**МИНОБРНАУКИ РОССИИ**

**Санкт-Петербургский государственный**

**электротехнический университет**

**«ЛЭТИ» им. В.И. Ульянова (Ленина)**

Доклад

**Тема: Методы редукции данных**

Выполнили: Чепасов Д.В, Щедрин А.А.

Группа:2382

Санкт-Петербург

2023

Оглавление

[PCA (Метод главных компонентов) 3](#_Toc130653684)

[**1.1)** **Математическое обоснование** 3](#_Toc130653685)

[**1.2)** **Роль в машинном обучении** 4](#_Toc130653686)

[**1.3)** **Реализация в прикладной задаче** 5](#_Toc130653687)

[**1.4)** **Плюсы и минусы метода** 6](#_Toc130653688)

[Корреляционный анализ 7](#_Toc130653689)

[**2.1) Математическая теория** 7](#_Toc130653690)

[**2.1.1) Основные понятия** 7](#_Toc130653691)

[**2.1.2) Различия корреляционных связей** 8](#_Toc130653692)

[**2.1.3) Вариационный ряд и ранговая корреляция.** 8](#_Toc130653693)

[**2.1.4) Коэффициент корреляция Пирсона.** 9](#_Toc130653694)

[**2.1.6) Коэффициент частной корреляции.** 10](#_Toc130653695)

[2.2) Роль в машинном обучении 12](#_Toc130653696)

[**2.3) Реализация для прикладной задачи** 13](#_Toc130653697)

[2.4) Плюсы и минусы метода 14](#_Toc130653698)

[Слабые стороны корреляционного анализа: 14](#_Toc130653699)

[Метод максимального правдоподобия 15](#_Toc130653700)

[**3.1) Математическая теория** 15](#_Toc130653701)

[**3.2) Роль в машинном обучении** 16](#_Toc130653702)

[**3.3) Реализация для прикладной задачи** 17](#_Toc130653703)

[**3.4) Плюсы и минусы метода** 18](#_Toc130653704)

[Сравнение методов 19](#_Toc130653705)

**PCA (Метод главных компонентов)**

* 1. **Математическое обоснование**
  2. **Роль в машинном обучении**
  3. **Реализация в прикладной задаче**
  4. **Плюсы и минусы метода**

**Корреляционный анализ**

**2.1) Математическая теория**

Основной задачей корреляционного анализа состоит в количественном определении тесноты связи между двумя признаками при парной связи и между результативной и несколькими факторными признаками при многофакторной связи и статистической оценке надежности установленной связи.Корреляционная связь существует там, где взаимосвязанные явления характеризуются только случайными величинами. При такой связи среднее значение (математическое ожидание) случайной величины результативного признака у закономерно изменяется в зависимости от изменения другой величины х или других случайных величин x1,x2,...,xn. Корреляционная связь проявляется не в каждом отдельном случае, а во всей совокупности в целом. Только при достаточно большом количестве случаев каждому значению случайного признака х будет соответствовать распределение средних значений случайного признака у. Наличие корреляционных связей присуще многим общественным явлениям.

**2.1.1) Основные понятия**

Корреляция – статистическая взаимосвязь двух или более переменных, в которой при изменении одной или нескольких величин, меняется другие величины или другая, но одна величина.

Корреляционный анализ – раздел математической статистки, которые изучает эти взаимосвязи.

Корреляционный связь – это согласованное изменение двух признаков, отражающее тот факт, что изменчивость одного признака находится в соответствии с изменчивостью другого. Т.е. каждой независимой переменной x соответствует определенное математическое ожидание значения y.

Группировочный признаки – признаки, по которым осуществляется разделение единиц изучаемой статистической совокупности на качественно однородные группы.

Факторный признаки – признаки, определяющие изменение показателей совокупности.

Результативные признаки – изменяются под влиянием факторных признаков.

Парная корреляция – связь между двумя признаками.

Частная корреляция – связь между двумя признаками, при фиксированных значениях других факторных признаков.

Множественная корреляция - связь между результативным и двумя или более факторными признаками, включенными в исследование.

**2.1.2) Различия корреляционных связей**

Корреляционные связи могут различаться по трем свойствам:

1. *По форме.* По форме корреляционная связь может быть прямолинейном или криволинейной. Примером прямолинейной связи можно назвать связь между количеством тренировок и количеством решенных задач на контрольной сессии.
2. *По направлению.* Корреляционная связь может быть положительной и отрицательной. При положительной корреляционной связи более высоким значениям одного признака соответствуют более высокие значения другого, при более низких значениях одного признака соответствуют более низкие значения другого. При отрицательной связи обратные соотношения: более высоким соответствуют более низкие, более низким соответствуют более высокие.
3. *По тесноте (силе или степень).* Определяется по коэффициенту корреляционной связи.

**2.1.3) Вариационный ряд и ранговая корреляция.**

Пусть из генеральной совокупности извлечена выборка, причем наблюдалось раз, — раз, — раз. — объем выборки. Наблюдаемые значения - называют вариантами, а последовательность

вариант, записанных в возрастающем порядке, — вариационным рядом. Числа наблюдений называют частотами, а их отношения к объему выборки— относительными частотами.

*Статистическим распределением выборки* называют перечень вариант и соответствующих им частот или относительных частот. Статистическое распределение можно задать также в виде последовательности интервалов и соответствующих им частот (в качестве частоты, соответствующей

интервалу, принимают сумму частот, попавших в этот интервал).

Заметим, что в теории вероятностей под *распределением*

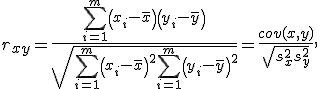
понимают соответствие между возможными значениями

случайной величины и их вероятностями, а в математической

статистике — соответствие между наблюдаемыми

вариантами и их частотами, или относительными частотами.

**2.1.4) Коэффициент корреляция Пирсона.**

Коэффициент корреляции Пирсона характеризует существование линейной зависимости между двумя величинами. Пусть даны две выборки x^m=\left( x_1, \cdots ,x_m  \right), \; y^m=\left( y_1, \cdots ,y_m  \right); коэффициент корреляции Пирсона рассчитывается по формуле: 

где \bar{x}, \bar{y} – выборочные средние x^m и y^m, s_x^2,  s_y^2 – выборочные дисперсии, r_{xy} \in \left[-1,1\right]. Коэффициент корреляции Пирсона называют также теснотой линейной связи. Свойства выборочного коэффициента корреляции:

* Знак выборочного коэффициента корреляции *ryx* совпадает со знаком выборочного коэффициента регрессии *byx*.
* Корреляция называется положительной, если переменные *Х*и *Y*изменяются в одном направлении, и отрицательной − если переменные *Х* и *Y* изменяются в разных направлениях.
* Абсолютная величина выборочного коэффициента корреляции *ryx* не больше единицы:|*ryx*|*≤*1.
* Если выборочный коэффициент корреляции равен 0, то переменные *Х* и *Y* не связаны линейной корреляционной зависимостью, т.е. при *ryx* *=*0 условные средние *yi* сохраняют постоянное значение при всех значениях *хi*.
* Если выборочный коэффициент корреляции равен 0, то переменные *Х* и *Y*могут быть связаны нелинейной корреляционной или даже функциональной зависимостью.
* Если абсолютная величина выборочного коэффициента корреляции равна 1, то наблюдаемые значения переменных *Х* и *Y* связаны линейной функциональной зависимостью.
* С возрастанием абсолютной величины выборочного коэффициента корреляции линейная корреляционная зависимость переменных *Х* и *Y* становится более тесной и при |*ryx*|= 1 переходит в функциональную зависимость.

**2.1.6) Коэффициент частной корреляции.**

Если удалось установить тесную зависимость между двумя исследуемыми величинами, отсюда ещё непосредственно не следует их причинная взаимообусловленность. Из причинной связи величин следует стохастическая связь, из стохастической связи не всегда следует причинная.

За счет эффектов одновременного влияния неучтенных факторов на исследуемые переменные может искажаться смысл истинной связи между переменными. Например, подсчеты приводят к положительному значению коэффициента корреляции между парой случайных величин, в то время как истинная связь между ними имеет отрицательный смысл. Такую корреляцию между двумя переменными часто называют «ложной». Более детально подобные ситуации — обнаружение и исключение «общих причинных факторов», расчет «очищенных» или частных коэффициентов корреляции — исследуют методами многомерного корреляционного анализа.

Исключить влияние третьей переменной позволяет частный коэффициент корреляции. Частным коэффициентом корреляции между случайными величинами X и Y при исключении влияния случайной величины Z называется

r_{XY|Z}=\frac{r_{XY}-r_{XZ}r_{YZ}}{\sqrt{(1-r_{XZ}^2)(1-r_{YZ}^2)}},

где r_{XY} — коэффициент корреляции Пирсона между случайными величинами X и Y.  
Ранговый [коэффициент корреляции Кенделла](http://www.machinelearning.ru/wiki/index.php?title=%D0%9A%D0%BE%D1%8D%D1%84%D1%84%D0%B8%D1%86%D0%B8%D0%B5%D0%BD%D1%82_%D0%BA%D0%BE%D1%80%D1%80%D0%B5%D0%BB%D1%8F%D1%86%D0%B8%D0%B8_%D0%9A%D0%B5%D0%BD%D0%B4%D0%B5%D0%BB%D0%BB%D0%B0) \tau (в отличие от [коэффициента Спирмена](http://www.machinelearning.ru/wiki/index.php?title=%D0%9A%D0%BE%D1%8D%D1%84%D1%84%D0%B8%D1%86%D0%B8%D0%B5%D0%BD%D1%82_%D0%BA%D0%BE%D1%80%D1%80%D0%B5%D0%BB%D1%8F%D1%86%D0%B8%D0%B8_%D0%A1%D0%BF%D0%B8%D1%80%D0%BC%D0%B5%D0%BD%D0%B0) \rho) переносится на случай частной корреляции с помощью аналогичной формулы:

<tex>\tau_{XY|Z}=\frac{\tau_{XY}-\tau_{XZ}\tau_{YZ}}{\sqrt{(1-\tau_{XZ}^2)(1-\tau_{YZ}^2)}},

Где \tau_{XY} — [коэффициент корреляции Кенделла](http://www.machinelearning.ru/wiki/index.php?title=%D0%9A%D0%BE%D1%8D%D1%84%D1%84%D0%B8%D1%86%D0%B8%D0%B5%D0%BD%D1%82_%D0%BA%D0%BE%D1%80%D1%80%D0%B5%D0%BB%D1%8F%D1%86%D0%B8%D0%B8_%D0%9A%D0%B5%D0%BD%D0%B4%D0%B5%D0%BB%D0%BB%D0%B0) между случайными величинами X и Y.

## 2.2) Роль в машинном обучении

**2.3) Реализация для прикладной задачи**

## 2.4) Плюсы и минусы метода

## Слабые стороны корреляционного анализа:

## 

**Метод максимального правдоподобия**

**3.1) Математическая теория**

**3.2) Роль в машинном обучении**

**3.3) Реализация для прикладной задачи**

**3.4) Плюсы и минусы метода**

**Сравнение методов**