Графы ближайших соседей и алгоритмы

Вотинов Антон
Научный руководитель - Панов Максим

09.12.2015

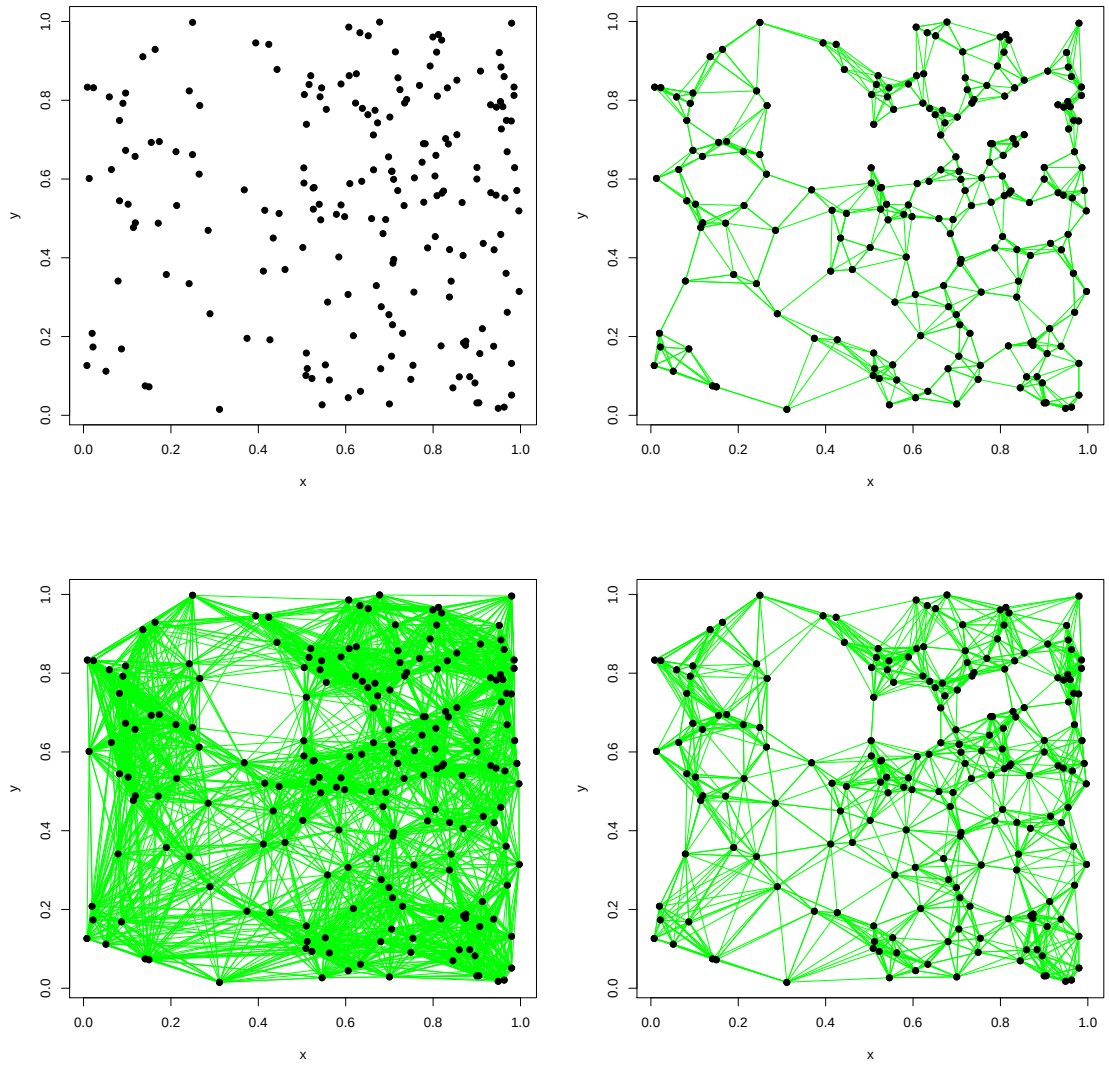
ВЫСШАЯ ШКОЛА ЭКОНОМИКИ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

kNN граф

- Пусть имеем исходное пространство $\chi \subseteq R^d$ и выборка X, семплированная какой-то функцией плотностью.
- На этом пространстве задана функция похожести $S(v_i, v_j)$ (similarity function), с помощью которой мы можем найти ближайших соседей для каждой точки.
- Ребро (v_i, v_j) существует, только если v_j является k-ближайшим соседом вершины v_i.
- Вес w_{ij} равен единице или определяется функцией (similarity function).
- Ориентированные/неориентированные графы.
- Графы взаимных друзей: (v_i, v_j) существует, если v_i - k-nn для v_j, и наоборот.

kNN граф

Примеры



Оценка размерности

Постановка задачи

- Низкоразмерное множество $\chi \subseteq R^d$.
- Функция $\phi : \chi \to M \subseteq R^D$.
- Функция плотности f_χ, заданная на χ.
- Наблюдения $x = (x_1, x_2, ..., x_n)$, семплированные функцией плотности f_χ, переводятся фунцией ϕ в пространство большей размерности с наблюдениями $\tilde{x} = (\tilde{x_1}, \tilde{x_2}, ..., \tilde{x_n})$:$\tilde{x_i} = \phi(x_i) + \nu_i$.
- Мы наблюдаем kNN граф G_k, построенный по наблюдениям $\tilde{x}$.

Задача: по kNN графу G_k оценить размерность d исходного пространства χ.
Мотивация: борьба с большой размерностью.

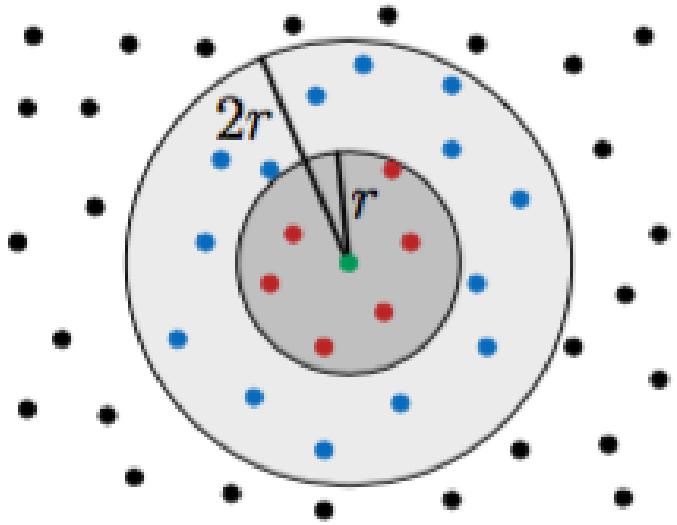
Оценка размерности

Алгоритм 1. Раздуваем шарик

Пусть λ_d - мера Лебега. Пусть $B(x, r)$ - количество точек, находящихся внутри d-мерной сферы с центром в точке x.

Doubling property:

$$\lambda_d(B(x, 2r)) = 2^d \lambda_d(B(x, r)).$$



Оценка размерности

Алгоритм 1. Раздуваем шарик

Определим $B_{SP}(i, r) = \{j \in V : d_{SP}(i, j) \leq r\}$, SP = Shortest Path. Тогда оценка размерности ищется из:

$$L_{DP}(i) = \frac{|B_{SP}(i, 1)|}{|B_{SP}(i, 2)|} \approx \frac{1}{2^d}.$$

Чтобы получить робастную статистику, усредним $L_{DP}(i)$ по всем вершинам

$$L_{DP}(A) = \frac{1}{|A|} \sum_{i \in V} L_{DP}(i)$$

$$E_{DP}(A) = -log_2 L_{DP}(A).$$

Оценка размерности

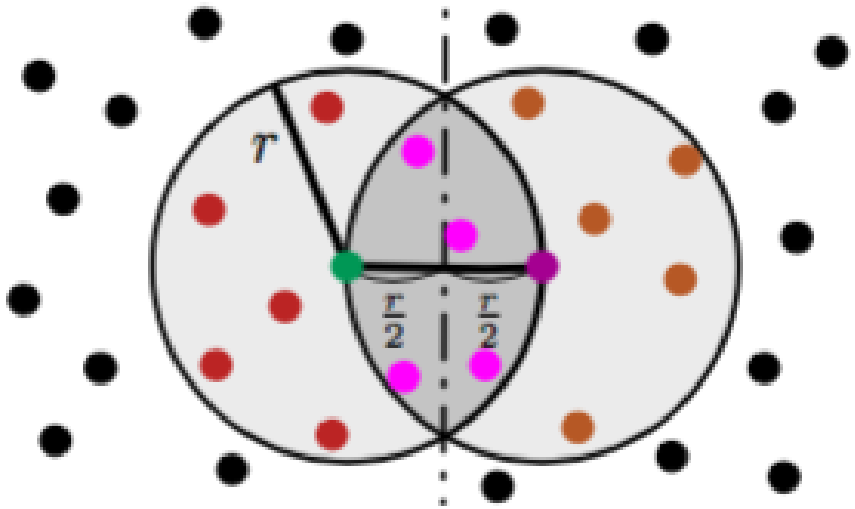
Алгоритм 2. Пересекаем шарики

Пусть $\eta_d = \lambda_d(B(0,1))$. Зафиксируем две точки x и y такие, что $d(x,y) = r$, тогда объём пересечения двух сфер $B(x,r) \cap B(y,r)$ определяется так:

$$\frac{1}{2}\eta_d r^d I_{\frac{3}{4}}(\frac{d+1}{1}, \frac{1}{2}).$$

Из равенства получаем:

$$\frac{\lambda_d(B(x,r) \cap B(y,r))}{\lambda_d(B(x,r))} = I_{\frac{3}{4}}(\frac{d+1}{1}, \frac{1}{2}) = S(d).$$



Оценка размерности

Алгоритм 2. Пересекаем шарики

Для i-й вершины получаем оценку:

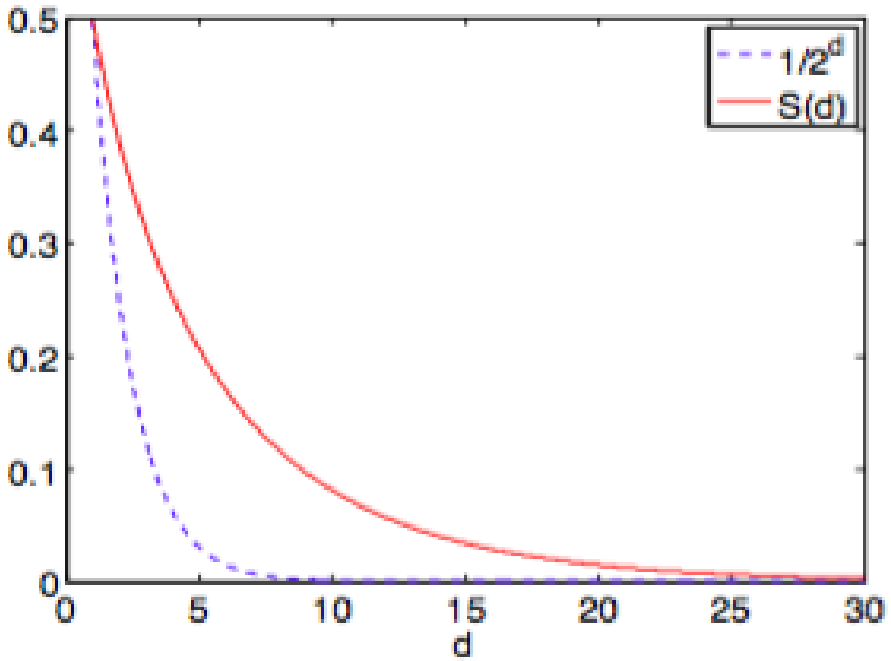
$$L_{CAP}(i) = \frac{min_{j \in V: i \to j} |B_{SP}(i,1) \cap B_{SP}(j,1)|}{k+1} \approx S(d).$$

Получаем следующую оценку размерности (как и выше, $L_{CAP}(A)$ - робастная оценка):

$$E_{CAP}(A) = S^{-1}(L_{CAP}(A)).$$

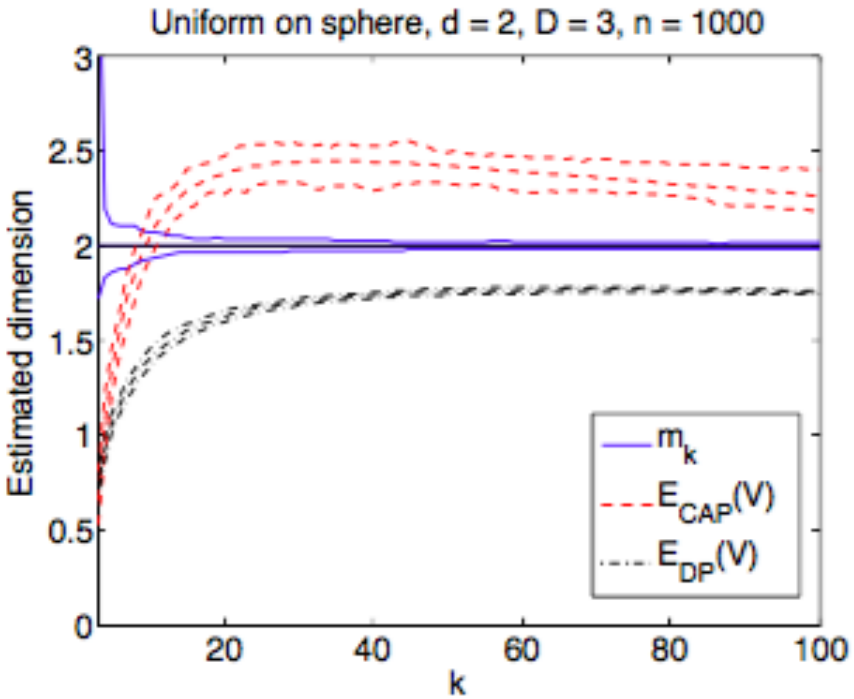
Оценка размерности

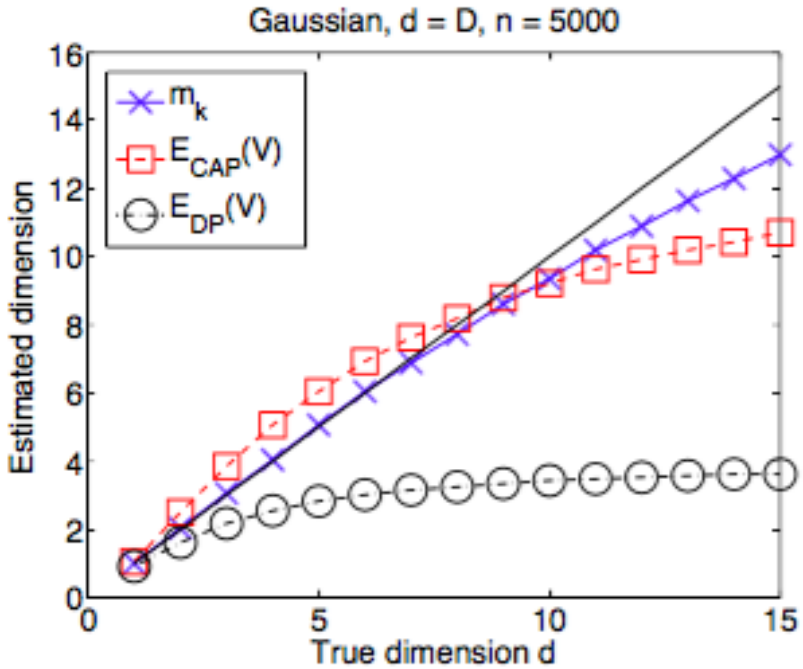
Сравнение оценок



Оценка размерности

Сравнение оценок





Оценка плотности

Постановка задачи. Определения

- Пусть $p(x)$ - функция плотности, определённая на $\chi \subseteq R^d$.
- По выборке $\chi_n = \{X_1, X_2, ..., X_n\} \subseteq R^d$ строится kNN граф $G_n =< V_n, E_n > (V_n = \chi_n)$.

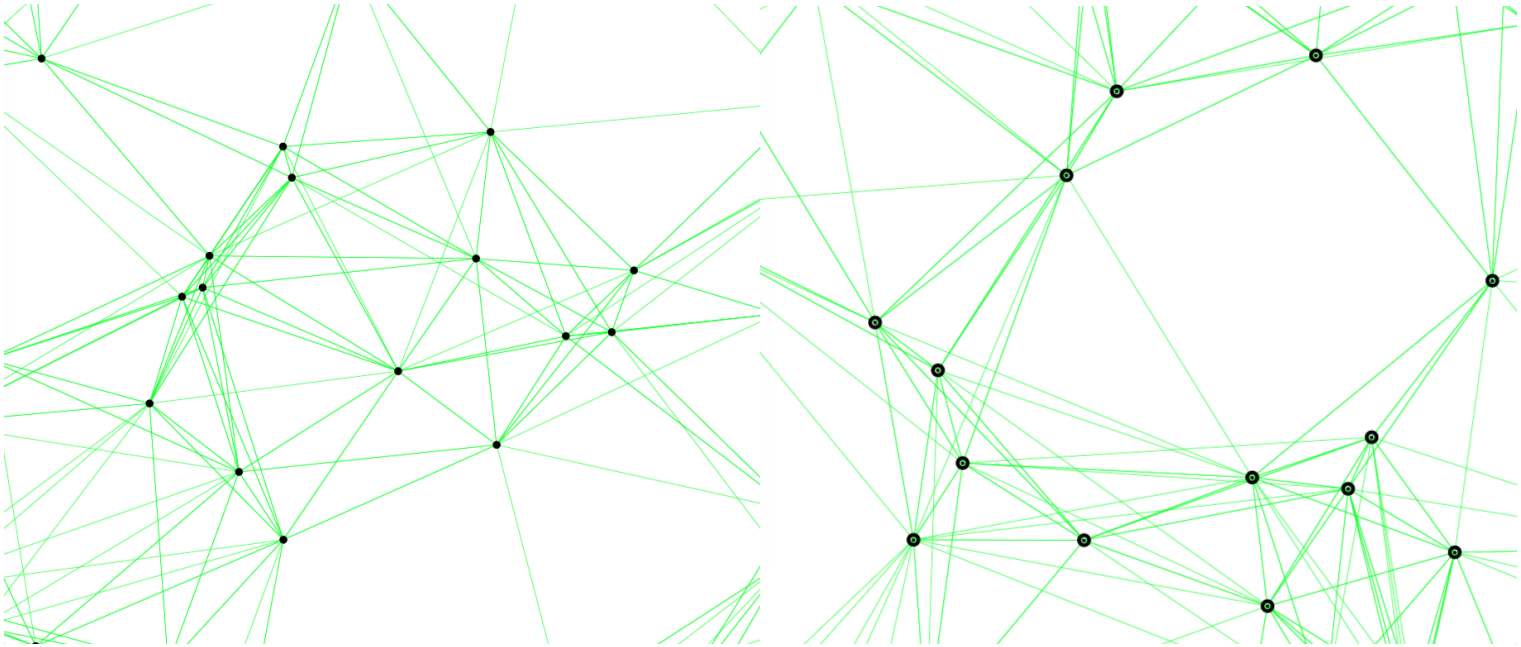
Задача: по kNN графу восстановить исходную функцию плотности $p(x)$.
Нужные определения:

- η_d - объём единичного d-мерного шарика.
- v_d - объём пересечения двух d-мерных шариков, расстояние между центрами у которых равно единице.
- $In(x) = \{y \in V_n : (y, x) \in E_n\}$.
- $Out(x) = \{y \in V_n : (x, y) \in E_n\}$.

Оценка плотности

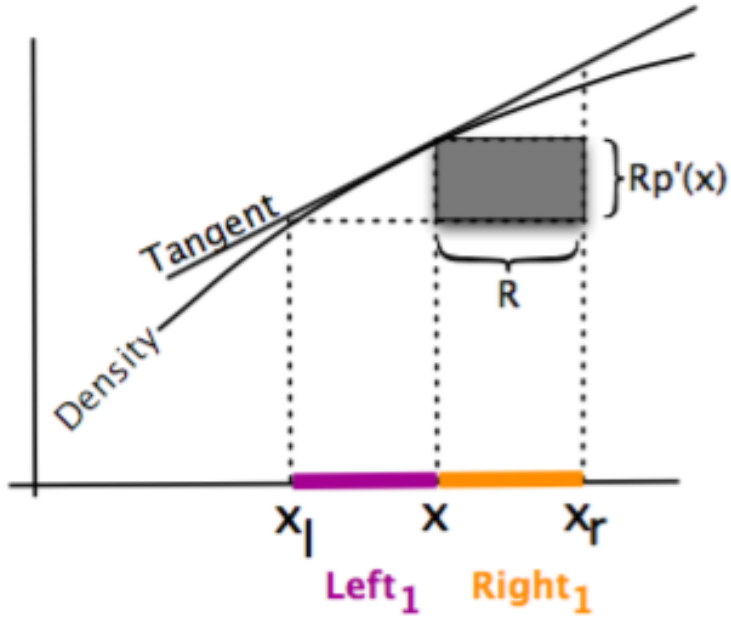
Проблема

Основная проблема связана с тем, что вес любого ребра равен единице: для взвешенного графа плотность легко находится. Более того, нельзя оценивать плотность, используя только локальные свойства.



Оценка плотности

Алгоритм. Идея



Рассмотрим одномерный случай. Знаем расстояние между точками.
Пусть $Left_1(x) = |\{y \in Out(x) : y < x\}|$ и $Right_1(x) = |\{y \in Out(x) : y > x\}|$.
Тогда:

$$Right_1(x) - Left_1(x) \approx n[P([x; x + R]) - P(x; x - R)].$$

$$n[P([x; x + R]) - P(x; x - R)] \approx R^2 p'(x).$$

$$R \approx \frac{k}{2np(x)}.$$

Оценка плотности

Алгоритм. Идея

Из полученных ранее соотношений получаем:

$$Right_1(x) - Left_1(x) \approx \frac{k^2}{4n^2}\frac{p'}{p^2}$$

, откуда получаем точечную оценку для $\frac{p'(x)}{p^2(x)}$.

От локального к глобальному:

1. Фиксируем некоторую точку X_0 (anchor point).
2. Для каждой точки $x \in [X_0, X_s]$ суммируем оценки $\frac{p'(x)}{p^2(x)}$ относительно, что равносильно интегрированию $\frac{p'(x)}{p(x)}$ на промежутке $x \in [X_0, X_s]$.
3. Полученная сумма - оценка $log(p(X_s)) - log(p(X_0))$.
4. Потенциируем и получаем оценку $p(x)$ в точке X_s.

Оценка плотности

Алгоритм. И вот Он!

Зафиксируем некоторую вершину X_0. Пусть $\gamma(X_s)$ - кратчайший путь от вершины X_0 до вершины X_s. Тогда $log(p(X_s)) - log(p(X_0))$ оценивается следующим образом:

$$\frac{\eta_d}{kv_d} \sum_{x \in \gamma} Right_\gamma(x) - Left_\gamma(x).$$

Где, для x_l - предшественник, x_r - последователь в пути γ:

$$Right_\gamma(x) = |Out(x) \cap In(x_r)|$$

$$Left_\gamma(x) = |Out(x) \cap In(x_l)|.$$

Оценка плотности

Результат