Что такое анализ главных компонентов?

Анализ главных компонентов, или PCA, представляет собой метод уменьшения размерности, который часто используется для уменьшения размерности больших наборов данных путем преобразования большого набора переменных в меньший, который по-прежнему содержит большую часть информации в большом наборе.

Уменьшение количества переменных в наборе данных, естественно, происходит за счет точности, но хитрость в уменьшении размерности заключается в том, чтобы пожертвовать небольшой точностью ради простоты. Потому что меньшие наборы данных легче исследовать и визуализировать, а также значительно упростить и ускорить анализ данных для алгоритмов машинного обучения без обработки посторонних переменных.

Подводя итог, идея PCA проста — уменьшить количество переменных в наборе данных, сохраняя при этом как можно больше информации.

Шаг 1: Стандартизация

Целью этого шага является стандартизация диапазона непрерывных исходных переменных, чтобы каждая из них вносила равный вклад в анализ.

Более конкретно, причина, по которой так важно проводить стандартизацию до PCA, заключается в том, что последний весьма чувствителен к отклонениям исходных переменных. То есть, если существуют большие различия между диапазонами исходных переменных, те переменные с большими диапазонами будут доминировать над переменными с малыми диапазонами (например, переменная, которая находится в диапазоне от 0 до 100, будет доминировать над переменной, которая находится в диапазоне от 0 до 1).), что приведет к необъективным результатам. Таким образом, преобразование данных в сопоставимые масштабы может предотвратить эту проблему.

$$z = \frac{value - mean}{standard\ deviation}$$

Шаг 2: Расчет ковариационной матрицы

Цель этого шага — понять, как переменные набора входных данных отличаются от среднего по отношению друг к другу, или, другими словами, увидеть, есть ли между ними какая-либо связь. Потому что иногда переменные сильно коррелированы таким образом, что содержат избыточную информацию. Итак, чтобы идентифицировать эти корреляции, мы вычисляем ковариационную матрицу.

Ковариационная матрица представляет собой симметричную матрицу размера p × p (где p — число измерений), в которой в качестве элементов используются ковариации, связанные со всеми возможными парами исходных переменных. Например, для трехмерного набора данных с тремя переменными x, y и z ковариационная матрица представляет собой матрицу 3 × 3 из этого:

$$\begin{bmatrix} Cov(x,x) & Cov(x,y) & Cov(x,z) \\ Cov(y,x) & Cov(y,y) & Cov(y,z) \\ Cov(z,x) & Cov(z,y) & Cov(z,z) \end{bmatrix}$$

Поскольку ковариация переменной с самой собой — это ее дисперсия (Cov(a,a)=Var(a)), на главной диагонали (сверху слева направо внизу) мы фактически имеем дисперсии каждой исходной переменной. А поскольку ковариация коммутативна (Cov(a,b)=Cov(b,a)), элементы ковариационной матрицы симметричны относительно главной диагонали, что означает, что верхняя и нижняя треугольные части равны.

Что ковариации, которые мы имеем в качестве элементов матрицы, говорят нам о корреляциях между переменными?

На самом деле важен знак ковариации:

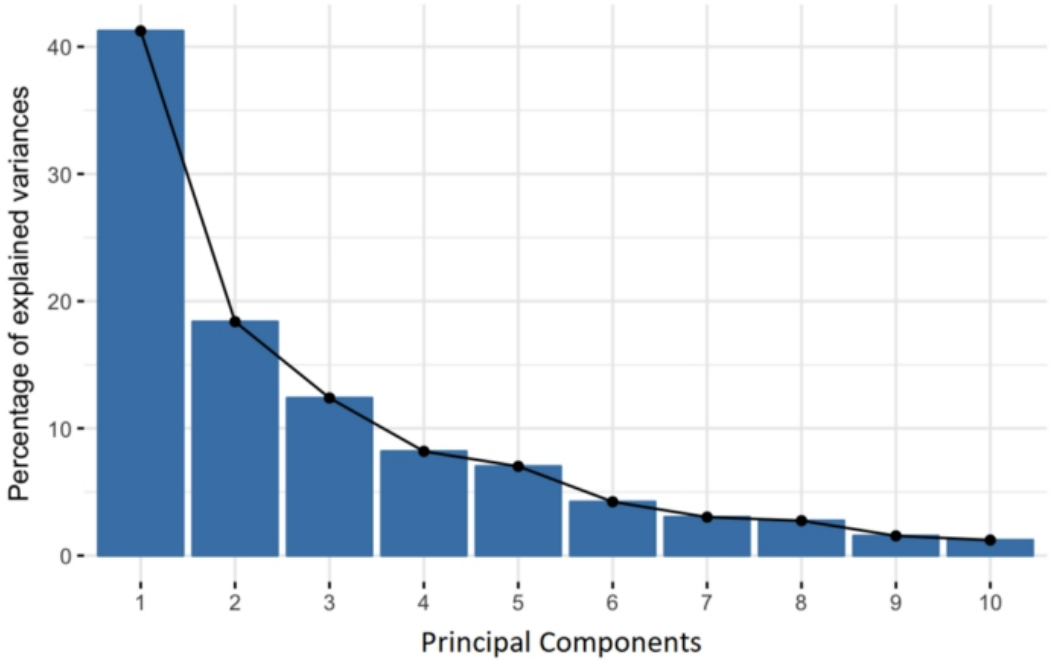
если положительный, то: две переменные увеличиваются или уменьшаются вместе (коррелируют)

если отрицательный, то: один увеличивается, когда другой уменьшается (обратно коррелирует)

Шаг 3: Вычислите собственные векторы и собственные значения ковариационной матрицы, чтобы определить главные компоненты.

Собственные векторы и собственные значения — это понятия линейной алгебры, которые нам нужно вычислить из ковариационной матрицы, чтобы определить главные компоненты данных. Прежде чем перейти к объяснению этих концепций, давайте сначала разберемся, что мы подразумеваем под основными компонентами.

Главные компоненты — это новые переменные, построенные как линейные комбинации или смеси исходных переменных. Эти комбинации выполняются таким образом, что новые переменные (т. е. главные компоненты) не коррелированы, а большая часть информации в исходных переменных сжата или сжата в первые компоненты. Итак, идея состоит в том, что 10-мерные данные дают вам 10 основных компонентов, но PCA пытается поместить максимально возможную информацию в первый компонент, затем максимально оставшуюся информацию во второй и так далее, пока не будет что-то вроде показанного на графике осыпи ниже.



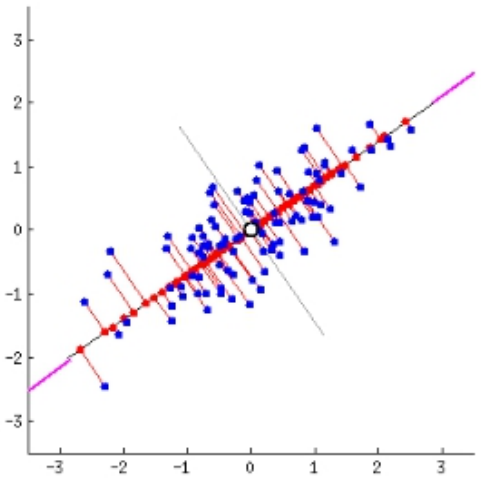
Такая организация информации в главных компонентах позволит вам уменьшить размерность без потери большого количества информации, и это за счет отбрасывания компонентов с малой информацией и рассмотрения оставшихся компонентов в качестве ваших новых переменных.

Здесь важно понимать, что основные компоненты менее интерпретируемы и не имеют никакого реального значения, поскольку они построены как линейные комбинации исходных переменных.

С геометрической точки зрения главные компоненты представляют направления данных, которые объясняют максимальное количество дисперсии, то есть линии, которые фиксируют большую часть информации о данных. Связь между дисперсией и информацией здесь заключается в том, что чем больше дисперсия, переносимая линией, тем больше дисперсия точек данных вдоль нее, и чем больше дисперсия вдоль линии, тем больше информации она содержит. Проще говоря, просто подумайте о главных компонентах как о новых осях, которые обеспечивают лучший угол для просмотра и оценки данных, чтобы различия между наблюдениями были лучше видны.

Как PCA создает основные компоненты

Поскольку основных компонентов столько же, сколько переменных в данных, главные компоненты строятся таким образом, что первый главный компонент учитывает максимально возможную дисперсию в наборе данных. Например, предположим, что диаграмма рассеяния нашего набора данных выглядит так, как показано ниже. Можем ли мы угадать первый главный компонент? Да, это примерно линия, которая соответствует фиолетовым меткам, потому что она проходит через начало координат, и это линия, в которой проекция точек (красные точки) наиболее разбросана. Или, говоря математическим языком, это линия, которая максимизирует дисперсию (среднее значение квадратов расстояний от спроецированных точек (красные точки) до начала координат).



Второй главный компонент рассчитывается таким же образом, но при условии, что он не коррелирует с первым главным компонентом (т. е. перпендикулярен ему) и учитывает следующую по величине дисперсию.

Это продолжается до тех пор, пока не будет вычислено общее количество основных компонентов p, равное исходному количеству переменных.

Теперь, когда мы поняли, что мы подразумеваем под главными компонентами, давайте вернемся к собственным векторам и собственным значениям. Прежде всего вам нужно знать о них то, что они всегда идут парами, так что каждый собственный вектор имеет собственное значение. И их количество равно количеству размерностей данных. Например, для трехмерного набора данных есть 3 переменные, следовательно, есть 3 собственных вектора с 3 соответствующими собственными значениями.

Без дальнейших церемоний, именно собственные векторы и собственные значения стоят за всей магией, объясненной выше, потому что собственные векторы матрицы ковариации на самом деле являются направлениями осей, где существует наибольшая дисперсия (наибольшая информация), и которые мы называем основными компонентами. А собственные значения — это просто коэффициенты, прикрепленные к собственным векторам, которые дают величину дисперсии, содержащейся в каждом основном компоненте.

Ранжируя собственные векторы в порядке их собственных значений, от самого высокого к самому низкому, вы получаете главные компоненты в порядке значимости.

Пример:

Предположим, что наш набор данных является двумерным с двумя переменными x, y и что собственные векторы и собственные значения ковариационной матрицы следующие:
Пример анализа главных компонентов

$$v1 = \begin{bmatrix} 0.6778736 \\ 0.7351785 \end{bmatrix} \qquad \lambda_1 = 1.284028$$

$$v2 = \begin{bmatrix} -0.7351785 \\ 0.6778736 \end{bmatrix} \qquad \lambda_2 = 0.04908323$$

Если мы ранжируем собственные значения в порядке убывания, мы получаем λ1>λ2, что означает, что собственный вектор, соответствующий первой главной компоненте (PC1), равен v1, а второй компонент (PC2) равен v2.

Получив основные компоненты, чтобы вычислить процент дисперсии (информации), приходящийся на каждый компонент, мы делим собственное значение каждого компонента на сумму собственных значений. Если мы применим это к приведенному выше примеру, мы обнаружим, что PC1 и PC2 несут соответственно 96% и 4% дисперсии данных.

Шаг 4: вектор признаков

Как мы видели на предыдущем шаге, вычисление собственных векторов и упорядочение их по собственным значениям в порядке убывания позволяет нам найти главные компоненты в порядке их значимости. На этом этапе мы выбираем, сохранять ли все эти компоненты или отбрасывать менее значимые (с низкими собственными значениями), и формируем с оставшимися матрицу векторов, которую мы называем вектором признаков.

Таким образом, вектор признаков — это просто матрица, столбцы которой содержат собственные векторы компонентов, которые мы решили сохранить. Это делает его первым шагом к уменьшению размерности, потому что, если мы решим оставить только p собственных векторов (компонентов) из n, окончательный набор данных будет иметь только p измерений.

Пример:

Продолжая пример с предыдущего шага, мы можем либо сформировать вектор признаков с обоими собственными векторами v1 и v2:

$$\begin{bmatrix} 0.6778736 & -0.7351785 \\ 0.7351785 & 0.6778736 \end{bmatrix}$$

Или отбросьте собственный вектор v2, который имеет меньшее значение, и сформируйте вектор признаков только с v1:

$$\begin{bmatrix} 0.6778736 \\ 0.7351785 \end{bmatrix}$$

Отбрасывание собственного вектора v2 уменьшит размерность на 1 и, следовательно, приведет к потере информации в окончательном наборе данных. Но учитывая, что v2 нес только 4% информации, потери не будут значительны, и мы по-прежнему будем иметь 96% информации, которую несет v1.

Итак, как мы видели в примере, вам решать, сохранять ли все компоненты или отбрасывать менее важные, в зависимости от того, что вы ищете. Потому что, если вы просто хотите описать свои данные в терминах новых переменных (главных компонентов), которые не коррелированы, не стремясь уменьшить размерность, не нужно исключать менее важные компоненты.

Последний шаг: преобразование данных по осям основных компонентов

На предыдущих шагах, кроме стандартизации, вы не вносите никаких изменений в данные, вы просто выбираете главные компоненты и формируете вектор признаков, но набор входных данных всегда остается в терминах исходных осей (т.е. в терминах исходные переменные).

На этом шаге, который является последним, цель состоит в том, чтобы использовать вектор признаков, сформированный с использованием собственных векторов ковариационной матрицы, чтобы переориентировать данные с исходных осей на те, которые представлены главными компонентами (отсюда и название «Анализ основных компонентов»).). Это можно сделать, умножив транспонирование исходного набора данных на транспонирование вектора признаков.

$$FinalDataSet = FeatureVector^T * StandardizedOriginalDataSet^T$$