Вики

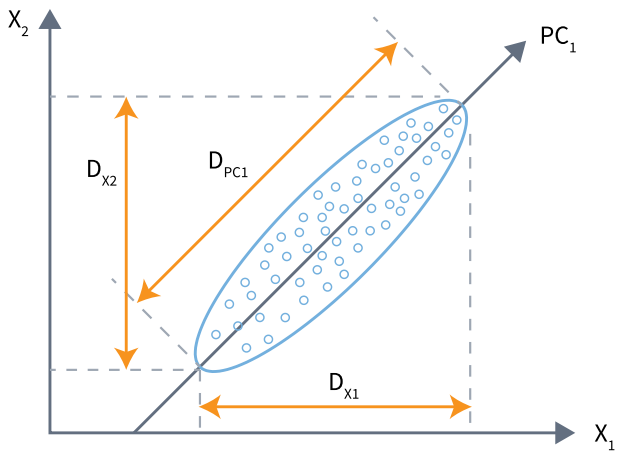
Метод главных компонент (Principal component analysis)

> Синонимы: Преобразование Хоттелинга, PCA
> Разделы: Алгоритмы
> Loginom: Факторный анализ (обработчик)

Метод главных компонент — это технология многомерного статистического анализа, используемая для сокращения размерности пространства признаков с минимальной потерей полезной информации. Предложен К. Пирсоном в 1901 г., а затем детально разработан американским экономистом и статистиком Г. Хоттелингом.

С математической точки зрения метод главных компонент представляет собой ортогональное линейное преобразование, которое отображает данные из исходного пространство признаков в новое пространство меньшей размерности.

При этом первая ось новой системы координат строится таким образом, чтобы дисперсия данных вдоль нее была бы максимальна. Вторая ось строится ортогонально первой так, чтобы дисперсия данных вдоль нее, была бы максимальной их оставшихся возможных и т.д. Первая ось называется первой главной компонентой, вторая — второй и т.д.



На рисунке показано снижение размерности исходного 2-мерного пространства (X_1, X_2) с помощью метода главных компонент до 1-мерного. Первая главная компонента PC_1 ориентирована вдоль направления наибольшей вытянутости эллипсоида рассеяния точек объектов исходного набора данных в пространстве признаков, т.е. с ней связана наибольшая дисперсия.

На рисунке, также, несложно увидеть, что проекция дисперсии данных на ось первой главной компоненты D_{PC1}, больше, чем ее проекции на исходные оси D_{X1} и D_{X2}, но меньше их суммы. Т.е. с помощью первой главной компоненты выразить всю дисперсию данных не удалось. Поэтому строят вторую, третью и т.д. главные компоненты, пока они суммарно не отразят всю дисперсию.

Таким образом, смысл метода заключается в том, что с каждой главной компонентой связана определенная доля общей дисперсии исходного набора данных (ее называют нагрузкой). В свою очередь, дисперсия, являющаяся мерой изменчивости данных, может отражать уровень их информативности.

Действительно, вдоль некоторых осей исходного пространства признаков изменчивость может быть большой, вдоль других — малой, а вдоль третьих вообще отсутствовать. Предполагается, что чем меньше дисперсия данных вдоль оси, тем менее значим вклад переменной, связанной с данной осью и, следовательно, исключив эту ось из пространства (т.е. переменную из модели), можно уменьшить размерность задачи почти не проиграв в информативности данных.

Следовательно, задача метода главных компонент заключается в том, чтобы построить новое пространство признаков меньшей размерности, дисперсия между осями которой будет перераспределена так, чтобы максимизировать дисперсию по каждой из них. Для этого выполняется последовательность следующих действий:

1. Вычисляется общая дисперсия исходного пространства признаков. Это нельзя сделать простым суммированием дисперсий по каждой переменной, поскольку они, в большинстве случаев, не являются независимыми. Поэтому суммировать нужно взаимные дисперсии переменных, которые определяются из ковариационной матрицы.
2. Вычисляются собственные векторы и собственные значения ковариационной матрицы, определяющие направления главных компонент и величину связанной с ними дисперсии.
3. Производится снижение размерности. Диагональные элементы ковариационной матрицы показывают дисперсию по исходной системе координат, а ее собственные значения — по новой. Тогда разделив дисперсию, связанную с каждой главной компонентой на сумму дисперсий по всем компонентам, получаем долю дисперсии, связанную с каждой компонентой. После этого отбрасывается столько главных компонент, чтобы доля оставшихся составляла 80-90%.

Следует отметить, что директивный подход к выбору числа компонент не всегда дает хорошие результаты. Это связано с тем, что часть дисперсии данных может быть обусловлена шумами, а не информативностью компонент. Тогда, задав порог, скажем, 80% может оказаться, что в них только 60% дисперсии связаны с информативностью, а 20 с шумом. Поэтому на практике часто используют различные специальные критерии для определения числа компонент, такие как критерий Кайзера, критерий сломанной трости и т.д.

Основными ограничениями метода главных компонент являются:

- невозможность смысловой интерпретации компонент, поскольку они «вбирают» в себя дисперсию от нескольких исходных переменных;
- метод может работать только с непрерывными данными.

Метод главных компонент включается в состав большинства аналитических платформ и широко используется для снижения размерности входных данных на этапе их предобработки.

Метод иногда рассматривают как часть более общего подхода к снижению размерности данных — факторного анализа. В аналитических платформах в модулях факторного анализа часто практически реализован именно метод главных компонент.