Stability of Clustering

— Wednesday Afternoon Session —

Hans U. Simon

Email: hans.simon@rub.de

Homepage: http://www.ruhr-uni-bochum.de/lmi

1

Part I: An Informal Start

- Illustration of Clustering and Algorithms for Clustering
- The Intuitive Concept of Clustering Stability
- Known Analytic Results about Clustering Stability

Illustration of Clustering

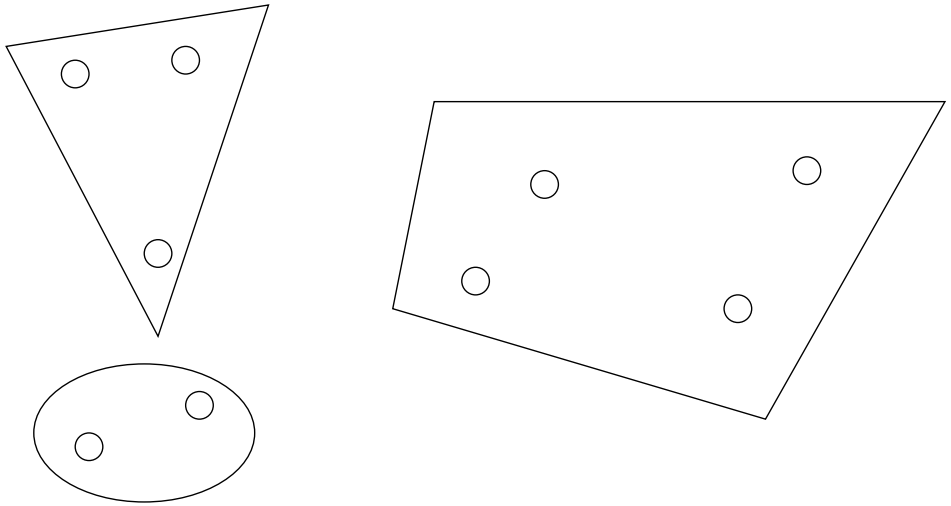

Figure 1: k-Clustering: a partition of data points into k classes (here, $k = 3$).

Illustration of k-Means Clustering

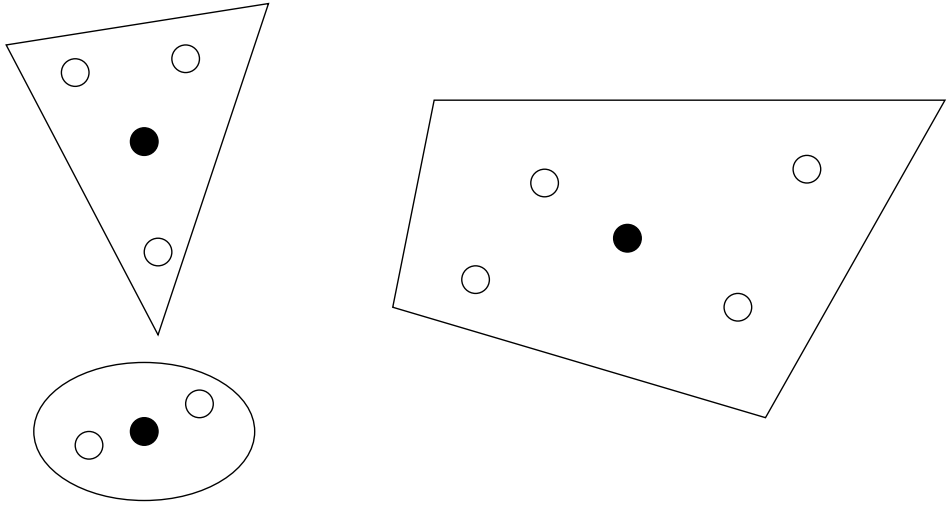

Figure 2: k-means aims at minimizing the average squared distance between a point and the corresponding center of gravity.

- NP-hard optimization problem in general
- efficient (EM-style) algorithms for the computation of a local optimum
- influence of large clusters potentially overemphasized
- "squared average distance" occasionally a bad choice as a "risk function"

A Nightmare Configuration for k-Means

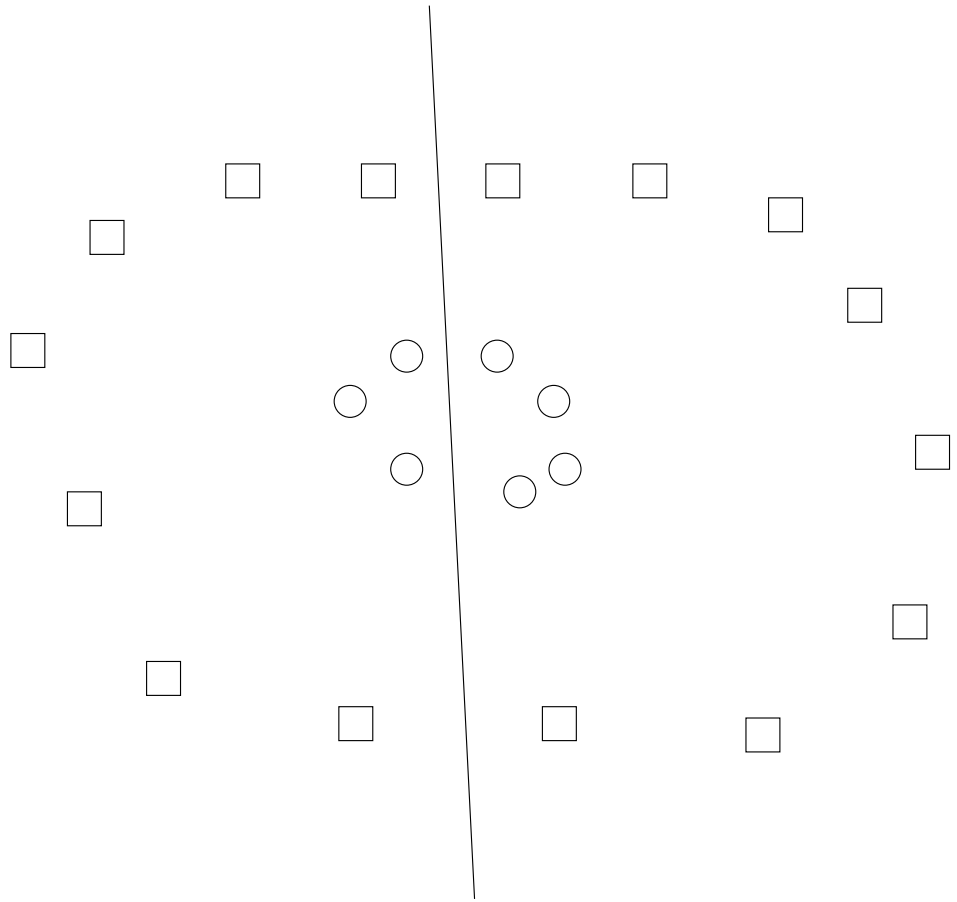

Figure 3: The 2-Clustering indicated by circles and squares looks pretty good (doesn't it?). k-Means would prefer the split indicated by the cut-line.

Some Alternatives to k-Means

k-Median: Choose k centers such as to minimize the average distance between a point and the center closest to it.

MST-Heuristic: Compute a minimum spanning tree and delete the $k-1$ edges with the largest weights.

Dissimilarity-based Heuristics: See second lecture on clustering stability.

The Intuitive Concept of Stability

A stable clustering algorithm should be robust against random fluctuations in the data !

As indicated in the following reasoning, choosing the "wrong" number of clusters should (hopefully) lead to instability.

Wrong Split: If k is a "good" number of clusters, and the algorithm produces $k+1$, then it has split at least one of the "true clusters". The wrong split is likely to be an over-sensitive reaction to "noise" in the data.

Wrong Merge: If k is a "good" number of clusters, and the algorithm produces $k-1$, then it has merged at least two of the "true clusters". Again this is likely to be an over-sensitive reaction to "noise" in the data.

Stability-based Clustering Decisions

Based on intuitive considerations (as above), stability is being widely used in practical applications as a heuristics for tuning parameters of clustering algorithms like

- the number of clusters
- or various stopping criteria.

Formally Unproven Intuitions May Go Wrong

The following example is taken from the COLT 2006 paper by Ben-David, von Luxburg, and Pál:

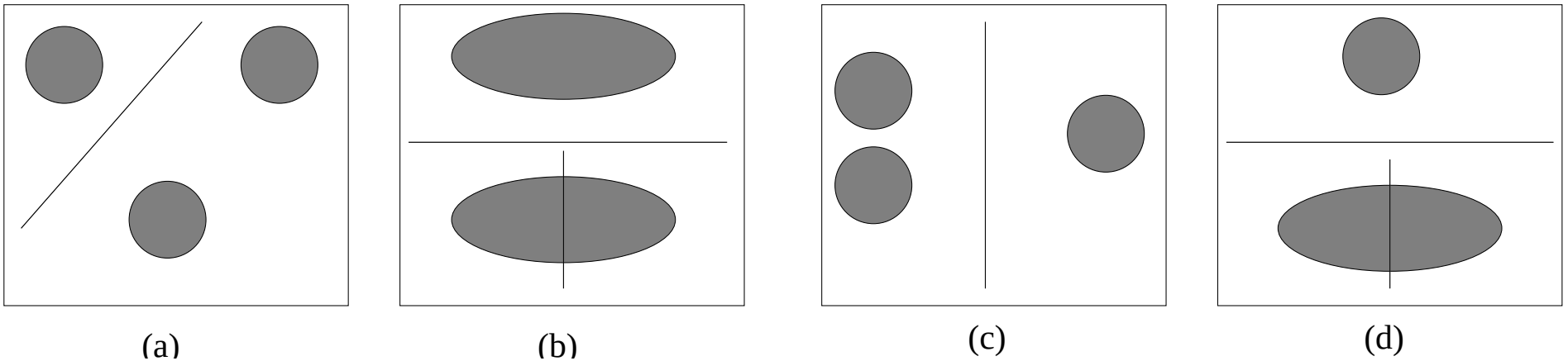

Figure 4: (a) wrong merge leading to instability (b) wrong split leading to instability (c) wrong merge leading to stability (d) wrong split leading to stability

Cartoons (a) and (b) support the intuition behind stability; cartoons (c) and (d) are in contradiction to it.

Some Pointers to the Literature

The following papers deal with clustering heuristics that follow the stability approach (without providing a rigorous analysis):

- Asa Ben-Hur, Andre Elisseeff, and Isabelle Guyon. A stability based method for discovering structure in clustered data. In *Pacific Symposium on Biocomputing 7*, pages 6–17, 2002.
- Sandrine Dudoit and Jane Fridlyand. A prediction-based resampling method for estimating the number of clusters in a data set. *Genome Biology*, 3(7):1–21, 2002.
- Tilman Lange, Mikio L. Braun, Volker Roth, and Joachim M. Buhmann. Stability-based model selection. In *Advances in Neural Information Processing Systems 15*, pages 617–624. MIT Press, 2003.
- Erel Levine and Eytan Domany. Resampling method for unsupervised estimation of cluster validity. *Neural Computation*, 13(11):2573–2593, 2001.

Papers Providing a More Formal Analysis

- Shai Ben-David, Ulrike von Luxburg, and Dávid Pál. A sober look at clustering stability. In *Proceedings of the 19th Annual Conference on Learning Theory*, pages 5–19, 2006
- Shai Ben-David, Dávid Pál, and H.U.S. Stability of k-means clustering. In *Proceedings of the 20th Annual Conference on Learning Theory*, pages 20–34, 2007.
- Alexander Rakhlin and Andrea Caponnetto. Stability of k-means clustering. In *Advances in Neural Information Processing Systems 19*. MIT Press, pages 1121–1128, 2007.

The last two papers (with the same title) refer to different notions of stability.

Known Analytic Results

The paper by Ben-David, von Luxburg, and Pál

- formalizes stability as a kind of robustness against independent resampling,
- shows that risk-minimizing algorithms are stable if the risk-minimizing clustering is unique,
- but the converse is proven only by making some additional symmetry assumptions.

The paper by Ben-David, Pál, and H.U.S. shows that k-means is stable if and only if the risk-minimizing clustering is unique.
The paper by Rakhlin and Caponnetto introduces another formal notion of stability (robustness against replacements of subsamples) and determines the degree of robustness of k-means w.r.t. this notion.

2

Part II: Towards a Theory of Clustering Stability

Data Space, Random Sample, Relative Frequencies

Data Space: $X = \{x_1, \ldots, x_n\}$

Probabilities: $\mu = (\mu_1, \ldots, \mu_n) \in \mathbb{R}^n_{++}$

Sample: $x_{j(1)}, \ldots, x_{j(m)} \in X$ (independently drawn according to μ)

Relative Frequencies: $W_i(x_{j(1)}, \ldots, x_{j(m)}) := \frac{1}{m}|l \in \{1, \ldots, m\} : j_l = i|$

Vector of Relative Frequencies: $\hat{W}_m = (W_1, \ldots, W_n)$

$\hat{W}_m$ can serve as a "guess" for the unknown vector μ of "true probabilities".

Admissible Clusterings and Risk Minimization

k-Clustering: a partition, $\mathcal{C}$, of X into k equivalence classes called "clusters"

Risk Function: $R_{\mathcal{C}}(w) \in \mathbb{R}_+$ (risk induced by $\mathcal{C}$ and "weight vector" w)

R-Minimizing Algorithm A: $(x_{j(1)}, \ldots, x_{j(m)}) \mapsto \hat{W}_m \mapsto \hat{\mathcal{C}}$,
where $\hat{\mathcal{C}}$ is a minimizer of $R_{\mathcal{C}}(w)$

Stability

Algorithm A is called stable if there is a clustering $\mathcal{C}_1$ such that

$$\lim_{m\to\infty} \Pr[A(\hat{W}_m) = \mathcal{C}_1] = 1 \ ,$$

i.e., A outputs $\mathcal{C}_1$ almost surely when m grows to infinity. Let $\mathcal{C}_1$ be the most likely output of A as m goes to infinity. Then

$$\text{instab}(A) := 1 - \lim_{m\to\infty} \Pr[A(\hat{W}_m) = \mathcal{C}_1]$$

can serve as a measure of instability (yielding zero iff A is stable).

Decision Function

- $\mathcal{C}_1$, the most likely output of A as m goes to infinity
- $\mathcal{C}_2$, the the second most likely output of A as m goes to infinity
- $f(w)$, the "decision function" given by

$$f(w) := R_{\mathcal{C}_2}(w) - R_{\mathcal{C}_1}(w) \ ,$$

  with ties broken in favor of $\mathcal{C}_1$ (for sake of clarity).

The following equivalence is obvious:

$$A \text{ is stable} \Leftrightarrow \lim_{m\to\infty} \Pr[f(\hat{W}_m) \geq 0] = 1$$

Assumptions

1. $f(w)$ is continuously differentiable at $w = \mu$ infinitely often.
2. There is an open ball around $w = \mu$ where the Taylor series

$$f(\mu + h) = \sum_{k \geq 0} T_k(h) \ , \tag{1}$$

$$T_k(h) = \frac{1}{k!} \sum_{1 \leq i_1, \ldots, i_k \leq n} \left. \frac{\partial^k f(w)}{\partial w_{i_1} \cdots \partial w_{i_k}} \right|_{\mu} h_{i_1} \cdots h_{i_k} \ , \tag{2}$$

$$= \sum_{\vartheta_1 + \cdots + \vartheta_n = k} \frac{1}{\vartheta_1! \cdots \vartheta_n!} \left. \frac{\partial^k f(w)}{\partial w_1^{\vartheta_1} \cdots \partial w_n^{\vartheta_n}} \right|_{\mu} h_1^{\vartheta_1} \cdots h_n^{\vartheta_n} \tag{3}$$

of f converges (where $\vartheta_1, \ldots, \vartheta_n \geq 0$ is always assumed implicitly).

Semidefiniteness versus Indefiniteness

$$\mathcal{U} := \{h \in \mathbb{R}^n : \ h_1 + \cdots + h_n = 0\} \ . \tag{4}$$

Let $k(\mathcal{U})$ be given by

$$k(\mathcal{U}) := \min\{k : \ T_k(h) \text{ does not vanish on } \mathcal{U}\} \ . \tag{5}$$

- T_k is called positive semidefinite on $\mathcal{U}$ if $T_k(h) \geq 0$ for every $h \in \mathcal{U}$.
- T_k is called negative semidefinite on $\mathcal{U}$ if $T_k(h) \leq 0$ for every $h \in \mathcal{U}$.
- T_k is called indefinite on $\mathcal{U}$ if it is neither positive nor negative semidefinite on $\mathcal{U}$.

T_k is indefinite on $\mathcal{U}$ (unless it vanishes on $\mathcal{U}$) for every odd k (because $T_k(-h) = -T_k(h)$ for every odd k).

Main Result 1

Theorem 2.1 *A is unstable if and only if $T_{k(\mathcal{U})}$ is indefinite on $\mathcal{U}$. Moreover, if $k(\mathcal{U})$ is odd, then*

$$\lim_{m\to\infty} \Pr[f(\hat{W}_m) > 0] = \lim_{m\to\infty} \Pr[f(\hat{W}_m) < 0] = \frac{1}{2}$$

which implies that $\mathrm{instab}(A) \geq 1/2$.

Sketch of Proof: a trivial case first

Exercise: Consider the case that $k(\mathcal{U}) = 0$. Argue that in this case $T_{k(\mathcal{U})}$ is positive definite (and thus not indefinite) and A is stable.

Since this is consistent with the theorem, we may now safely assume that $k(\mathcal{U}) \geq 1$.

Bringing the Central Limit Theorem into Play

- Consider the decomposition

$$\hat{W}_m = \mu + \hat{h}_m \ .$$

  Then, random vector $\hat{h}_m$ takes values in $\mathcal{U}$, is normally distributed with mean $\vec{0}$, and approaches $\vec{0}$ as m goes to infinity.

- The equidensity levels are concentric ellipsoids. Let E_0 be the unique such ellipsoid with a surface of volume 1.

- A random draw of $\hat{h}_m$ can be performed in two stages:
  - Pick a point h uniformly at random from the surface of E_0.
  - Pick a scaling factor λ_m at random and set $\hat{h}_m = \lambda_m h$.

- λ_m approaches zero as m goes to infinity.

First Non-vanishing Taylor-term Determines the Sign

Consider the decomposition

$$f(\mu + h) = T_{k(\mathcal{U})}(h) + \sum_{k > k(\mathcal{U})} T_k(h) \ .$$

If $f(\mu + h) \neq 0$, then

$$\mathrm{sign} f(\mu + \lambda \cdot h) = \mathrm{sign} T_{k(\mathcal{U})}(\lambda \cdot h)$$

for a sufficiently small $\lambda > 0$ (depending on h).

Indefiniteness = Instability

The following illustration is explained during the talk:

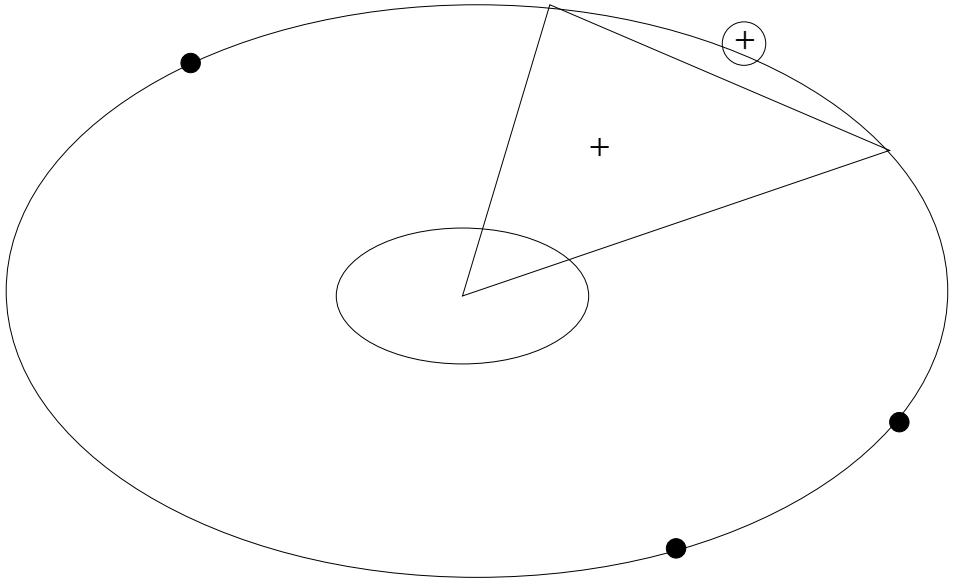

Figure 5: The outer ellipsoid is E_0. The black bullets mark areas on the surface of E_0 where $|T_{k(\mathcal{U})}|$ is smaller than a given threshold value. The sector marked "+" denotes an area where $T_{k(\mathcal{U})}$ is strictly positive. In the intersection of this sector and the inner ellipsoid, function f is strictly positive too.

Main Result 2

Recall that a function g defined on an open domain is called *homogeneous of degree* α if, for every λ that is sufficiently close to 1, the following holds:

$$g(\lambda x) = \lambda^{\alpha} g(x) \tag{6}$$

Theorem 2.2 *If the decision function $f(w)$ is homogeneous, then*

$$k(\mathcal{U}) = k(\mathbb{R}^n) \ .$$

Moreover, for $k := k(\mathcal{U}) = k(\mathbb{R}^n)$, T_k is positive semidefinite (or negative semidefinite, indefinite, respectively) on $\mathcal{U}$ if and only T_k is positive semidefinite (or negative semidefinite, indefinite, respectively) on $\mathbb{R}^n$.

Euler's Homogeneity Relation

The following result is a well-known fact:

Lemma 2.3 *For a continuously differentiable function $f : D \to \mathbb{R}$, $D \subseteq \mathbb{R}^n$, the following holds. $f(w)$ is homogeneous of degree α iff f satisfies the following condition (called "Euler's Homogeneity Relation") on D:*

$$\nabla f(w)^\top w = \alpha f(w) \tag{7}$$

Lemma 2.4 *Assume that f is continuously differentiable infinitely often on its domain D. Then the following holds. If $f(w)$ is homogeneous of degree α, then, for every $k \geq 0$ and every sequence $1 \leq i_1, \ldots, i_k \leq n$, function $(\nabla^k f(w))_{i_1,\ldots,i_k}$ is homogeneous of degree $\alpha - k$.*

Exercise: Prove Lemma 2.4 by means of Lemma 2.3 and by induction on k.

A Generalization of Main Result 2

Lemma 2.5 *Assume that $\mathcal{U}$ is a linear subspace of $\mathbb{R}^n$ of dimension $n-1$, $\mu \notin \mathcal{U}$, and f is homogeneous (say of degree α). Then, $k(\mathcal{U}) = k(\mathbb{R}^n)$. Moreover, for $k := k(\mathcal{U}) = k(\mathbb{R}^n)$, T_k is positive semidefinite (or negative semidefinite, indefinite, respectively) on $\mathcal{U}$ if and only T_k is positive semidefinite (or negative semidefinite, indefinite, respectively) on $\mathbb{R}^n$.*

Clearly, $k(\mathbb{R}^n) \leq k(\mathcal{U})$. As for the converse direction, it suffices to show that the following holds for every $l \geq 0$:

$$T_0, T_1, \ldots, T_k \text{ vanish on } \mathcal{U} \implies T_0, T_1, \ldots, T_k \text{ vanish on } \mathbb{R}^n. \qquad (8)$$

Exercise:
- For $k = 1$, (8) is fairly easy to show. Why ?
- Prove (8) for arbitrary $k \geq 0$ by induction.
- Argue that Lemma 2.5 applies to $\mathcal{U}$ and μ from Main Result 2.

3

Part III: Exercises

- two equivalent notions of stability
- uniqueness of the minimizing clustering as a sufficient condition for stability
- discussion of the case $k(\mathcal{U}) = 0$ in the proof of Main Result 1
- the covariance matrix for the random vector of relative frequencies
- the homogeneity of the partial derivatives of a homogeneous function
- the proof of (the generalized) Main Result 2

Exercise 1

The definition of stability based on independent resampling is as follows:

- Let $\hat{W}_m$ be the vector of relative frequencies derived from a first random sample of size m, and let $\hat{W}'_m$ be the vector derived from a second independent random sample of size m. Let $A(\hat{W}_m)$ and $A(\hat{W}'_m)$ denote the k-clusterings output by A when it is independently applied to the two random samples, respectively.
- The μ-Hamming-distance, d_μ, between two clusterings is defined as the probability (according to μ) that a random pair of data points falls into different clusters in one clustering and in the same cluster for the other-one.
- A is called stable if

$$\lim_{m\to\infty} \mathbb{E}[d_\mu(A(\hat{W}_m), A(\hat{W}'_m))] = 0 \ .$$

Argue that this definition of stability is equivalent to the definition given on page 16.

Exercise 2

Let R be a risk function and A an R-minimizing clustering algorithm. Argue that the following holds: if the clustering of minimum risk is unique, then A is stable.

Exercise 3

Solve the exercise on page 21 (trivial case within the proof of Main Result 1).

Exercise 4

- Let $\vec{e_i} \in \mathbb{R}^n$ denotes the vector with a 1 in component i and zeros in the remaining components, and let $\vec{e} \in \mathbb{R}^n$ denote the all-ones vector.
- Let $W \in \mathbb{R}^n$ denote the random vector that takes value $\vec{e_i}$ with probability μ_i.
- Note that the subspace $\mathcal{U} := \{h \in \mathbb{R}^n : \ h_1 + \cdots + h_n = 0\}$ from (4) is precisely the subspace spanned by the eigenvectors of C with strictly positive eigenvalues. Moreover, the random vector $\hat{W}_m$ of relative frequencies can be expressed as

$$\hat{W}_m = \frac{1}{m}(W_1 + \cdots + W_m) \ , \tag{9}$$

where $W_1, \ldots, W_m$ are i.i.d. with the same distribution as W.

Exercise 4 (continued)

Prove the following result:
$\mathbb{E}[W] = \mu$ and the covariance matrix C of W is given by

$$C[i,j] = \begin{cases} \mu_i(1-\mu_i) & \text{if } i = j \\ -\mu_i\mu_j & \text{if } i \neq j \end{cases} . \tag{10}$$

Moreover, the eigenvalues $\lambda_1 \geq \cdots \geq \lambda_N \geq 0$ of C are as follows:

1. $\lambda_N = 0$ (with eigenvector $(1,\ldots,1)^\top$).
2. Eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{N-1}$ are the extrema of the function

$$h(\lambda) = \prod_{i=1}^{N} \mu_i(\lambda - \mu_i)$$

so that

$$\mu_i \geq \lambda_i \geq \mu_{i+1}$$

for $i = 1,\ldots,N-1$.

Exercise 5

Solve the exercise on page 26 (homogeneity of partial derivatives of a homogeneous function).

Exercise 6

Solve the exercise on page 27 (proof of the lemma that generalizes Main Result 2).