2021

Лекция 14. Метод главных компонент

МЕТОДЫ МАШИННОГО ОБУЧЕЧНИЯ

ПАПУЛИН С.Ю. (papulin.study@yandex.ru)

Дополнительные материалы доступны по ссылке: https://github.com/mlmethods

Содержание

1. Метод главных компонент (PCA)

1.1. Общие сведения

Метод главных компонент (Principal Components Analysis – PCA) – популярная техника для уменьшения размерности. На вход поступает множество p размерных данных, и задача PCA заключается в поиске линейного подпространства размерностью m при $m < p$ такого, что точки данных лежали бы преимущественно в этом подпространстве с сохранением вариабельности.

PCA относится к классу методов обучения без учителя, так как используется только множество признаков без целевого значения.

Направления главных компонент представляются как направления в пространстве признаков, вдоль которых исходные данные имеют наибольшую дисперсию.

Метод главных компонент применяют для:

- Уменьшения размерности данных при наличии избыточности (большое количество коррелируемых переменных): может использоваться для сжатия данных или для предобработки данных с последующим применением, например, методов регрессии или классификации
- Визуализации многомерных данных (за счет уменьшения размерности до 2 или 3)

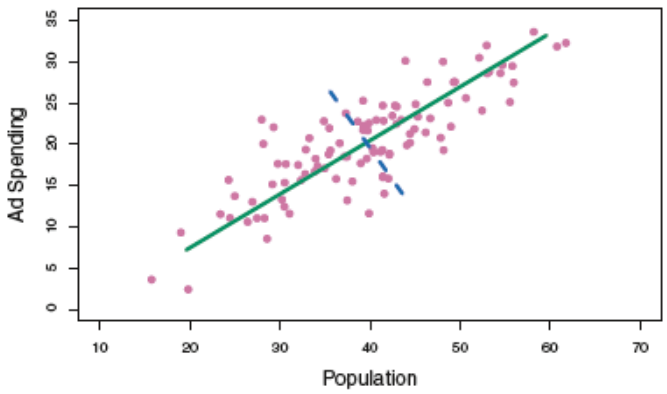

Рисунок – Преобразование исходных данных в новое пространства посредством метода главных компонент [1]

Особенность PCA является его линейность преобразований исходного пространства. Для нелинейных преобразований используется PCA с ядрами (kernel PCA)

1.2. Матрица признаков

Как правило, признаки представляются в виде матрицы

$$X' = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{np} & \cdots & x_{np} \end{bmatrix}^{n \times p} = [\mathrm{x}_1 \quad \cdots \quad \mathrm{x}_p] = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix},$$

где n – количество наблюдений; p – количество признаков.

Для дальнейших рассуждений транспонируем матрицу X', то есть

$$X = X'^T = [x_1 \quad \cdots \quad x_n] = \begin{bmatrix} | & & | \\ x_1 & \cdots & x_n \\ | & & | \end{bmatrix}^{p \times n} = \begin{bmatrix} - & \mathrm{x}_1 & - \\ & \vdots & \\ - & \mathrm{x}_p & - \end{bmatrix}^{p \times n}$$

1.3. Линейное преобразование

PCA использует линейное преобразование пространства. Преобразование данных X в новое пространство записывается как

$$Z = PX$$

где P – матрица трансформации X в Z (геометрически P – матрица вращения и растяжения).

Матрицу трансформации можно представить как множество базисных векторов p_i нового пространства:

$$P = \begin{bmatrix} p_1 \\ \vdots \\ p_p \end{bmatrix} = \begin{bmatrix} - & p_1 & - \\ & \vdots & \\ - & p_p & - \end{bmatrix}^{p \times p},$$

где

$$p_i = [p_{i1} \quad \cdots \quad p_{ip}].$$

Таким образом, линейное преобразование можно записать как

$$Z = PX = \begin{bmatrix} - & p_1 & - \\ & \vdots & \\ - & p_p & - \end{bmatrix}^{p \times p} \begin{bmatrix} | & & | \\ x_1 & \cdots & x_n \\ | & & | \end{bmatrix}^{p \times n} = \begin{bmatrix} p_1 \\ \vdots \\ p_p \end{bmatrix} [x_1 \quad \cdots \quad x_n] = \begin{bmatrix} p_1 \cdot x_1 & \cdots & p_1 \cdot x_n \\ \vdots & \ddots & \vdots \\ p_p \cdot x_1 & \cdots & p_p \cdot x_n \end{bmatrix}$$
$$= \begin{bmatrix} | & & | \\ z_1 & \cdots & z_n \\ | & & | \end{bmatrix} = [z_1 \quad \cdots \quad z_n],$$

где z_i – представление x_i в новом пространстве, т. е. проекция x_i в новый базис $p_1, \ldots, p_p$; z_{ij} – проекция x_i по направлению вектора p_j.

При PCA задача заключается в том, чтобы найти некоторое преобразование P, при котором признаки нового пространства будут независимы. Рассмотрим, что это означает.

1.4. Матрица ковариации

На рисунке 1(a) представлены данные по двум признакам без избыточности, что соответствует отсутствию корреляции.

На рисунке 1(c) наблюдается ярко выраженная корреляция двух переменных. В этом случае было бы более правильным использовать одну переменную в виде линейной комбинации (например, $r_2 - kr_1$) вместо двух отдельных переменных.

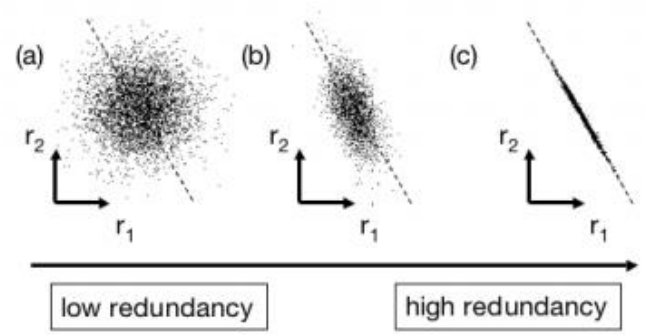

Рисунок – Корреляция случайных величин [3]

Чтобы выразить избыточность между признаками, мы можем вычислить ковариационную матрицу.

Ковариация признаков i и j представляется как

$$\sigma_{i,j}^2 = \frac{1}{n-1}(\mathrm{x}_i - \bar{\mathrm{x}}_i)\left(\mathrm{x}_j - \bar{\mathrm{x}}_j\right)^T$$

где x_i и x_j – векторы значений признаков i и j по всем n наблюдениям; $\bar{\mathrm{x}}_i$ и $\bar{\mathrm{x}}_j$ – средние значения по признакам i и j.

Векторы значений признаков i и j есть

$$\mathrm{x}_i = [x_{11} \quad x_{12} \quad \dots \quad x_{1n}],$$
$$\mathrm{x}_j = [x_{21} \quad x_{22} \quad \dots \quad x_{2n}].$$

Замечание

В данном случае векторы x_i и x_j представляются в виде строки, чтобы согласовать с ранее введенными обозначениями для матрицы признаков:

$$X = [x_1 \quad \cdots \quad x_n] = \begin{bmatrix} | & & | \\ x_1 & \cdots & x_n \\ | & & | \end{bmatrix}^{p\times n} = \begin{bmatrix} - & \mathrm{x}_1 & - \\ & \vdots & \\ - & \mathrm{x}_p & - \end{bmatrix}^{p\times n}$$

Далее будем считать, что признаки центрированы, то есть

$$\sigma_{i,j}^2 = \frac{1}{n-1}\mathrm{x}_i\mathrm{x}_j{}^T.$$

Ковариационная матрица содержит ковариации между всеми p признаками:

$$S_X = \frac{1}{n-1} XX^T = \frac{1}{n-1} \begin{bmatrix} - & \mathrm{x}_1 & - \\ & \vdots & \\ - & \mathrm{x}_p & - \end{bmatrix}^{p \times n} \begin{bmatrix} | & & | \\ \mathrm{x}_1 & \cdots & \mathrm{x}_\mathrm{n} \\ | & & | \end{bmatrix}^{n \times p} = \begin{bmatrix} \sigma_{1,1}^2 & \cdots & \sigma_{1,p}^2 \\ \vdots & \ddots & \vdots \\ \sigma_{p,1}^2 & \cdots & \sigma_{p,p}^2 \end{bmatrix}^{p \times p}$$

Свойства ковариации:

- $\sigma_{i,j}^2 = 0$ тогда и только тогда, когда признаки i и j совершенно не коррелируют.
- $\sigma_{i,j}^2 = \sigma_i^2$, если $i = j$

Свойства матрицы ковариации:

- S_X – симметричная матрица
- элементы по диагонали есть дисперсия признаков σ_i^2
- элементы вне диагонали есть ковариации признаков $\sigma_{i,j}^2$

1.5. Диагонализация матрицы

Если наша цель заключается в уменьшении избыточности, то необходимо, чтобы ковариация между признаками стремилась к нулю. В этом случае после преобразования P, в результате которого получаем Z (то есть признаки X в новом пространстве, что соответствует $Z = PX$), матрица ковариации S_Z должна содержать нулевые значения вне диагонали.

С учетом этого обозначим нашу цель следующим образом.

Необходимо найти такое преобразование P, при котором матрица ковариации S_Z будет диагональной.

При использовании метода главным компонент мы предполагаем, что базисные векторы $\{p_1, p_2, \ldots, p_m\}$ образуют систему ортонормированных векторов, то есть

$$p_i . p_j \; = \; \delta_{ij} = \begin{cases} 1, \text{если } i = j \\ 0 \text{ иначе} \end{cases}$$

Соответственно, P является ортонормальной матрицей.

Кроме того, полагаем, что направления с наибольшей дисперсией являются наиболее значимыми или главными.

В простой форме суть процесса работы метода главных компонент можно представить следующим образом. PCA сначала выбирает нормализованное направление в p размерном пространстве, вдоль которого наблюдается наибольшая дисперсия. Данное направление представляется как вектор p_1 – первая главная компонента. После этого он ищет следующее направление с максимальной дисперсией, при этом оно должно быть ортогонально p_1. В результате получаем вектор p_2 – вторая главная компонента. Продолжаем в той же манере для всех p признаков, получим в итоге упорядоченное множество p главных компонент.

Таким образом, дисперсия по направлению p_i определяет важность самого направления, и в соответствие с этим значением необходимо упорядочить все базисные векторы.

Условия и ограничения

- линейность преобразований
- вероятностное распределении по x_i должно быть нормальным
- главные компоненты – ортогональны

1.6. Задача PCA

Найти главные компоненты

Данная задача сводится к преобразованию наблюдений X в новое пространство:

$$Z = PX$$

При этом

$$S_Z = \frac{1}{n-1} ZZ^T$$

должна быть диагональной.

Строки итоговой матриц P есть главные компоненты.

Решение

Запишем ковариационную матрицу S_Z в следующем виде

$$S_Z = \frac{1}{n-1} ZZ^T = \frac{1}{n-1} (PX)(PX)^T = \frac{1}{n-1} PXX^T P^T = \frac{1}{n-1} P \underbrace{(XX^T)}_{A} P^T$$

Получаем

$$S_Z = \frac{1}{n-1} PAP^T,$$

где A – симметричная матрица, так как $A = XX^T$.

Симметричную матрицу A можно разложить на матрицы собственных векторов и чисел следующим образом

$$A = EDE^{-1},$$

где E – ортогональная матрица собственных векторов $e_1, \dots, e_p$:

$$E = [e_1 \quad \cdots \quad e_p] = \begin{bmatrix} | & & | \\ e_1 & \cdots & e_p \\ | & & | \end{bmatrix}^{p \times p};$$

D – диагональная матрица собственных значений:

$$D = \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_p \end{bmatrix}$$

С учетом того, что обратная ортогональная матрица равна транспонированной матрице, то есть

$$E^{-1} = E^T,$$

получаем

$$A = EDE^T.$$

Матрица A имеет $r \leq p$ ортонормальных векторов, где r – ранг матрицы. Ранг матрицы A меньше чем p, когда A – вырожденная, т.е. есть линейная зависимость, или все данные умещаются в подпространстве размерностью $r \leq p$.

Предположим, что

$$P = E^T$$

Тогда

$$S_Z = \frac{1}{n-1} PAP^T = \frac{1}{n-1} P(P^T DP)P^T = \frac{1}{n-1} (PP^T)D(PP^T) = \frac{1}{n-1} (PP^T)D(PP^T)$$
$$= \frac{1}{n-1} (PP^{-1})D(PP^{-1}).$$

В итоге получаем

$$S_Z = \frac{1}{n-1} D$$

Таким образом, получаем диагональную ковариационную матрицу S_Z после преобразования P над исходными данными X.

- Так как в этом случае в качестве матрицы преобразования P использовалась транспонированная матрица собственных векторов симметричной матрицы $A = XX^T$, то главными компонентами будут собственные векторы.
- Диагональные значения матрицы S_Z соответствуют дисперсии вдоль направлений главных компонент.

Главные компоненты $p_1, \dots, p_p$:

$$P = \begin{bmatrix} | & & | \\ e_1 & \cdots & e_n \\ | & & | \end{bmatrix}^T = \begin{bmatrix} - & p_1 & - \\ & \vdots & \\ - & p_p & - \end{bmatrix}^{p \times n}$$

Замечание

Собственные векторы упорядочены по значению собственных чисел, поэтому первая главная компонента это та, вдоль которой наблюдается наибольшая дисперсия.

В результате вычисление главных компонент сводится к вычислению собственных векторов матрицы XX^T.

2. Уменьшение размерности

Уменьшение размерности посредством метода главных компонент заключается в исключении отдельных компонент. Так если необходимо уменьшить размерность с p до m при $m < p$, то исключаются все компоненты от $m+1$ до p, то есть оставляем первые m компонент, которые соответствуют направлениям с наибольшей дисперсией:

$$\begin{bmatrix} - & p_1 & - \\ & \vdots & \\ - & p_p & - \end{bmatrix} \mapsto \begin{bmatrix} - & p_1 & - \\ & \vdots & \\ - & p_m & - \end{bmatrix}$$

Таким образом, матрица главных компонент примет вид

$$P^* = \begin{bmatrix} - & p_1 & - \\ & \vdots & \\ - & p_m & - \end{bmatrix}^{m \times p}$$

Преобразование исходных данных X в уменьшенное пространства можно представить как

$$Z^* = P^* X = \begin{bmatrix} - & p_1 & - \\ & \vdots & \\ - & p_m & - \end{bmatrix}^{m \times p} \begin{bmatrix} | & & | \\ x_1 & \cdots & x_n \\ | & & | \end{bmatrix}^{p \times n} = \begin{bmatrix} p_1 \cdot x_1 & \cdots & p_1 \cdot x_n \\ \vdots & \ddots & \vdots \\ p_m \cdot x_1 & \cdots & p_m \cdot x_n \end{bmatrix}^{m \times n} = \begin{bmatrix} | & & | \\ z_1^* & \cdots & z_n^* \\ | & & | \end{bmatrix}$$

Как определить значение m (размерность нового пространства)? Для этого введем понятия доли объяснимой дисперсии и кумулятивной энергии.

Доля объяснимой дисперсии для i компоненты:

$$\frac{\lambda_i}{\sum_{k=1}^{p} \lambda_k}$$

Доля объяснимой дисперсии для первых m компонент:

$$\frac{\sum_{j=1}^{m} \lambda_j}{\sum_{k=1}^{p} \lambda_k}$$

Выражение выше можно переписать в следующем виде

$$\frac{g_m}{g_p},$$

где g_i называется кумулятивной энергией и вычисляется как

$$g_i = \sum_{k=1}^{i} \lambda_k$$

Как правило, используют значение объяснимой дисперсии в районе 0.9, т. е.

$$\frac{g_m}{g_p} \geq 0.9$$

Обратное преобразование в исходное пространство

TODO

3. Сингулярное разложение (SVM)

TODO

4. Линейная регрессия с использованием PCA

Главные компоненты регрессии (Principal Components Regression – PCR) – подход для задачи регрессии с применением метода главных компонент, который заключается в использовании первых m главных компонент в качестве предикторов в модели линейной регрессии.

В данном случае предполагаем, что небольшое количество главных компонент вполне достаточно для объяснения большей части разброса данных и взаимосвязи с откликом y. Кроме того, обобщение данных с использованием пространства с меньшим количеством признаков позволяет избежать переобучения.

В целом техники по уменьшению размерности для задач регрессии и классификации работаю в два этапа:

- Трансформация исходного пространства наблюдений X с p признаками в новое пространство Z с m признаками, такое что $m < p$.
- Обучение модели посредством Z

Для выбора значения m может быть использована доля объяснимой дисперсии или кросс-валидация. Кроме того, перед применением PCA желательно стандартизовать признаки, чтобы они были в одном масштабе.

Рассмотрим более подробно случай с линейной регрессией. В общем виде линейная регрессия имеет вид

$$y_i = \theta_0 + \theta^T x_i + \epsilon_i = \theta_0 + \sum_{k=1}^{p} \theta_k x_{ik} + \epsilon_i\,, \quad i = 1, \dots, n$$

В качестве модели предсказания используем запись

$$h_i = \theta_0 + \theta^T x_i = \theta_0 + \sum_{k=1}^{p} \theta_k x_{ik}, \quad i = 1, \dots, n$$

В этом случае неизвестными являются параметры θ, которые необходимо оценить посредством одной из техник, например, методом наименьших квадратов.

Множество наблюдений представим как матрицу

$$X = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}^{n \times p}$$

Применив PCA преобразование и уменьшив размерности до m, что можно записать как

$$Z^* = {P^*}^T X,$$

получим наблюдения в новом пространстве

$$Z^* = \begin{bmatrix} z_1 \\ \vdots \\ z_n \end{bmatrix}^{n \times m}$$

Матрицу преобразования запишем как

$$P^* = [p_1 \quad \cdots \quad p_m]^{p \times m}$$

В данном случае для удобства мы трансформировали ранее рассмотренные значения для X, Z^* и P^*.

Уменьшение размерности

Уменьшим исходное пространство из p признаков до m размерного пространства посредством PCA. Тогда преобразование для некоторого наблюдения x будет иметь вид

$$z = P^* x = \begin{bmatrix} {p_1}^T x \\ \vdots \\ {p_m}^T x \end{bmatrix},$$

где

$$p_j = \begin{bmatrix} p_{1j} \\ \vdots \\ p_{pj} \end{bmatrix}$$

Проекция x на координату j в новом m-размерном пространстве есть

$$z_j = {p_j}^T x = \sum_{k=1}^{p} p_{kj} x_k$$

Получив значения наблюдений в новом пространстве, можно переписать выражение для линейной регрессии как

$$y_i = \beta_0 + \beta^T z_i + \epsilon_i = \beta_0 + \sum_{j=1}^{m} \beta_j z_{ij} + \epsilon_i, \quad i = 1, \dots, n$$

где β_j – параметр регрессии в новом пространстве.

Можно показать взаимосвязь между θ и β следующим образом

$$\sum_{j=1}^{m} \beta_j z_{ij} = \sum_{j=1}^{m} \beta_j \sum_{k=1}^{p} p_{kj} x_{ik} = \sum_{k=1}^{p} \sum_{j=1}^{m} \beta_j p_{kj} x_{ik} = \sum_{k=1}^{p} \theta_k x_{ik}$$

Таким образом получаем, что при уменьшении размерности посредством PCA мы добавляем ограничения на значения оценки параметров θ, т.е.

$$\theta_k = \sum_{j=1}^{m} \beta_j p_{kj}$$

Такого рода ограничение может привести к смещению оценок параметров. Однако в ситуации, когда p имеет большое значение по сравнению с n, выбор $m \ll p$ может существенно уменьшить дисперсию оценок параметров модели. Случай при $m = p$ соответствует отсутствию ограничений.

Список литературы

1. Chapter 6. Dimension Reduction Methods // An Introduction to Statistical Learning by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshir. pp. 228–237. URL: http://faculty.marshall.usc.edu/gareth-james/ISL/
2. Chapter 10. Principal Components Analysis // An Introduction to Statistical Learning by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshir. pp. 374–284. URL: http://faculty.marshall.usc.edu/gareth-james/ISL/
3. Shlens, J. (2003). A tutorial on principal component analysis, URL: https://www.cs.princeton.edu/picasso/mats/PCA-Tutorial-Intuition_jp.pdf