Алгоритмы анализа данных

Урок 8. Снижение размерности данных

Практическое задание

Задание 1: Можно ли отобрать наиболее значимые признаки с помощью PCA?

Опишем суть метода PCA в терминах преобразования линейного n-мерного пространства признаков, где n-количество признаков содержащихся в матрице признаков X нашей задачи.

Пусть имеется некоторе n-мерное пространство S, в котором определён n-мерный базис, состоящих из n базисных векторов $e_i, i = 1, \ldots, n.$

Для простоты и наглядности будем считать, что базис является ортонормированным, то есть

$$e_i \cdot e_j = \delta_{ij}$$

Здесь δ_{ij} - дельта символ Кронекера

$$\delta_{ij} = \begin{cases} 1, i=j \\ 0, i \neq j \end{cases}$$

Пусть вектор X_k - n-мерный вектор нашего пространства признаков, проекции которого x_i^k на базис e_i и есть **строка нашей матрицы признаков** X, то есть набор значений признаков соответствующих y_k, конкретному k-му значению вектора значений y. В нашем базисе вектор X_k можно записать как

$$X_k = \sum_{i=1}^{n} x_i^k e_i$$

.

Зафиксируем, что нашей матрице параметров X, в нашем n-мерном пространстве, соответствует некоторая n-мерная гиперплоскость G. Точками этой гиперплоскости являются точки заданные векторами X_k.

Вращая наш базис e_i в пространстве S, мы можем перейти в новый ортонормированнй базис e'_i. Каждый базисный вектор новой системы координат, может быть выражен как линейная комбинация базисных векторов старой системы координат

$$e'_i = \sum_{j=1}^{n} a_{ij} e_j$$

и наоборот, каждый базисный вектор старой системы координат, может быть выражен как линейная комбинация базисных векторов новой системы координат

$$e_i = \sum_{j=1}^{n} a'_{ij} e'_j$$

.

Матрицы перехода между базисами A и A' связаны между собой ($A' = A^T$).

Каждый n-мерный вектор нашего пространства X_k, в новом базисе вектор X_k можно записать как

$$X_k = \sum_{i=1}^{n} x'^k_i e'_i$$

где x'^k_i - координаты вектора X_k в новой системе координат.

Можно выбрать такой базис, что часть базисных векторов новой системы координат окажутся "ортогональными" нашей n-мерной гиперплоскости параметров G. Пусть количество таких веторов равно m.

Под "ортогональнальностью" мы будем понимать, то что проекция вектора параметров X_k будет равна 0 или близка к нему, то есть $x'^k_s \approx 0$ для некоторых m номеров из набора n.

Соответственно, в этом случае, в новом базисе e' мы можем перейти от n-мерного пространства, к подпространству размерности $n - m < n$, так как

$$X_k = \sum_{i=1}^{n-m} x'^k_i e'_i + \sum_{i=m}^{n} x'^k_i e'_i = \sum_{i=1}^{n-m} x'^k_i e'_i$$

где

$$\sum_{i=m}^{n} x'^k_i e'_i = 0$$

То есть сделать именно то, что нам необходимо, понизить размерность до n-m не теряя качества модели.

В этом как я понимаю и заключается смысл метода главных компонент PCA.

Вывод

Как показали наши рассуждения, метод PCA не выявляет наиболее значимые признаки, а позволяет понизить размерность переходом к новому базису и затем переходу к подпространству меньшей размерности, без потери качества модели.

B [ ]: