

NIPS Workshop on Theoretical Foundations of Clustering

**Shai Ben-David, Ulrike von Luxburg, Naftali Tishby,
John Shawe-Taylor**

Whistler, 10 Dec 2005

Why do we need this workshop?

Clustering...

- ▶ is being applied in a variety of disciplines and contexts
- ▶ has been studied for over 50 years, in many communities
- ▶ there exist hundreds of algorithms
- ▶ intuitively, everybody knows what clustering is

So why do we need a workshop on theoretical foundations of clustering?

We cannot answer the simplest questions!

We still do not know how to ...

- ▶ compare clustering algorithms in a meaningful way
- ▶ measure the quality of a clustering algorithm (for a given task)
- ▶ measure the quality of a particular clustering (for a given data set)
- ▶ distinguish clustered from non-clustered data
- ▶ determine the number of clusters

▶ **New applications and types of data:**

- data is too complex (huge data sets, high dimensional data, streaming data)
- we can no longer check the results of a clustering algorithm by inspection or visualization

▶ **Clustering in a chain of tasks:**

- not only a stand alone tool for exploratory data analysis
- often used as an intermediate step in a data analysis chain (for example in semi-supervised learning). Here good performance is crucial.

Machine learning: reveal general rules by observing examples.

Observations/ examples can be:

- ▶ **individual objects**. Here we want to learn something about the structure of the space of those objects.
- ▶ **individual features** of some particular object. Here we want to learn more about this one object.
- ▶ **individual measurements** of differences between some objects. Here we want to learn something about the relations of the objects to each other, and do this by observing those relations.

- ▶ The more observations we have, the more accurate should the derived knowledge be (convergence)
- ▶ Often we assume that there is some “best solution” that we want to learn (consistency)
- ▶ The choice of the examples should not influence the result too much (stability).

↪ **Statistical properties of clustering (algorithms).**

Quality of a clustering on a particular data set:

- ▶ How can this be measured (if at all)?
- ▶ How can we compare clusterings on the same data set?
- ▶ Can *every* clustering task be expressed as an optimization of some explicit, readily computable, objective cost function?
- ▶ Can stability be considered a first principle for meaningful clustering?
- ▶ Can we distinguish clusterable from structureless data?

Quality of a clustering algorithm:

- ▶ Is there a principled way to measure this?
- ▶ Are there necessary/sufficient properties of “good algorithms”?
- ▶ What type of performance guarantees should we aim at?

More fundamental questions (with many answers)

What is the purpose of clustering?

Is the main purpose of clustering to discover new features in the data? Or to throw away unimportant information?

How can clustering be defined?

Is clustering just data compression? Is clustering just estimating modes of a density? Is clustering just (*fill in your favorite definition here*)?

What is a class / category?

Note that we do not need to answer this question in classification!

Goals of the workshop

- ▶ Collect different viewpoints
- ▶ What are the most important questions?
- ▶ Where should we start?