

Методы понижения размерности.

- 1 Доказать, что в рамках метода главных компонент проекции объектов на первую главную компоненту (т.е. соответствующую максимальному собственному значению) имеют наибольшую выборочную дисперсию среди проекций на всевозможные направления в пространстве \mathbb{R}^k .
- 2 Доказать, что сумма квадратов расстояний от объектов до их проекций на линейное подпространство размерности $m < k$ минимально, когда это подпространство есть линейное подпространство с базисом из первых m главных компонент. Также доказать, что в таком случае эта сумма равна $n \sum_{i=m+1}^k \lambda_i$, где λ_i – собственные значения матрицы $\frac{1}{n} X^T X$.
- 3 Выданы данные $\{(y_i, x_{ij}), i = 1, \dots, n + q, j = 1, \dots, k\}$, причем y_{n+1}, \dots, y_{n+q} неизвестны. Используя пройденные методы регрессионного анализа и понижения размерности, учитывая кластеризацию данных, по первым n объектам в рамках линейной регрессионной модели предсказать значения откликов объектов с номерами $n + 1, \dots, n + q$. Описать и объяснить проделанные процедуры.
- 4 Выданы данные $\{(y_i, x_{ij}), i = 1, \dots, n + q, j = 1, \dots, k\}$, причем y_{n+1}, \dots, y_{n+q} неизвестны. Используя пройденные методы регрессионного анализа и понижения размерности, учитывая кластеризацию данных, по первым n объектам предсказать значения откликов объектов с номерами $n + 1, \dots, n + q$. Описать и объяснить проделанные процедуры.
- 5 Выданы данные $\{(y_i, x_{ij}), i = 1, \dots, n, j = 1, \dots, k\}$, причём y_i могут принимать значения 1, 2 или 3. С помощью методов понижения размерности кластеризовать данные, разделить данные на 2 части, визуализировать данные по первой части и классифицировать вторую часть объектов с помощью метода KNN, положив \hat{y}_i равным тому значению отклика, которое встречается чаще среди k ближайших (после кластеризации) к классифицируемому объекту соседей (если 2 значения встречается одинаковое количество раз, то увеличивайте k , пока неоднозначность не пропадёт). Посчитайте относительную долю ошибок классификации на пройденных методах понижения размерности и объясните полученные результаты.
- 6 Загрузить в Python'e датасет `fetch_olivetti_faces` (данные представляют собой монохромные изображения лиц размера 64×64). Спроецировать изображения на пространство, натянутое на k главных компонент. После какого k фотографии теряют индивидуальные различия? С помощью методов понижения размерности спроецировать объекты в \mathbb{R}^2 (использование PCA и t-SNE обязательно) и визуализировать получившиеся данные (образы разных лиц на получившейся картинке отметить разными цветами). Для каких методов получилось четкое разделение на кластеры? Как вы это объясните?