^{N} (\underline{x}_n - \underline{\mu})(\underline{x}_n - \underline{\mu})^T$

Properties

- $\underline{\underline{\Sigma}}_T$ fixed ; $\underline{\underline{\Sigma}}_W$, $\underline{\underline{\Sigma}}_B$ depend on S_i
- $\underline{\underline{\Sigma}}_T = \underline{\underline{\Sigma}}_W + \underline{\underline{\Sigma}}_B$

5-6

Goal of clustering

$$\min tr(\underline{\underline{\Sigma}}_W) \hat{=} \max tr(\underline{\underline{\Sigma}}_B)$$

trace : $tr(\underline{\underline{A}}) = \sum_i a_{ii}$

$tr(\underline{\underline{A}}\,\underline{\underline{B}}) = tr(\underline{\underline{B}}\,\underline{\underline{A}})$

Problem

$$\min_{S_1 \dots S_{C-1}} J(S_1 \dots S_{C-1}) = tr(\underline{\underline{\Sigma}}_W)$$

$$= \sum_{i=1}^{C} \sum_{\underline{x}_n \in S_i} \left\| \underline{x}_n - \frac{1}{|S_i|} \sum_{x_n \in S_i} \underline{x}_n \right\|^2$$

Discrete valued optimization problem

Q How many possibilities

5-7

5.1.2 K-means algorithm

(+) much faster

(−) don't find the best soln

Idea:

- If $\underline{\mu}_i$ are well known: classify all $\underline{x}_n$ to their nearest mean (4.2.1) $\Rightarrow$ cluster S_i

- If S_i known: recompute (improved) centers.

$$\underline{\mu}_i = \frac{1}{|S_i|} \sum_{\underline{x}_n \in S_i} \underline{x}_n$$

5-8

5-9

5-10

⋮

5-13

5.1.3 GMM

$\{\underline{x}(n)\}$, M = # modes $\hat{=}$ # clusters = C

$\downarrow$ EM

$\underline{\hat{\mu}}_m$, $\hat{\underline{\underline{C}}}_m$, $q_{mn} = p(\underline{x}(n)$ belongs to mode/cluster $m)$

$1 \leq m \leq M$

5.2 Clustering for an unknown number of clusters

5.1 : known number of clusters C,

k-means, GMM → sensitive to wrong C

In practice : C often unknown

⇒ need other clustering algorithms