Lecture # 9: 10-2-18

- So far: clustering with respect to shape (K-means) and distance between clusters at different scales (hierarchical methods)
- A different approach is to consider a notion of "density" as correlating with clusterability.

ex:

- The two circles are sample roughly uniformly.
- The density on each is roughly the same, and there is no easy way to step between them.
- We want a method that can identify these clusters.

"SIGKDD test of time" award

↓

- DBSCAN: "Density-based spatial clustering of applications with noise"
- Let $X \in \mathbb{R}^d$, and let $B_r(x) = \{y \mid \|x-y\|_2 \leq r\}$.
- Let $\{X_1, \dots, X_n\} \subset \mathbb{R}^d$ be the data to be clustered. We say X_i is a core point if $|B_\varepsilon(x)| \geq \text{MinPts}$, where ε, MinPts are parameters.
- ε may be thought of as local neighborhood size: $\varepsilon \to 0 \Rightarrow |B_\varepsilon(x)| = 0, \forall x$
  $\varepsilon \to \infty \Rightarrow |B_\varepsilon(x)| = n, \forall x$.
- MinPts is a threshold for significance: MinPts small suggests every neighborhood is meaningful.

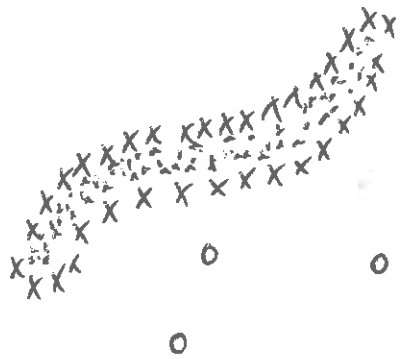
- MinPts large suggests no neighborhood is meaningful.

- A ~~core~~ border point is x s.t. 1) $|B_\varepsilon(x)| < \text{MinPts}$

2.) $b_\varepsilon(z) \ni x$, for some core point z.

- A noise point is everyone else.

ex:



$\circ$ = core point
x = border point
$\circ$ = noise point

} with respect to some ε, MinPts.

- DBSCAN works by linking contiguous regions of "core points" together.

- More precisely, we say:

1.) X is directly density reachable from y if a.) y is a core point
b.) $x \in B_\varepsilon(y)$

2.) X is density reachable from y if there exists a path

$X_0, X_1, \ldots, X_L$ such that: a.) $X_0 = X$, $X_L = y$

b.) X_i is directly density reachable from X_{i-1}, $i = 2, \ldots, L$.

3.) X and y are density connected if there exist a core point z such that both x and y are density reachable from z.

- The clusters learned by DBSCAN are the maximal sets of density connected points.

- This method has the benefit of being shape-agnostic, meaning it doesn't favor spherical clusters the way K-means does, for example.
- On the other hand, it depends heavily on ε, MinPts, and has large complexity in high dimensions.
- Indeed, computing $B_\varepsilon(x)$ requires a nearest neighbors search. A naive implementation is $O(n)$ for each point, so overall $O(n^2)$.

Remark: Indexing structures like k-d trees (~1975) and covertrees (~2005) speed up these searches to $O(n \log(n))$ with a big constant, if certain mathematical assumptions on the data hold.

- Note that "density" was implicit here, in the notion of core point, border point, and noise point.

- We can also explicitly estimate the density of the points, using a kernel-density estimate.

- Let $K: \mathbb{R}^d \to \mathbb{R}$ be a kernel, i.e. 1.) $K(x) \geq 0 \quad \forall x$

2.) $\int_{\mathbb{R}^d} K(x)dx = 1$

3.) $K(-x) = K(x)$.

- Let $\rho(x) = \frac{1}{nh^k} \sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)$ for some scalar h, which can be understood as the scale of the density estimator. As $n \to \infty$, $h \to 0$, $\frac{\rho(x)}{\int_{\mathbb{R}^d} \rho(x)\,dx}$ estimates the underlying probability distribution of the data (under certain mathematical assumptions we do not discuss).

- Choosing $K: \mathbb{R}^d \to \mathbb{R}$ the kernel is important.

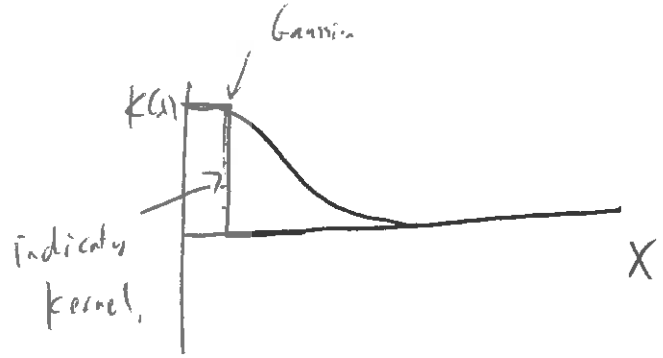
ex. Fix $T > 0$. Let $k_T(x) = \begin{cases} 1, & \|x\|_2 \leq T \\ 0, & \text{else} \end{cases}$

↑ Indicator kernel

- This gives max wherever $\|x - x_i\| \leq h_T$

- To soften the transition, a smooth kernel could be used

ex: $K_\sigma(x) = e^{-\|x\|_2^2 / 2\sigma^2}$ for some $\sigma > 0$. This is the Gaussian kernel.



- Supposing the density is well estimated by $\hat{p}(x)$ (happens if n is large, say $n \gg 2^D$ and h small)

this can be used to cluster: associated points to nearby regions of high density.

- Several variations on the theme:
  - DENCLUE
  - Mean-Shift Clustering
  - Fast search and finding density peaks clustering (FSFDPC)

- Generally good, but may need a lot of points, or computational resources