Работа № 3.3 Анализ главных компонент

1. Теоретические сведения

Это один из основных алгоритмов машинного обучения. Позволяет уменьшить размерность данных, потеряв наименьшее количество информации. Вычисление главных компонент сводится к вычислению собственных векторов и собственных значений ковариационной матрицы исходных данных или к сингулярному разложению матрицы данных.

Метод главных компонент (МГК, англ. principal component analysis, PCA) — это техника машинного обучения, которая используется для изучения взаимосвязей между наборами переменных. Другими словами, МГК изучает наборы переменных для того, чтобы определить базовую структуру этих переменных. МГК еще иногда называют факторным анализом.

Основные приложения

- Dimensionality reduction. Снижение размерности данных при сохранении всей или большей части информации

Метод главных компонент используется для преобразования набора данных с множеством параметров в новый набор данных с меньшим количеством параметров и каждый новый параметр этого набора данных — это линейная комбинация раннее существующих параметров. Эти преобразованные данные стремятся обосновать большую часть дисперсии оригинального набора данных с гораздо большей простотой.

- Feature extraction. Выявление и интерпретация скрытых признаков

Нередко в машинном обучении встречаются ситуации, когда данные собираются априори, и лишь затем возникает необходимость разделить некоторую выборку по известным классам. Как следствие часто может возникнуть ситуация, когда имеющийся набор признаков плохо подходит для эффективной классификации. По крайней мере, при первом приближении.

В такой ситуации можно строить композиции слабо работающих по отдельности методов, а можно начать с обогащения данных путём выявления скрытых зависимостей между признаками. И затем строить на основе найденных зависимостей новые наборы признаков, некоторые из которых могут потенциально дать существенный прирост качества классификации.

Различия линейной регрессии и МГК

Линейная регрессия определяет линию наилучшего соответствия через набор данных. Метод главных компонентов определяет несколько ортогональных линий наилучшего соответствия для набора данных.

2. Задача: проанализировать заемщиков банка на основе различных данных

Пример источника данных: GiveMeSomeCredit
https://www.kaggle.com/c/GiveMeSomeCredit

Variable Name	Description	Type
RevolvingUtilizationOfUnsecuredLines	Total balance on credit cards and personal lines of credit except real estate and no installment debt like car loans divided by the sum of credit limits Общий баланс по кредитным картам и личным кредитным линиям, за исключением долга по недвижимости	percentage

	и без рассрочки, например автокредитов, деленный на сумму кредитных лимитов	
Age	Age of borrower in years Возрастзаемщика в годах	integer
NumberOfTime30-59DaysPastDueNotWorse	Number of times borrower has been 30-59 days past due but no worse in the last 2 years. Количество просроченных платежей заемщика на 30-59 дней, но не больше чем за последние 2 года.	integer
DebtRatio	Monthly debt payments, alimony, living costs divided by monthy gross income Ежемесячные выплаты по долгу, алименты, расходы на жизнь, разделенные на ежемесячный валовой доход	percentage
MonthlyIncome	Monthly income Ежемесячный доход	real
NumberOfOpenCreditLinesAndLoans	Number of Open loans (installment like car loan or mortgage) and Lines of credit (e.g. credit cards) Количество открытых займов (рассрочка, например, автокредит или ипотека) и кредитных линий (например, кредитные карты)	integer
NumberOfTimes90DaysLate	Number of times borrower has been 90 days or more past due. Количество просроченных платежей заемщика на 90 дней или более.	integer
NumberRealEstateLoansOrLines	Number of mortgage and real estate loans including home equity lines of credit Количество ипотечных кредитов и ссуд на недвижимость, включая кредитные линии под залог собственного капитала	integer
NumberOfTime60-89DaysPastDueNotWorse	Number of times borrower has been 60-89 days past due but no worse in the last 2 years. Количество раз, когда заемщик просрочил платеж на 60-89 дней, но не больше чем за последние 2 года.	integer
NumberOfDependents	Number of dependents in family excluding themselves (spouse, children etc.) Количество иждивенцев в семье, исключая их самих (супруга, дети и т. д.)	integer

Пример: Give Me Some Credit

Revolving Utilization Of Unsecured Lines	age	Number Of Time 30-59 Days Past Due Not Worse	Debt Ratio	Monthly Income	Number Of Open Credit Lines And Loans	Number Of Times 90 Days Late	Number Real Estate Loans Or Lines	Number Of Time 60-89 Days Past Due Not Worse	Number Of Dependents
0.766126609	45	2	0.802982129	9120	13	0	6	0	2
0.957151019	40	0	0.121876201	2600	4	0	0	0	1
0.65818014	38	1	0.085113375	3042	2	1	0	0	0
0.233809776	30	0	0.036049682	3300	5	0	0	0	0
0.9072394	49	1	0.024925695	63588	7	0	1	0	0
0.213178682	74	0	0.375606969	3500	3	0	1	0	1
0.305682465	57	0	5710	NA	8	0	3	0	0
0.754463648	39	0	0.209940017	3500	8	0	0	0	0
0.116950644	27	0	46	NA	2	0	0	0	NA
0.189169052	57	0	0.606290901	23684	9	0	4	0	2
0.644225962	30	0	0.30947621	2500	5	0	0	0	0
0.01879812	51	0	0.53152876	6501	7	0	2	0	2
0.010351857	46	0	0.298354075	12454	13	0	2	0	2
0.964672555	40	3	0.382964747	13700	9	3	1	1	2

3. Задача снижения размерности

Представить набор данных меньшим числом признаков таким образом, чтобы потеря информации, содержащейся в оригинальных данных, была минимальной.

Принципы компонентного анализа

Данные заданы матрицей $X = (x_i^j)$ размерности $n \times m$, где $i = \overline{1,n}$ и $j = \overline{1,m}$, n – число наблюдений (объектов), m – число признаков.

Обозначим за $C\ (m \times m)$ матрицу ковариаций признаков матрицы X:

$$c_{ij} = \frac{\sum_{p=1}^{n} x_k^i x_k^j}{n} - \mu_i \mu_j, \forall i,j \in \{1 \dots m\},$$

$$\mu_i - среднее значение признака i, i \in \{1 \dots m\}$$

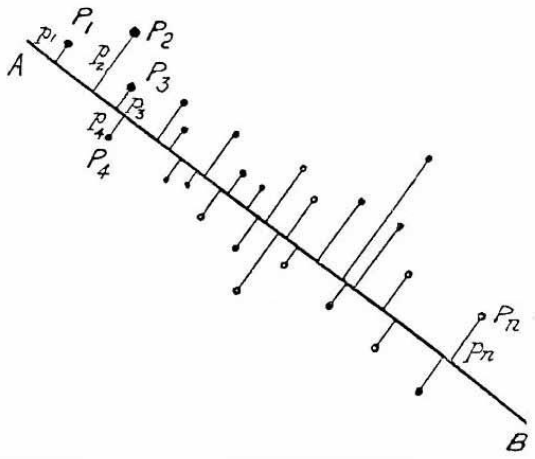
В матричном виде:

$$C = \frac{X^T X}{n} - \mu^T \mu, \mu = (\mu_1 \dots \mu_m)$$

Вариация i-го признака: $Var(x^i) = c_{ii}$

Общая вариация данных: $Var(X) = \sum_{i=1}^{m} c_{ii}$

Задача: найти ортогональные векторы такие, что $v^T C v \to max,$ т.е. проекция данных, на которые позволит сохранить наибольшую вариацию



Матрица C симметричная и положительно определена. Имеет место равенство:

$$C = V \Lambda V^T$$

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \lambda_m \end{bmatrix},$$

$$\lambda - собственные значения матрицы C, \sum_{i=1}^{m} \lambda_i = \sum_{i=1}^{m} c_{ii}, \lambda_1 > \lambda_2 \geq \dots \geq \lambda_m \geq 0$$

$$V(m \times m) - матрица собственных векторов матрицы C$$

Главные компоненты:

$$U = X \cdot \lfloor v^1, v^2, \dots, v^k \rfloor^T, k < m$$

Доля объясненной вариации:

$$\frac{\sum_{i=1}^{k} \lambda_i}{Var(X)}$$

Singular value decomposition

- Данные заданы матрицей $X = (x_i^j)$ размерности $n \times m$, где $i = \overline{1,n}$ и $j = \overline{1,m}$, n – число наблюдений (объектов), m – число признаков.
- Требуется среди всех матриц такого же размера $n \times m$и ранга $\leq k$найти матрицу Y, для которой норма матрицы $\|X - Y\|$ будет минимальной.
- Решение зависит от матричной нормы
- Наиболее подходящие: Евклидова норма и норма Фробениуса:
    - Евклидова норма: $\|A\|_2 = \sqrt{\lambda T_{max}}$, где λ_{max} - максимальное собственное значение матрицы A
    - Норма Фробениуса: $\|A\|_F = \sqrt{\sum_i \sum_i a_{ij}^2}$

Существуют такие матрицы Uи V, что выполняется равенство $X = U \cdot S \cdot V^T$,

где U – матрица собственных векторов матрицы $X \cdot X^T$, V – матрица собственных векторов матрицы $X^T \cdot X$, а матрица S размерности $n \times m$имеет на главной диагонали элементы $\sigma_1, \sigma_2, \ldots, \sigma_m$ и все остальные нули, где σ_i - сингулярные числа матрицы X, а σ_i^2 - собственные числа матрицы $X^T \cdot X$.

Запишем матрицы Uи V в векторном виде:

$$U = \lfloor u^1, u^2, \ldots, u^n \rfloor, \rightleftarrows V = \lfloor v^1, v^2, \ldots, v^m \rfloor$$

Тогда SVD разложение можно представить как

$$X = \sigma_1 u^1 (v^1)^T + \sigma_2 u^2 (v^2)^T + \cdots + \sigma_m u^m (v^m)^T$$

Теорема Шмидта-Мирского:

Решением матричной задачи наилучшей аппроксимации в норме Евклида и в норме Фробениуса является матрица $X^* = \sigma_1 u^1 (v^1)^T + \sigma_2 u^2 (v^2)^T + \cdots + \sigma_k u^k (v^k)^T$

Ошибки аппроксимации:

$$\|X - X^*\|_2 = \sigma_{k+1}$$

$$\|X - X^*\|_F = \sqrt{\sigma_{k+1}^2 + \sigma_{k+2}^2 + \cdots \sigma_m^2}$$

Выбор числа k главных факторов

Общая вариация данных:

$$Var(X) = \sigma_1^2 + \sigma_2^2 + \ldots + \sigma_m^2$$

Доля объясненной вариации:

$$\frac{\sigma_1^2 + \sigma_2^2 + \ldots + \sigma_k^2}{Var(X)}, k < m$$

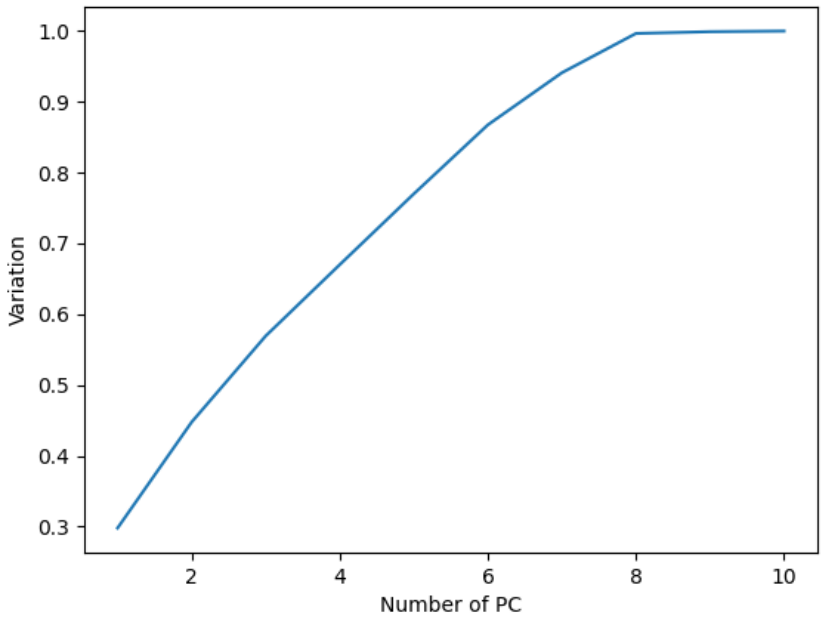
Хорошим значением считается доля объясненной вариации ≥ 80%

4. Решениевscikit-learn

```
import numpyas np
import scipyas sp
from sklearn.decompositionimport PCA
import matplotlib.pyplotas plt
from sklearn.preprocessingimport scale

np.set_printoptions(precision=10,
threshold= 10000,
suppress= True)
```

```
# Загружаем данные и удаляем наблюдения с пропущенными значениями
data = np.genfromtxt("cs-data.csv", delimiter = ',',
skip_header= 1, usecols=list(range(1, 11)))
data = data[~np.isnan(data).any(axis = 1)]

# Выполняем метод главных компонент
data = scale(data)
pca = PCA(svd_solver='full')
pca.fit(data)

print("Размерность данных \n", data.shape, "\n")
# Вклад каждого фактора в объяснение вариации
print("Вклад каждого фактора в объяснение вариации \n",
pca.explained_variance_ratio_, "\n")
# Рост доли объясненной вариации с увеличением числа главных факторов
var = np.round(np.cumsum(pca.explained_variance_ratio_), decimals=4)
print("Рост доли объясненной вариации с увеличением числа главных факторов
\n", var, "\n")
plt.figure()
plt.plot(np.arange(1,11), var)
plt.ylabel('Variation')
plt.xlabel('Number of PC')
plt.show()
```

```
Размерность данных
 (201669, 10)

Вклад каждого фактора в объяснение вариации
 [0.2979766397 0.1496007962 0.1217110055 0.1007219879 0.0999739517
 0.0975640598 0.0735527529 0.055468798  0.0024871325 0.0009428757]

Рост доли объясненной вариации с увеличением числа главных факторов
[0.298  0.4476 0.5693 0.67   0.77   0.8675 0.9411 0.9966 0.9991 1.    ]
```



5. Задание

1. Воспроизведите вычисления, представленные в теоретическом материале практической работы. Подтвердите выводы.

2. Рассмотрите набор данных Turkiye Student Evaluation:
    a) Опишите исследуемые данные
    b) Выберите данные по одному предмету (class) и выполните анализ главных компонент. Выделите главные факторы, дайте интерпретацию (или покажите, что этого сделать нельзя).
    c) Выберите два предмета, которые проводил один и тот же преподаватель. Снова выполните анализ главных компонент, выделите главные факторы, постарайтесь дать интерпретацию. Сравните результаты с предыдущим пунктом.
    d) Выполните PCA для всего набора данных. Также сравните результаты с пунктами выше.
    e) Повторите вычисления из пунктов b - d, но для нестандартизованных данных. Сравните с соответствующими результатами, полученными на стандартизованных данных.