Stability of Clustering Methods

Sasha Rakhlin

Ph.D. candidate, MIT

A procedure is $stable$ if

$$\mathbb{P}\left(\left\|\text{solution} \;-\; \text{perturbed solution}\right\| > \varepsilon\right) \to 0$$

This talk:

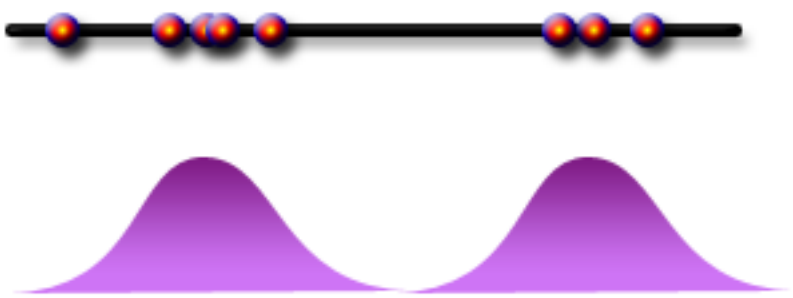
A tool for *theoretical* analysis of stability of clustering algorithms.

Idea: phrase *clustering* as *empirical risk minimization* and use stability of ERM.

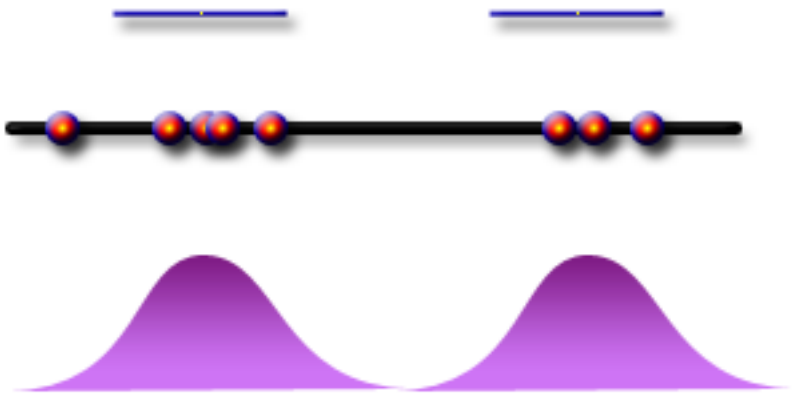
Based on work with A. Caponnetto: "Some properties of ERM over Donsker classes," submitted to JMLR.

Stability for model selection

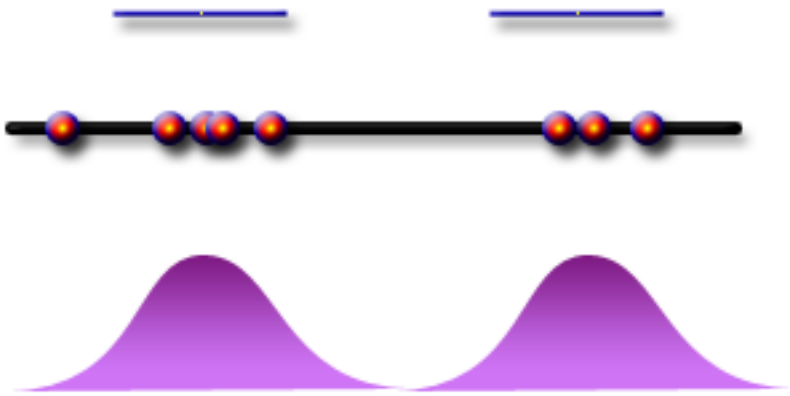
Stability for model selection



Stability for model selection



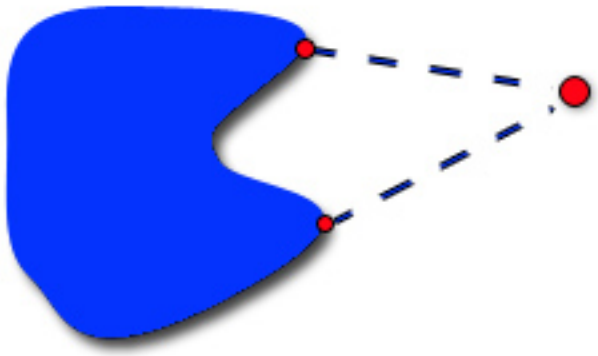
Stability for model selection



If the 2-cluster solution is in our hypothesis space ("realizable" case), we get stability with respect to perturbations of the whole dataset.

Stability for model selection

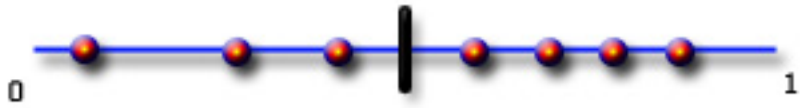
Instability (w.r.t. complete change of dataset) arises in the "non-realizable" case when there are two or more clusterings of similar "distance" to the underlying density.



What can we say about "non-realizable"? We will show that natural algorithms are stable w.r.t. change of $o(\sqrt{n})$ points.

Toy example

Choose, according to majority, either left or right half as the cluster.

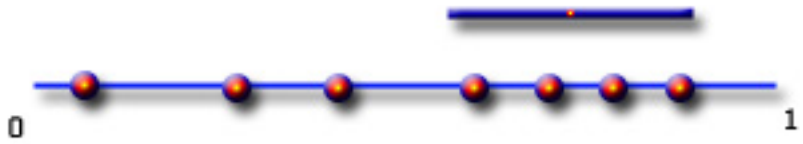


Probability that one point changes the cluster is $\Omega(n^{-1/2})$.

This procedure is stable with respect to changes of $o(\sqrt{n})$ points.

Much harder

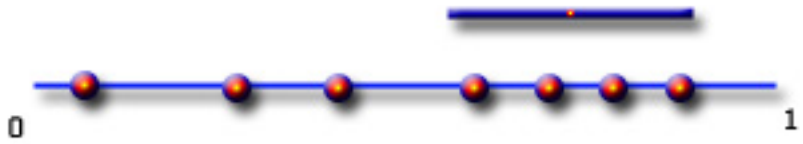
Choose, according to majority, a cluster of fixed size.



Does the probability of jumps by ε decrease as $n \to \infty$?

Much harder

Choose, according to majority, a cluster of fixed size.

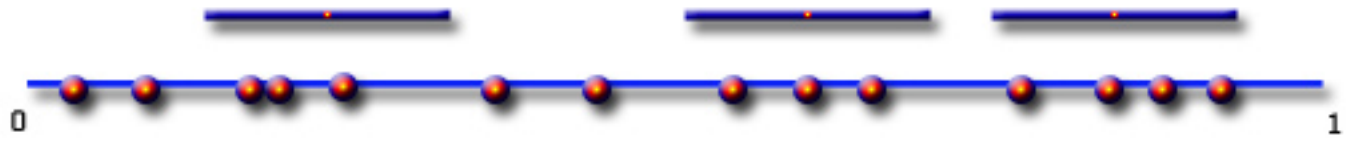


Does the probability of jumps by ε decrease as $n \to \infty$?

Yes, this procedure is stable w.r.t. changes of $o(\sqrt{n})$ points, no matter what P is.

Similar problem

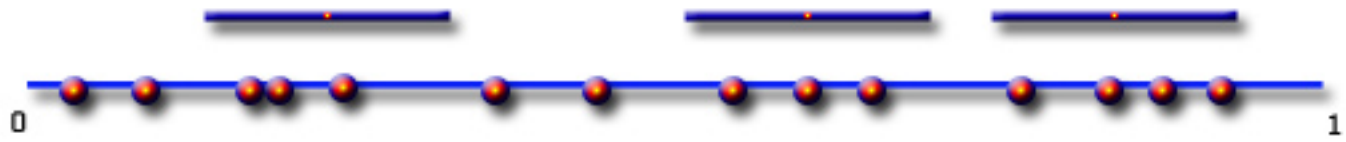
Choose, according to majority, k clusters of fixed size.



Does the probability of jumps (in L_1 distance) by ε decrease as $n \to \infty$?

Similar problem

Choose, according to majority, k clusters of fixed size.



Does the probability of jumps (in L_1 distance) by ε decrease as $n \to \infty$?

Yes, this procedure is stable w.r.t. changes of $o(\sqrt{n})$ points.

Empirical Risk Minimization

A. Caponnetto and A. Rakhlin "Some properties of ERM over Donsker classes," submitted to JMLR.

Empirical Risk Minimization

A. Caponnetto and A. Rakhlin "Some properties of ERM over Donsker classes," submitted to JMLR.

These examples are instances of empirical risk minimization. The following general result holds for any fixed distribution P:

$$\forall \varepsilon > 0, \quad \mathbb{P}\left(\|f_S - f_T\|_{L1} \geq \varepsilon\right) \to 0$$

where S and T differ on $o(\sqrt{n})$ points, and f_S, f_T are respective almost-minimizers over a P-Donsker class.

Empirical Risk Minimization

A. Caponnetto and A. Rakhlin "Some properties of ERM over Donsker classes," submitted to JMLR.

These examples are instances of empirical risk minimization. The following general result holds for any fixed distribution P:

$$\forall \varepsilon > 0, \quad \mathbb{P}\left(\|f_S - f_T\|_{L1} \geq \varepsilon\right) \to 0$$

where S and T differ on $o(\sqrt{n})$ points, and f_S, f_T are respective almost-minimizers over a P-Donsker class.

For binary functions, Donsker = VC.

k-means clustering

We can now study stability of other clustering procedures which optimize an objective function.

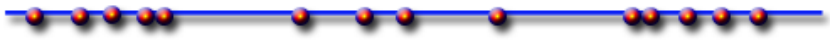
k-means clustering is

$$\min_C \sum_{i=1}^{n} \|x_i - m_{C(x_i)}\|^2$$

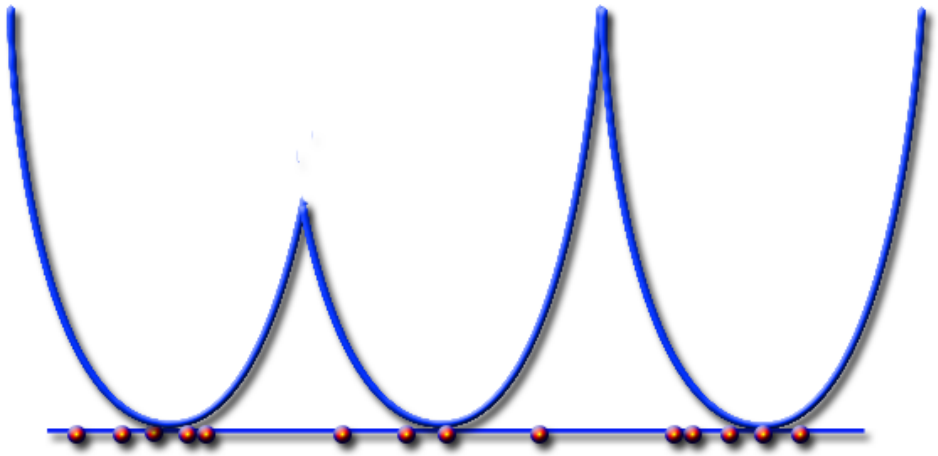
which is empirical risk minimization over the class

$$\mathcal{F} = \{\|x - m_{C(X)}\|^2 : C \text{ is a } k\text{-partition and } m_{C(x)} \text{ are centers}\}$$

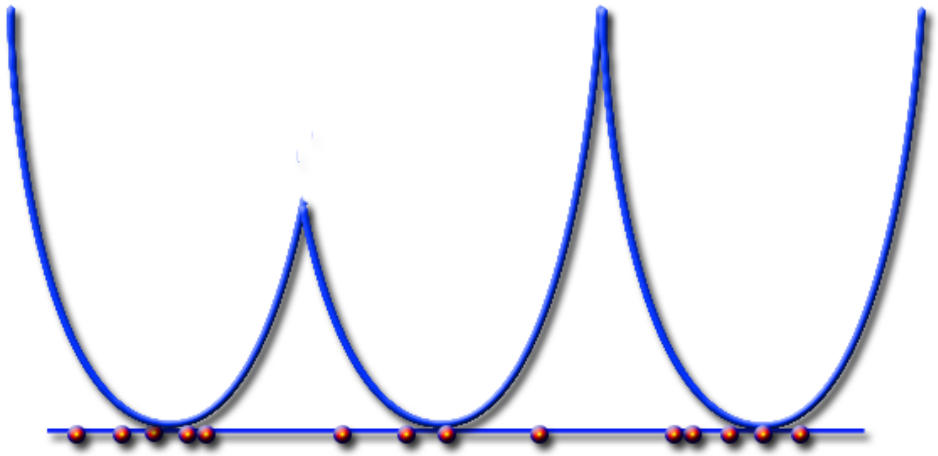
k-means clustering



k-means clustering



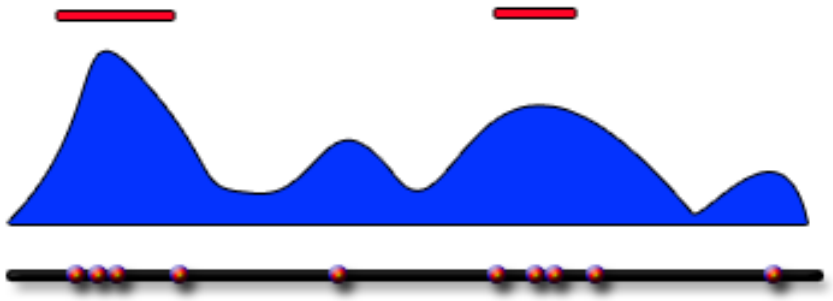
k-means clustering



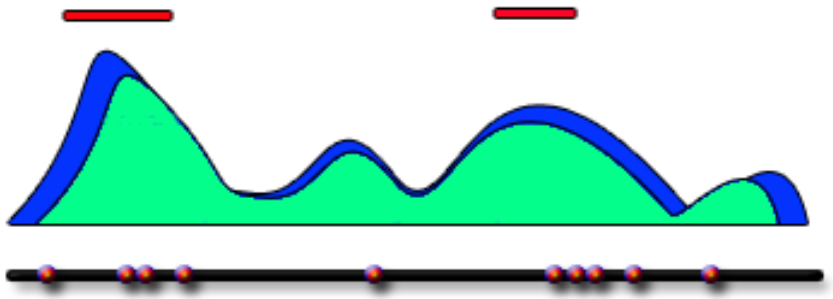
$$\mathcal{F} = \{\|x - m_{C(X)}\|^2 : C \text{ is a } k\text{-partition and } m_{C(x)} \text{ are centers}\}$$

If $\mathcal{F}$ is Donsker (e.g. domain is compact), then L_1 stability implies stability of centers $m_{C(x)}$.

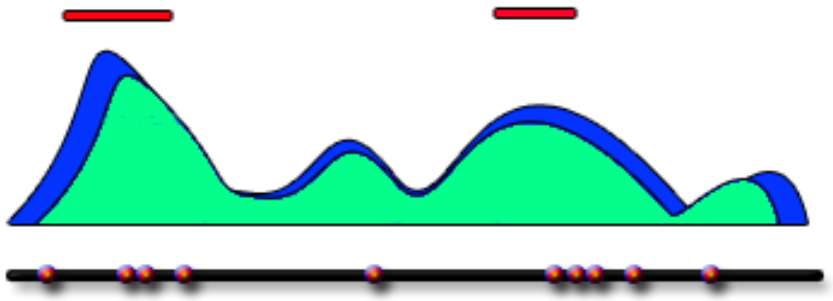
MLE density estimation



MLE density estimation



MLE density estimation



$$\max_{f \in \mathcal{F}} \sum_{i=1}^{n} \log f(x_i)$$

Under some assumptions on the class $\mathcal{F}$ of densities, this should imply stability of modes/clusters.

That's all