

Towards a Statistical Theory of Clustering

Ulrike von Luxburg, Shai Ben-David

Basic intuition in data-driven inference in science:

The more data we get the more accurate are the results we can derive from this data.

- ▶ Underlying assumption: data is generated by a random process
- ▶ In classification: we use generalization bounds
- ▶ In clustering: ???

Goal: raise basic questions, point out interesting problems, and discuss which techniques could (not) solve them.

Discuss difference between classification and clustering.

Separating two major questions

Question 1: How does a desirable clustering look like if we have the **complete knowledge** about our data generating process?

- ▶ conceptual question about the goal of clustering itself
- ▶ answer is a definition

Question 2: How can we approximate such an optimal clustering if we have **incomplete knowledge** or if we have **limited computational resources**?

- ▶ refers to a clustering *algorithm*
- ▶ answer is a statement with proof

Our basic setting

Given: X_1, \dots, X_n drawn iid from \mathbb{X} according to P ,
some extra knowledge (e.g. distances, “relevant structure”)

Goal: construct “best clustering” on (\mathbb{X}, P) from sample

To compute a distance between different clusterings:

Clusterings need to be defined on the **same space**.

- ▶ $C_1(\mathbb{X}_1), C_2(\mathbb{X}_2)$ clustering of subspaces $\mathbb{X}_1, \mathbb{X}_2 \subset \mathbb{X}$
- ▶ **Either extend** $C(\mathbb{X}_1), C(\mathbb{X}_2)$ to clusterings on \mathbb{X}
or restrict to clusterings on $\mathbb{X}_1 \cap \mathbb{X}_2$
- ▶ Then can define a distance $d(C_1, C_2)$ (e.g. by comparing for all pairs of points whether they are in the same group)

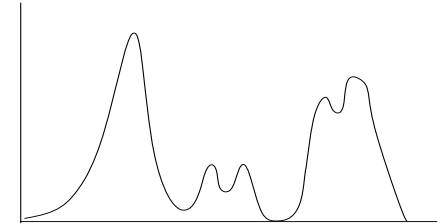
Question 1:

**Given a space \mathbb{X} with some probability distribution P ,
what is the best clustering on this space?**

Some definitions of “best” clustering

Best clustering is a mapping: $(\mathbb{X}, P) \mapsto C_*(\mathbb{X}, P)$

- ▶ Maximize a **quality criterion** q (e.g. k-means)
 - Is rather ad hoc, makes strong implicit assumptions
- ▶ identify **high density regions**
 - perform density estimation for clustering?
- ▶ **axiomatic** approaches
 - which choice of axioms?
- ▶ **Information theoretic** approaches
- ▶ ... many more ...



Which definition should we use?

- ▶ Different applications suggest different definitions
- ▶ None of them is clearly superior
- ▶ All of them have drawbacks

This question does not have a unique answer. Instead:

- ▶ What are our minimal requirements for such a definition from a statistical point of view?
- ▶ What can we prove if we don't have such a definition?

Continuity of “best” clustering

Best clustering is a mapping: $P \mapsto C_*(P)$

Would like to have **continuity of this mapping ...**

$$P_n \rightarrow P \quad \Rightarrow \quad C_*(P_n) \rightarrow C_*(P)$$

or

$$|P_1 - P_2| \leq \delta \Rightarrow d(C_*(P_1), C_*(P_2)) \leq \varepsilon$$

... at least for certain special cases:

P_n sequence of empirical distributions corresponding to P

Example: k-means criterion is continuous

$C_*(P)$ minimizes P-mean distance to cluster centers:

$$q(C) = \sum_i |x_i - \text{closest center}|^2$$

Pollard 1981: $(P_n)_n$ sequence of empirical distributions.
Then: optimal centers for $P_n \rightarrow$ optimal centers for P

Thus definition of “best clustering” is continuous.

For most clustering quality measures such an analysis has not been done yet!

Often we cannot even compute best clustering on (\mathbb{X}, P_n) :

- ▶ **Computational reasons**, e.g. k-means: computing optimal cluster centers can only be done in theory. In practice we can only approximate the global minimum.
- ▶ To **evaluate quality function**, might need to know **complete space** \mathbb{X} rather than points $\{X_1, \dots, X_n\}$, e.g. diameter based criterion. Here we need to estimate the quality based on $(\{X_1, \dots, X_n\}, P_n)$ instead of (\mathbb{X}, P_n) .

In both cases need to **estimate the best clustering on** (\mathbb{X}, P_n) .

Question 2:

How can we estimate or approximate the optimal clustering if we have **incomplete knowledge** or if we have **limited computational resources**?

Generalization bounds for clustering?

If we want to minimize a quality measure

Here a **standard generalization bound approach could work:**

- ▶ Compute an estimator $q_{\text{emp}}(f)$ of the quality function on (\mathbb{X}, P_n)
- ▶ Want to prove: $\min q_{\text{emp}}(f) \rightarrow \min q(f)$
- ▶ Need uniform convergence of $q_{\text{emp}}(f) \rightarrow q(f)$ over whole function class, for all probability measures P
- ▶ This can be done by standard techniques used in generalization bounds for classification

As far as I know: has not been done for most clustering algorithms!!!

If we don't have a quality measure

Definition of $C_*(P)$ cannot be expressed in terms of a quality function (Example: density based criterion)

On first glance: We don't know P , hence don't know $C_*(P)$.
Instead have P_n . Sounds similar to classification.

Overall goal: minimize $d(C(P_n), C_*(P))$

Problem: we cannot estimate it directly as we do not have any information on C_* ! **This is different from classification!**

But can we estimate it indirectly?

If P_n is close to P , then C should be close to C_* ...

Need additional assumptions on P

Estimating $d(C, C_*)$ indirectly:

- ▶ assume we know that $|P - P_n| < \delta$ with high probability
- ▶ assume that $C_*(P)$ is continuous with respect to P
- ▶ Then: $d(C_*(P_n), C_*(P)) < \varepsilon$ with high probability

To be able to prove that $|P_n - P| < \delta$, whp: **need to restrict the class of admissible probability distributions!!!**

Bounds will be bad as we do density estimation as intermediate step.

(Side question: is clustering easier than density estimation?)

Statement we would get

- ▶ given a function class F
- ▶ given a subset of “nice” probability measures P on \mathbb{X}
- ▶ if n is large enough, then with high probability:
the clustering computed from the finite sample will be close to the one computed from P

Techniques we would need to use:

- ▶ density estimation bounds (explain how to choose P)
- ▶ continuity of “best clustering”
- ▶ standard generalization bounds don’t work here

Question 3:

The most likely setting:

**We don't have a definition of “best clustering”,
we just want to use some given algorithm...**

Any theoretical guarantees?

Question 3: Turning the tables

Question 1: What is the best clustering?

Question 2: How can we approximate it on finite samples?

Now ask the other way round:

- ▶ Do the results of the algorithm converge for $n \rightarrow \infty$?
- ▶ If yes, is the limit clustering a useful clustering of the space $(\mathbb{X}, \mathcal{P})$?
- ▶ On a given sample of size n , how good is my result already?

- ▶ Convergence: Clusterings computed on n -sample converge for $n \rightarrow \infty$
 - results only known for very few algorithms (mixture models; not even k-means)
 - spectral clustering
- ▶ Stability analysis: Clusterings on independent n -samples should lead to similar results
 - used in practice, very few theoretical results
 - see talk of Shai

Note: convergence and stability are complementary aspects.

Spectral clustering: uses eigenvectors of graph Laplacians (\sim similarity matrix) to compute a clustering

Normalized spectral clustering (Luxburg, Bousquet, Belkin, COLT 04)

- ▶ always converges
- ▶ the limit clustering has a nice interpretation

Unnormalized spectral clustering (Luxburg, Bousquet, Belkin, NIPS 04)

- ▶ can fail to converge
- ▶ it can converge to trivial solutions
- ▶ we can construct basic examples where this happens
- ▶ the convergence conditions cannot be checked on the sample

- ▶ **Define** stability of algorithm A
- ▶ **Estimate** stability of different algorithms on the given sample
- ▶ Choose the most stable model
- ▶ **Statistical question:** decision about the best model is based on a test statistic computed on a random sample. How reliable is this?
- ▶ see Shai's talk 😊

For most clustering algorithms, most statistical questions are unsolved!

- ▶ *Clustering with known best clustering:*
 - defining “best clustering” requires continuity
 - consistency of estimators of best clustering
 - generalization bounds: only in case of quality fct.
 - estimating $d(C, C_*)$: only via density estimation
 - (Side question: is clustering easier than density est.?)
- ▶ *Analysis of clustering algorithms:*
 - convergence
 - stability

Pick your favorite algorithm and start 😊