Out-of-Sample extensions for LLE, Isomap, MDS, Eigenmaps, and Spectral Clustering

Расширения алгоритмов LLE, Isomap, MDS, Eigenmaps, и Spectral Clustering для точек вне обучающей выборки

Постановка задачи снижения размерности

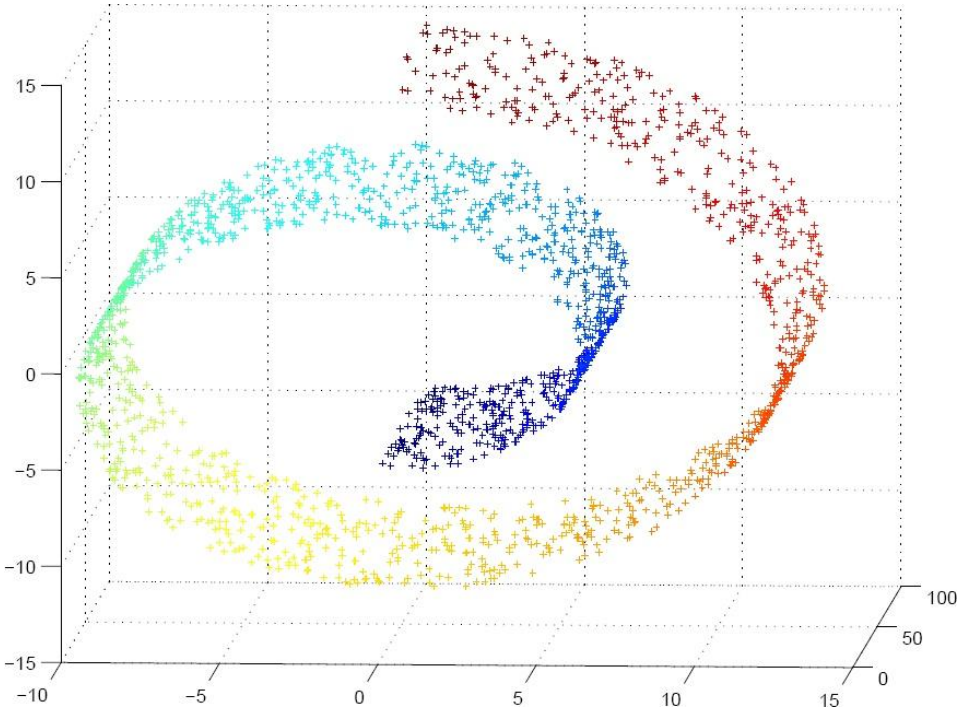
Пусть дано множество точек $D = \{x_1, ..., x_n\}$ в R^d.
Необходимо описать D меньшим количеством координат:

$$Y = \{y_1, ..., y_n\} \text{ в } R^m,\ m < d.$$

Зачем нужно снижение размерности?

- Сокращение вычислительных затрат при обработке данных
- Сжатие данных для более эффективного хранения данных
- Визуализация данных
- и др.

Пример



Входные и выходные данные

Вход:

$D = \{x_1, ...x_n\} \in R^d$ - выборка

$K(\cdot,\cdot): \quad R^d \times R^d \to [0, \inf)$ - функция близости

Выход:

$Y = \{y_1, ..., y_n\} \in R^m (m < d)$ - низкоразмерное описание D

Общий алгоритм

1. Строим матрицу близости M размера $n \times n$: $M_{ij} = K_D(x_i, x_j)$
2. (Опционально) Транформируем M, получаем 'нормализованную' $\tilde{M}$. ($\tilde{M}ij = \tilde{K}_D(x_i, x_j)$)
3. Вычисляем m наибольших положительных собственных значений λ_k и собственных векторов v_k матрицы $\tilde{M}$.
4. Вложение каждой точки x_i:

   $y_{ik} = v_{ki}$ (LLE, Eigenmaps, Spectral Clustering)

   $e_{ik} = \sqrt{\lambda_k} y_{ik}$ (MDS, Isomap)

   Если первые m собственных чисел положительны, тогда $e_i e_j$ - лучшее приближение $\tilde{M}_{ij}$, используя m координат.

Пример (Spectral Clustering)

$$K(a,b) = exp(-\frac{||a-b||^2}{2\sigma^2})$$
$$M_{ij} = K(x_i, x_j)$$
$$\tilde{M}_{ij} = \frac{M_{ij}}{\sqrt{S_i S_j}}, \text{ где } S_i = \sum_{j=1}^{n} M_{ij}$$

Что делать с новыми точками?

Обозначения

Рассмотрим пространство функций $\mathcal{H}_p$ со скалярным произведением
$(f, g)_p = \int f(x)g(x)p(x)dx$, где $p(x)$ - функция плотности.
Определим линейный оператор K_p:
$(K_p f)(x) = \int K(x, y)f(y)p(y)dy$, где $K(\cdot, \cdot)$ - функция близости

Обозначения

Рассмотрим пространство функций $\mathcal{H}_p$ со скалярным произведением
$(f, g)_p = \int f(x)g(x)p(x)dx$, где $p(x)$ - функция плотности.
Определим линейный оператор K_p:
$(K_p f)(x) = \int K(x, y)f(y)p(y)dy$, где $K(\cdot, \cdot)$ - функция близости
Пусть $p = \hat{p}$ - эмпирическая плотность, тогда:

$$(f, g)_{\hat{p}} = \frac{1}{n} \sum_{i=1}^{n} f(x_i)g(x_i)$$

$$(K_{\hat{p}} f)(x) = \frac{1}{n} \sum_{i=1}^{n} K(x, x_i)f(x_i)$$

Предложение 1

Пусть $\tilde{M}_{ij} = \tilde{K}(x_i, x_j)$
(v_k, λ_k) - собственный вектор и собственное значение матрицы $\tilde{M}$
(f_k, λ'_k) - собственная функция и собственное значение оператора $\tilde{K}_{\hat{p}}$
Тогда:

$$\lambda'_k = \frac{1}{n}\lambda_k, \qquad f_k(x) = \frac{\sqrt{n}}{\lambda_k}\sum_{i=1}^{n} v_{ki}\tilde{K}(x, x_i)$$

$$f_k(x_i) = \sqrt{n}v_{ki}, \qquad y_k(x) = \frac{f_k(x)}{\sqrt{n}} = \frac{1}{\lambda_k}\sum_{i=1}^{n} v_{ki}\tilde{K}(x, x_i)$$

$$y_k(x_i) = y_{ik}, \qquad e_k(x_i) = e_{ik}$$

Пример (Laplacian Eigenmaps)

Algorithm 2. *1. Вычисляем* $K(x, x_i)$, $i = 1, \ldots, n$. *Для* i: $x = x_i$ *присваиваем* $K(x, x_i) = \infty$.

2. Определим

$$\bar{K}(a, b) := \frac{1}{n} \frac{K(a, b)}{\sqrt{E_x\left(K(a, x)\right) E_{x'}\left(K(b, x')\right)}}.$$

Так как ни носитель выборки, ни мера неизвестны, то оцениваем математические ожидания простым усреднением по точкам обучающей выборки D.

3. Вычисляем вектор $\hat{M} = (\bar{K}(x, x_1), \ldots, \bar{K}(x, x_n))^T \in \mathbb{R}^n$.

4. Вычисляем искомое y *по формуле* $y = (V_1 | \ldots | V_m)^T \cdot \hat{M}$.