Stability of Clustering

— Thursday Morning Session —

Hans U. Simon

Email: hans.simon@rub.de

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

Risk Functions

For sake of simplicity, we identify a k-clustering over a data space $X = \{x_1, \ldots, x_n\}$ as a partition of $[n] := \{1, \ldots, n\}$ into k classes (=clusters).

We consider risk functions $R_{\mathcal{C}}(w)$, where $\mathcal{C}$ is a k-clustering, with the following additive structure:

$$R_{\mathcal{C}}(w) = \sum_{C \in \mathcal{C}} R_C(w)$$

$R_C(w)$ is what cluster C contributes to the total risk.

Often $R_C(w)$ has the form

$$R_C(w) = S_C(w) \cdot \bar{R}_C(w) \text{ where } S_C(w) = \sum_{i \in C} w_i \ .$$

Here, $S_C(w)$ is the "total weight" of C and $\bar{R}_C(w)$ represents the weighted average risk within cluster C so that $R_{\mathcal{C}}(w)$ is the weighted sum of these average risk terms.

The k-Means Risk

$X \subseteq \mathbb{R}^d$, $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^d$, and

$$z_C(w) = \frac{1}{S_C(w)} \sum_{j \in C} w_j x_j$$

is the "center of gravity" within C:

$$\begin{aligned} R_C(w) &= \sum_{i \in C} w_i \cdot \|x_i - z_C(w)\|^2 \\ &= S_C(w) \cdot \underbrace{\sum_{i \in C} \frac{w_i}{S_C(w)} \|x_i - z_C(w)\|^2}_{=:\bar{R}_C(w)} \end{aligned}$$

Note that $\bar{R}_C(w)$ is the weighted-average squared distance between a point in C and the center $z_C(w)$ of C.

Exercise

Show the following:

- The centers of gravities are the optimal positions for centers, i.e., for fixed $\mathcal{C}$, any other choice of centers leads to a strictly increased risk.
- Optimal clusterings coincide with the partition induced by the Voronoi diagram for the center points.
- Optimal clusterings exclude "ties", that is, for for every point there is a unique closest center.

The first two observations give rise to an EM-style algorithm for the computation of a locally optimal clustering.

An Alternative Representation of the k-Means Risk

Exercise: Show that

$$\begin{aligned}\bar{R}_C(w) &:= \sum_{i\in C} \frac{w_i}{S_C(w)} \|x_i - z_C(w)\|^2 \\ &\stackrel{!}{=} \frac{1}{2} \cdot \sum_{i,j\in C} \frac{w_i}{S_C(w)} \frac{w_j}{S_C(w)} \|x_i - x_j\|^2\end{aligned}$$

so that

$$R_C(w) = S_C(w) \cdot \frac{1}{2} \cdot \sum_{i,j\in C} \frac{w_i}{S_C(w)} \frac{w_j}{S_C(w)} \|x_i - x_j\|^2$$

Some Risk Functions of a Similar Structure

$$R_C(w) = S_C(w) \cdot \frac{1}{2} \cdot \sum_{i,j \in C} \frac{w_i}{S_C(w)} \frac{w_j}{S_C(w)} \|x_i - x_j\|^2 \quad (1)$$

$$R_C(w) = \frac{1}{2} \cdot \sum_{i,j \in C} w_i w_j \|x_i - x_j\|^2 \quad (2)$$

$$R_C(w) = S_C(w) \cdot \frac{1}{2} \cdot \sum_{i,j \in C} \frac{w_i}{S_C(w)} \frac{w_j}{S_C(w)} \|x_i - x_j\| \quad (3)$$

$$R_C(w) = \frac{1}{2} \cdot \sum_{i,j \in C} w_i w_j \|x_i - x_j\| \quad (4)$$

Exercise: All risk functions shown above are homogeneous (of degree 1 or 2). Why might homogeneity be a desirable property for risk functions ?

Dissimilarity Matrix

A matrix $D \in \mathbb{R}^{n \times n}$ is called *dissimilarity matrix* if

- it is symmetric,
- has zeros on the main diagonal,
- and has strictly positive entries elsewhere.

Some remarks:

- $d_{i,j}$ represents the "dissimilarity" between data points x_i, x_j (often estimated by averaging over human judgments).
- D need not satisfy the triangle inequality and is therefore not necessarily a metric.
- Clustering algorithms will have access to data points only indirectly via D.

Dissimilarity-based Risk Functions

The *(first) risk function induced by* D is defined as

$$R_{\mathcal{C}}(w) = \sum_{C \in \mathcal{C}} S_C(w) \cdot \bar{R}_C(w)$$

where

$$\bar{R}_C(w) = \frac{1}{2} \cdot \sum_{i,j \in C} \frac{w_i}{S_C(w)} \frac{w_j}{S_C(w)} d_{i,j}$$

Note that $\bar{R}_C(w)$ represents the weighted-average dissimilarity within C.

Alternatively, we could consider the *second risk function induced by* D which is defined as follows:

$$R_{\mathcal{C}}(w) = \frac{1}{2} \cdot \sum_{C \in \mathcal{C}} \sum_{i,j \in C} w_i w_j d_{i,j}$$

All four risk functions from page 6 are special cases !

Reminder

Let $f = R_{\mathcal{C}_2} - R_{\mathcal{C}_1}$ be the decision function for the two distinguished clusterings $\mathcal{C}_1, \mathcal{C}_2$ (introduced in the Wednesday session). Let

$$\begin{aligned} f(\mu + h) &= \sum_{k \geq 0} T_k(h) \\ T_k(h) &= \frac{1}{k!} \sum_{1 \leq i_1, \ldots, i_k \leq n} \left. \frac{\partial^k f(w)}{\partial w_{i_1} \cdots \partial w_{i_k}} \right|_{\mu} h_{i_1} \cdots h_{i_k} \end{aligned}$$

be the Taylor-expansion around μ. Let $k(\mathbb{R}^n)$ be the smallest k such that $T_{k(\mathbb{R}^n)}(h)$ does not vanish. Then (provided that the risk function is homogeneous) the following holds for any risk-minimizing algorithm A:

- A is unstable if and only if $T_{k(\mathbb{R}^n)}$ is indefinite (on $\mathbb{R}^n$).
- Moreover, if $k(\mathbb{R}^n)$ is odd, then $\mathrm{instab}(A) \geq 1/2$.

Goal

The paper by Ben-David, Pál, and H.U.S. from COLT 2006 shows:

$$\begin{aligned} k\text{-means is stable} &\overset{!}{\Leftrightarrow} k(\mathbb{R}^n) = 0 \\ &\Leftrightarrow \text{the risk-minimizing clustering is unique} \end{aligned}$$

We conjecture that the analogous statement holds for any dissimilarity-based risk minimizing algorithm. We proceed in stages:

Stage 1: The conjecture is true and easy to verify for the *second* risk function induced by a dissimilarity matrix.

Stage 2: As for the *first* risk function, there is a "partial proof" that covers all but one weird case.

Stage 3: We show how one can cope with the weird case when we specialize our considerations to k-means clustering.

Stage 1: The Second Dissimilarity-based Risk Function

For $R_{\mathcal{C}}(w) = \sum_{C \in \mathcal{C}} R_C(w)$ and

$$R_{\mathcal{C}}(w) = \frac{1}{2} \cdot \sum_{C \in \mathcal{C}} \sum_{i,j \in C} w_i w_j d_{i,j} \ ,$$

the partial derivatives are easy to determine:
Let $D(\mathcal{C})$ be the matrix resulting from D by setting entry (i,j) to zero for all pairs whose components are not in the same cluster. Then

$$R_{\mathcal{C}}(w) = \frac{1}{2} w^\top D(\mathcal{C}) w \ , \ \nabla R_{\mathcal{C}}(w) = D(\mathcal{C}) w \ , \text{ and } \nabla^2 R_{\mathcal{C}}(w) = D(\mathcal{C})$$

Now consider the decision function $f = R_{\mathcal{C}_2} - R_{\mathcal{C}_1}$. It follows that:

- $\nabla^2 f(w)$ has zeros on the main diagonal.
- Since $\mathcal{C}_1 \neq \mathcal{C}_2$, $\nabla^2 f(w)$ has non-zero entries.
- $\nabla^2 f(w)$ is indefinite.

This completes stage 1 !

Exercise

Explain the reasoning on the preceding page in more detail.

Stage 2

Stage 2 is structured as follows:

- partial derivatives of the first dissimilarity-based risk function
- a central lemma:
  first non-vanishing term in the Taylor series for f is of order at most 3
- case-analysis excluding the weird case

Derivatives for the 1st Dissimilarity-based Risk Function

Using short-notation $(\textit{statement}) = 1$ if the statement is true and 0 otherwise, and

$$(\nabla^k R_{\mathcal{C}}(\mu))_{p_1,\ldots,p_k} := \left.\frac{\partial^k R_{\mathcal{C}}(w)}{\partial w_{p_1} \cdots \partial w_{p_k}}\right|_{\mu} , \tag{5}$$

we get:

$$(\nabla R_{\mathcal{C}}(w))_p = \sum_{j \in C_p} \frac{w_j}{S_{C_p}(w)} d_{p,j} - \frac{1}{2} \sum_{i,j \in C_p} \frac{w_i}{S_{C_p}(w)} \frac{w_j}{S_{C_p}(w)} d_{i,j} \tag{6}$$

$$(\nabla^2 R_{\mathcal{C}}(w))_{p,q} = \frac{(C_p = C_q)}{S_{C_p}(w)} \cdot (d_{p,q} - (\nabla R_{\mathcal{C}}(w))_p - (\nabla R_{\mathcal{C}}(w))_q) \tag{7}$$

From (7), we extrapolate that

$$(\nabla^2 R_{\mathcal{C}}(w))_{p,p} = \frac{-2}{S_{C_p}(w)} \cdot (\nabla R_{\mathcal{C}}(w))_p \tag{8}$$

Partial Derivatives of Higher Order

$$(\nabla^3 R_{\mathcal{C}}(w))_{p,q,r} = \frac{-(C_p = C_q = C_r)}{S_{C_p}(w)} \tag{9}$$

$$\Big((\nabla^2 R_{\mathcal{C}}(w))_{p,q} + (\nabla^2 R_{\mathcal{C}}(w))_{p,r} + (\nabla^2 R_{\mathcal{C}}(w))_{q,r}\Big) \tag{10}$$

More generally, for every $k \geq 3$, the following holds:

$$(\nabla^k R_{\mathcal{C}}(w))_{p_1,\ldots,p_k} = \frac{-(C_{p_1} = \cdots = C_{p_k})}{S_{C_{p_1}}(w)} \sum_{l=1}^{k} (\nabla^{k-1} R_{\mathcal{C}}(w))_{p_1,\ldots,p_{l-1},p_{l+1},\ldots,p_k} \tag{11}$$

Exercise

- Check the formulas for the partial derivatives.
- Try to give them a nice interpretation.
- Show the following: if R is the k-means risk function, then

$$\begin{aligned}(\nabla R_{\mathcal{C}}(w))_p &= \|x_p - z_C(w)\|^2 \\ (\nabla^2 R_{\mathcal{C}}(w))_{p,q} &= (C_p = C_q) \cdot \frac{-2\langle x_p - z_C(w), x_q - z_C(w)\rangle}{S_C(w)}\end{aligned}$$

A Central Lemma

We claim that

$$k_* := k(\mathbb{R}^n) \leq 3 \ .$$

The proof makes use of the following observations:

- Assume that $\nabla f(\mu)$ and $\nabla^2 f(\mu)$ vanish. It suffices to show that $\nabla^3 f(\mu)$ does not vanish.
- For any k-clustering $\mathcal{C}$:
  - $R_{\mathcal{C}}$ is not a linear function.
  - $\nabla^2 R_{\mathcal{C}}(\mu)$ does not vanish because, otherwise, all higher order terms would vanish as well (according to the recursion on page 15), which is impossible.
- We may assume wlog that $\mathcal{C}_1$ and $\mathcal{C}_2$ have no clusters in common. (Otherwise, remove common clusters and diminish k accordingly.)

Proof of the Central Lemma

- Pick $C \in \mathcal{C}_1$ and $p, q \in C$ such that $(\nabla^2 R_C(\mu))_{p,q} \neq 0$.
- Since $\nabla^2 f(\mu)$ vanishes, we conclude that there exists $C' \in \mathcal{C}_2$ such that $(\nabla^2 R_{C'}(\mu))_{p,q} = (\nabla^2 R_C(\mu))_{p,q} \neq 0$.
- Pick r from the symmetric difference of C and C'. For reasons of symmetry, we may assume that $r \in C \setminus C'$.

Proof of the Central Lemma (continued)

Clearly,

$$(\nabla^3 R_{C'}(\mu))_{p,q,r} = \nabla^2 R_{C'}(\mu))_{p,r} = \nabla^2 R_{C'}(\mu))_{q,r} = 0 \ .$$

On the other hand, we get

$$\begin{aligned}(\nabla^3 R_C(\mu))_{p,q,r} &= \frac{-1}{S_C(\mu)} \left((\nabla^2 R_C(\mu))_{p,q} + (\nabla^2 R_C(\mu))_{p,r} + (\nabla^2 R_C(\mu))_{q,r}\right) \\ &\stackrel{*}{=} \frac{-1}{S_C(\mu)} (\nabla^2 R_C(\mu))_{p,q} \neq 0 \ ,\end{aligned}$$

where the equation marked "$*$" follows because the Hessian of f vanishes at μ so that

$$(\nabla^2 R_C(\mu))_{p,r} = (\nabla^2 R_{C'}(\mu))_{p,r} = (\nabla^2 R_C(\mu))_{q,r} = (\nabla^2 R_{C'}(\mu))_{q,r} = 0 \ .$$

This settles the proof for the central lemma.

Analysis of "easy-to-handle" Cases

Let T_k be the term in the Taylor-expansion of $f = R_{\mathcal{C}_2} - R_{\mathcal{C}_1}$ around μ. Recall that $k_* = k(\mathbb{R}^n)$ denotes the order of the first term that does not vanish. The following is obvious from our general results about stability:

- If $k_* = 0$, then $\mathcal{C}_1$ is a unique minimizer of $R_{\mathcal{C}}(\mu)$ and any R-minimizing algorithm is stable.
- If $k_* \in \{1, 3\}$, then, for any R-minimizing algorithm A, instab$(A) \geq 1/2$ (implying that A is unstable).
- If $k_* = 2$ and $\nabla^2 f(\mu)$ has zeros on the main diagonal, then $\nabla^2 f(\mu)$ is indefinite (since there must be a non-zero entry outside the main diagonal). It follows that any R-minimizing algorithm is unstable.

Stage 2 is finished !!

Stage 3: The Remaining Case (focusing on k-means)

We are left only with the

Weird Case: $k_* = 2$ and $\nabla^2 f(\mu)$ has at least one non-zero entry on its main diagonal.

Claim: Assume that we are in the weird case. If R is the k-means risk function, then any R-minimizing algorithm is unstable.

Proof of the Claim

- Let $g := \nabla R_{\mathcal{C}_1} = \nabla R_{\mathcal{C}_2}$ denote the common gradient of $R_{\mathcal{C}_1}$ and $R_{\mathcal{C}_2}$.
- For any $1 \leq p \leq n$, let $C(p) \in \mathcal{C}_1$ and $C'(p) \in \mathcal{C}_2$ denote the respective cluster that contains p. Furthermore, let $W(p)$ and $W'(p)$ denote the total weight of $C(p)$ and $C'(p)$, respectively.
- With this notation:

$$(\nabla^2 f(\mu))_{p,p} = 2g_p(W(p) - W'(p))$$

It suffices to show that

$$(\exists 1 \leq p \leq n : (\nabla^2 f(\mu))_{p,p} > 0) \stackrel{!}{\Longrightarrow} (\exists 1 \leq q \leq n : (\nabla^2 f(\mu))_{q,q} < 0) \ .$$

(This would imply indefiniteness of $\nabla^2 f(\mu)$, which, in turn, implies instability.)

Exercise

Show the following:

- $(\exists 1 \leq p \leq n : (W(p) > W'(p)) \Rightarrow (\exists 1 \leq q \leq n : (W'(q) > W(q))$.
- $g_i = 0 \Rightarrow C(i) = C'(i) \Rightarrow W(i) = W'(i)$ so that $W'(q) > W(q) \Rightarrow g_q \neq 0$.

As for the second part, it is helpful to remember that the gradient gives us the squared distance between a point and the corresponding center of gravity, and that the clusters of optimal clusterings correspond to Voronoi-cells.

Proof of the Claim (continued)

Now we are ready to complete the analysis of the weird case:

$$
\begin{aligned}
(\exists 1 \leq p \leq n : (\nabla^2 f(\mu))_{p,p} > 0) \quad &\Longrightarrow \quad (\exists 1 \leq p \leq n : (W(p) > W'(p)) \\
&\Longrightarrow \quad (\exists 1 \leq q \leq n : (W'(q) > W(q)) \\
&\Longrightarrow \quad (\exists 1 \leq q \leq n : (W'(q) > W(q) \wedge g_q \neq 0) \\
&\Longrightarrow \quad (\exists 1 \leq q \leq n : (\nabla^2 f(\mu))_{q,q} < 0)
\end{aligned}
$$

Example

Consider the dissimilarity matrix

$$D = \begin{bmatrix} 0 & 1 & 1/4 & \infty & \infty & 1/4 \\ 1 & 0 & 1/4 & \infty & \infty & 1/4 \\ 1/4 & 1/4 & 0 & 1/4 & 1/4 & \infty \\ \infty & \infty & 1/4 & 0 & 1 & 1/4 \\ \infty & \infty & 1/4 & 1 & 0 & 1/4 \\ 1/4 & 1/4 & \infty & 1/4 & 1/4 & 0 \end{bmatrix} ,$$

where ∞ could be replaced by a sufficiently large value. Let μ represent the uniform distribution on $[6]$. It is easy to see that there are precisely two optimal 2-partitions,

$$\mathcal{C} = \{\{1,2,3\},\{4,5,6\}\} \text{ and } \mathcal{C}' = \{\{1,2,6\},\{4,5,3\}\} \ ,$$

each-one leading to risk 1/6.

Exercise

- Look at the example from the preceding page. Calculate the first order Taylor-terms of the decision function and show that the first non-vanishing term is order 3.

Final Remark

Exercise: Let R be the risk-function for k-means. Show that $\mathcal{C}_1 \neq \mathcal{C}_2$ implies that $\nabla^2 R_{\mathcal{C}_1}(\mu) \neq \nabla^2 R_{\mathcal{C}_2}(\mu)$. In other words, the clustering is uniquely determined by the Hesse matrix of its risk-function.

This shows $k(\mathbb{R}^n) \leq 2$ in the special case of k-means (in contrast to the example that we have seen above).