МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«ТУЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

Институт *Прикладной математики и компьютерных наук*

Кафедра *Информационной безопасности*

СБОРНИК МЕТОДИЧЕСКИХ УКАЗАНИЙ К ЛАБОРАТОРНЫМ РАБОТАМ

по дисциплине

МЕТОДЫ АНАЛИЗА ДАННЫХ

Направление подготовки: 09.04.01 *«Информатика и вычислительная техника»*
Профиль : *«Компьютерный анализ и интерпретация данных»*

Квалификация (степень) выпускника: *магистр*

Формы обучения: *очная*

Тула 2015 г.

Методические указания к лабораторным работам учебной дисциплины *(модуля) «Методы анализа данных»* разработаны *проф. С.Д. Двоенко* и обсуждены на заседании кафедры Информационной безопасности института Прикладной математики и компьютерных наук (протокол заседания кафедры №____ от «____» _____ 20___ г.)

Разработчик(и) МУ ЛР дисциплины *(модуля)* _____.

личная подпись(и)

Лабораторная работа №1

СТАНДАРТИЗАЦИЯ И ПРЕОБРАЗОВАНИЕ ДАННЫХ

Цель и задача работы

Приведение экспериментальных данных к стандартизованному виду. Преобразования матрицы данных.

Теоретические положения

МАТРИЦА ДАННЫХ

Рассмотрим традиционный вид представления результатов эксперимента - матрицу данных. Пусть исследователь располагает совокупностью из N наблюдений над состоянием исследуемого явления. Пусть при этом явление описано набором из n характеристик, значения которых тем или иным способом измерены в ходе эксперимента. Данные характеристики носят название признаков, показателей или параметров. Такая информация представляется в виде двухмерной таблицы чисел $\mathbf{X}$ размерности $N \times n$ или в виде матрицы $\mathbf{X}(N \times n)$:

$$\begin{array}{c} \\ \mathbf{x}_1 \\ \cdot \\ \mathbf{x}_i \\ \cdot \\ \mathbf{x}_N \end{array} \begin{array}{c} \begin{array}{ccccc} X_1 & \dots & X_j & \dots & X_n \end{array} \\ \begin{pmatrix} x_{11} & \dots & x_{1j} & \dots & x_{1n} \\ \cdot & \dots & \cdot & \dots & \cdot \\ x_{i1} & \dots & x_{ij} & \dots & x_{in} \\ \cdot & \dots & \cdot & \dots & \cdot \\ x_{N1} & \dots & x_{Nj} & \dots & x_{Nn} \end{pmatrix} \end{array}.$$

Строки матрицы $\mathbf{X}$ соответствуют наблюдениям или, другими словами, объектам наблюдения. В качестве объектов наблюдения выступают, например, в социологии - респонденты (анкетируемые люди), в экономике - предприятия, виды продукции и т.д. Столбцы матрицы $\mathbf{X}$ соответствуют признакам, характеризующим изучаемое явление. Как правило, это наиболее легко измеряемые характеристики объектов. Например, предприятие характеризуется численностью, стоимостью основных фондов, видом выпускаемой продукции и т.д. Очевидно, что элемент x_{ij} представляет собой значение признака j, измеренное на объекте i.

Часто матрица данных приводится к стандартной форме преобразованием

$$x'_{ij} = \left(x_{ij} - \bar{x}_j\right) \Big/ \sigma_j, \quad \bar{x}_j = \frac{1}{N}\sum_{i=1}^{N} x_{ij}, \quad \sigma_j^2 = \frac{1}{N}\sum_{i=1}^{N}\left(x_{ij} - \bar{x}_j\right)^2, \quad i = 1,\dots N;\ j = 1,\dots n,$$

где $\bar{x}_j, \sigma_j^2 -$ среднее и дисперсия по столбцу j, после которого стандартная матрица $\mathbf{X}'$ обладает свойствами

$$\bar{x}'_j = \frac{1}{N}\sum_{i=1}^{N} x'_{ij} = 0, \qquad \sigma'^2_j = \frac{1}{N}\sum_{i=1}^{N} x'^{\,2}_{ij} = 1, \quad i = 1,\dots N;\ j = 1,\dots n.$$

В дальнейшем будем использовать для матрицы данных обозначение $\mathbf{X}$, полагая, что это стандартизованная матрица, без дополнительного упоминания. Для пояснения заметим, что часто признаки, описывающие некоторый объект, имеют существенно различный физический смысл. Это приводит к тому, что величины в различных столбцах исходной матрицы трудно сопоставлять между собой, например, *кг* и *м*. Поэтому

получение стандартизованной матрицы можно понимать как приведение всех признаков к некоторой единой условной физической величине, измеренной в одних и тех же условных единицах.

ГИПОТЕЗЫ КОМПАКТНОСТИ И СКРЫТЫХ ФАКТОРОВ

Рассмотрим n-мерное пространство, где оси координат соответствуют отдельным признакам матрицы данных **X**. Тогда каждую строку матрицы данных можно представить как вектор в этом пространстве. Следовательно, каждый из N объектов наблюдения представлен своей изображающей точкой в n-мерном пространстве признаков (Рис. 1.1).

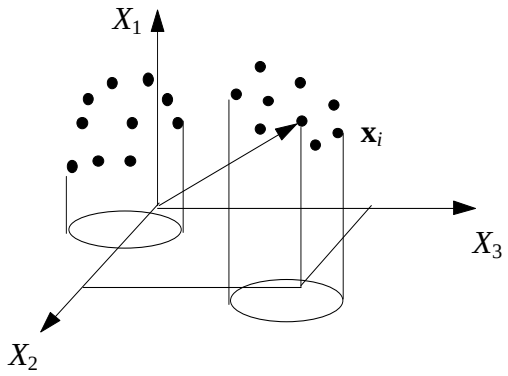

Рис. 1.1. Пространство признаков.

Отметим, что в основе различных методов анализа матрицы данных лежит неформальное предположение, условно названное “гипотезой компактности”. Предполагается, что объекты наблюдения в различной степени “похожи” друг на друга. Предполагается, что все множество большого числа объектов представимо в виде небольшого числа достаточно сильно различающихся подмножеств, внутри которых объекты наблюдения “сильно похожи”. Например, сильно различающиеся подмножества характеризуют типы различных состояний изучаемого явления, а похожие объекты внутри них являются зафиксированными состояниями явления, где разброс значений объясняется ошибками измерения, изменением условий эксперимента и т.д.

Такие компактные множества называются классами, кластерами, таксонами. При справедливости такой гипотезы задача обработки в наиболее общей формулировке неформально ставится как задача разбиения исходного множества объектов в признаковом пространстве на конечное число классов. Не вдаваясь глубоко в суть различных постановок задачи классификации, отметим следующие важные моменты.

Во-первых, при известном числе классов, как правило, требуется получить наиболее удаленные друг от друга в пространстве признаков компактные классы.

Во-вторых, часто число классов заранее неизвестно, поэтому нужно его определить, исходя из априорных соображений, или, пробуя разные варианты разбиения на классы.

В-третьих, важно, чтобы результат разбиения был устойчивым. Например, методы, используемые в одном из направлений обработки данных - кластер-анализе - могут порождать различные разбиения для небольших изменений матрицы данных. Так, если в исходную матрицу добавить новые объекты, то результат кластеризации изменится. Если он изменится незначительно по составу кластеров, удаленности кластеров друг от друга, их размеру в пространстве, то результат можно считать устойчивым.

В-четвертых, другие методы классификации, например, в распознавании образов, направлены не на получение таксономии (перечисление принадлежности объектов каждому из классов), а на получение способа определять класс каждого добавляемого к

матрице данных объекта. Данный метод реализуется в виде так называемого решающего правила. Оно представляет собой функцию $g(\mathbf{x})$, принимающую значения на конечном множестве из m классов $\{\Omega_1,\dots\Omega_m\}$. Тогда при предъявлении объекта $\mathbf{x}\in\Omega_i$, решающая функция примет значение $g(\mathbf{x})=\Omega_i$.

Заметим, что разбиение объектов наблюдения на классы означает разделение матрицы данных на горизонтальные полосы, т.е. перегруппировку строк матрицы так, что внутри каждой из групп строк объекты принадлежат одному классу и не принадлежат другим классам.

С другой стороны, можно рассмотреть *N*-мерное пространство, оси которого соответствуют отдельным объектам. Тогда каждый столбец X_j матрицы **X** представляет собой вектор в данном пространстве, а вся матрица - совокупность *n* векторов (Рис. 1.2).

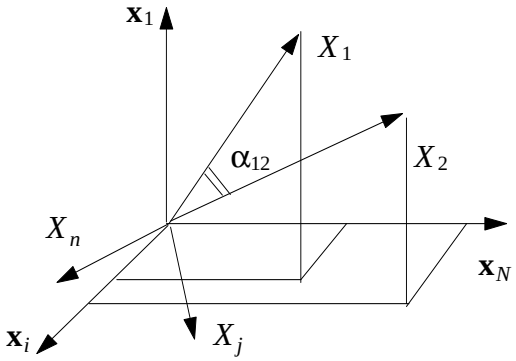

Рис. 1.2. Пространство объектов.

Такое пространство называется пространством объектов. В нем все векторы X_j одинаковы по длине, вычисляемой как евклидова норма

$$\|X_j\|=\sqrt{\sum_{i=1}^{N}x_{ij}^2}=\sqrt{N\sigma_j^2}=\sqrt{N}\ .$$

Тогда характеристикой близости признаков X_i и X_j в таком пространстве служит близость направлений их векторов, измеряемая $\cos\alpha_{ij}$, где α_{ij} - угол между ними. В этом смысле векторы близки, если угол между ними близок к нулю или к 180^0, и, следовательно, косинус угла близок по модулю к единице. Равенство $\cos\alpha_{ij}$ по модулю единице означает совпадение векторов и линейную связь, так как в стандартизованной матрице данных значения по одному признаку в точности соответствует значениям по другому признаку, или совпадение векторов с точностью до наоборот, то есть противоположные направления, и, следовательно, также линейную связь. Тогда перпендикулярные векторы и нулевое значение косинуса угла между ними соответствуют наиболее далеким признакам. В этом случае можно предположить противоположную ситуацию, когда признаки наименее зависимы друг от друга - линейно независимы.

Из теории вероятностей и математической статистики известно, что линейная связь между двумя переменными характеризуется коэффициентом корреляции. Случаю двух переменных, где значения каждой из них представлены в виде ряда наблюдений, соответствует выбор двух столбцов и $X_j=\left(x_{1j},\dots x_{Nj}\right)^T$ в матрице данных. Коэффициент корреляции есть просто скалярное произведение двух векторов признаков в пространстве объектов, нормированное к их длине, то есть просто косинус угла между стандартизованными векторами:

$$r_{ij} = \frac{1}{N}\sum_{k=1}^{N} x_{ki}x_{kj} = \frac{1}{N} X_i^T X_j = \frac{1}{N}\|X_i\|\cdot\|X_j\|\cos\alpha_{ij} = \cos\alpha_{ij}.$$

В статистическом смысле корреляционная связь означает, что значения одного признака имеют тенденцию изменяться синхронно значениям другого признака. Отсутствие связи означает, что изменение значений одного признака никак не сказывается на изменении значений другого признака. Такие признаки считаются статистически независимыми и, в частности, при отсутствии корреляционной связи, линейно независимыми.

Отметим, что в основе понятия о взаимосвязи между признаками лежит неформальное предположение, условно названное “гипотезой скрытых факторов”. А именно, предполагается, что состояние некоторого изучаемого явления определяется “скрытым”, “существенным” фактором, который нельзя измерить непосредственно. Можно лишь измерить набор некоторых других признаков, косвенно отражающих состояние скрытого фактора. Предполагается также, что множество скрытых факторов невелико и значительно меньше набора измеряемых признаков. Тогда группа признаков, испытывающая преимущественное влияние некоторого из факторов, будет более или менее синхронно изменять свои значения при изменении состояния этого скрытого фактора. Чем сильнее влияние скрытого фактора, тем синхроннее меняют свои значения признаки, тем сильнее связь.

В пространстве объектов это означает, что векторы признаков образуют достаточно компактную группу, в которой пучок направлений векторов можно охватить некоторым выпуклым конусом с острой вершиной в начале координат.

При справедливости гипотезы о факторах задача обработки в наиболее общей формулировке неформально ставится как задача выделения конечного числа групп наиболее сильно связанных между собой признаков и построения для каждой из них (либо выбора среди них) одного, наиболее сильно связанного с ними (наиболее близкого к ним) признака, который считается фактором данной группы. Успешное решение такой задачи означает, что в основе сложных взаимосвязей между внешними признаками лежит относительно более простая скрытая структура, отражающая наиболее характерные и часто повторяющиеся взаимосвязи.

Отметим следующие важные моменты.

Во-первых, различные методы выделения скрытых факторов объединены в группу методов - факторный анализ. Сюда же многие исследователи относят и метод главных компонент.

Во-вторых, существенным в этих методах является то, что число найденных факторов k должно быть много меньше числа признаков n, а найденные факторы должны быть как можно более ортогональны друг другу.

В-третьих, как правило, система факторов должна быть ориентирована так, чтобы факторы были упорядочены по масштабу разброса значений объектов на их осях. В статистических терминах это означает, что факторы должны быть упорядочены по дисперсии объектов на их осях. Необходимость получения именно такой конфигурации объясняется следующим обстоятельством. Возьмем в пространстве факторов главный фактор - фактор с наибольшей дисперсией объектов по его оси. Очевидно, что чем больше дисперсия значений объектов по его оси, тем легче выделить локальные сгущения значений и интерпретировать их как группы похожих объектов, то есть классифицировать их. Такое же предположение применимо и к оставшимся факторам. Если система факторов ортогональна или близка к ней, то факторы считаются независимыми. Тогда разброс значений по оси каждого из факторов можно объяснить влиянием только этого фактора.

Пусть, например, ряды наблюдений двух случайных величин $X_i = \left(x_{1i},\ldots x_{Ni}\right)^T$ и

$X_j = \left(x_{1j}, \ldots x_{Nj}\right)^T$ являются выборками из генеральной совокупности с нормальным законом распределения. Изобразим пространство двух признаков в виде плоскости с осями координат X_i и X_j (Рис. 1.3).

Плотность вероятности нормального распределения по оси каждого признака есть $f(x) = \left(1/\sigma\sqrt{2\pi}\right)\exp\left[-\left(x-\bar{x}\right)^2/2\sigma^2\right] = \left(1/\sqrt{2\pi}\right)\exp\left(-x^2/2\right)$ при $\bar{x} = 0, \sigma = 1$.

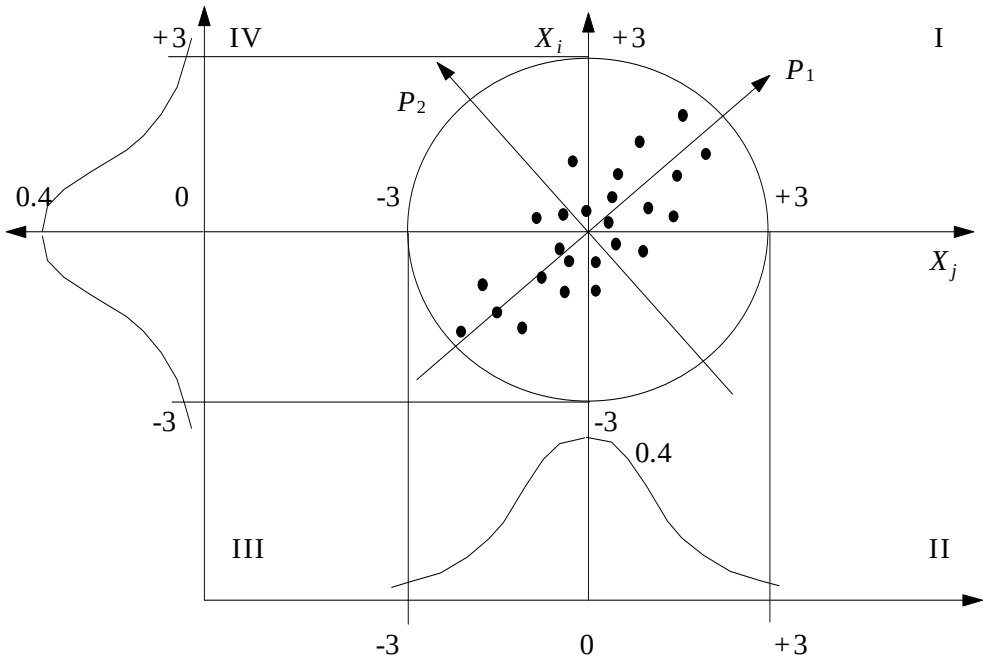

Рис. 1.3. Распределение наблюдений на плоскости.

Согласно хорошо известному правилу “трех сигм”, 99.73% наблюдений нормально распределенной случайной величины попадет в интервал значений по оси аргумента от -3σ до $+3\sigma$, или при $\sigma = 1$ от -3 до +3. Следовательно, на плоскости в координатах X_i и X_j все 99.73% наблюдений будут сосредоточены внутри окружности радиуса 3. При наличии корреляционной связи между признаками наблюдения будут сосредоточены внутри эллипса рассеивания. Чем сильнее окажется связь, тем уже будет эллипс рассеивания. В случае положительной связи, изображенной на рисунке, большие значения одного признака имеют тенденцию соответствовать большим значениям другого признака и наоборот. Поэтому, в большинстве случаев совместные наблюдения значений этих признаков более часто попадают в I и III квадранты плоскости и реже - во II и IV. Кривые равных вероятностей имеют форму вложенных эллипсов с двумя осями P_1 и P_2. Из рисунка легко заметить, что проекции изображающих точек на горизонтальную ось X_j в среднем расположены более плотно, чем проекции тех же точек на ось P_1. Математически доказан факт, что проекции точек на главную ось P_1 эллипса рассеивания расположены наименее плотно по сравнению с другими возможными положениями оси. Если кластеры представляют собой локальные сгущения в эллипсе рассеивания, то переход к системе координат P_1 и P_2 дает наилучшую возможность для их выделения. При достаточно сильной корреляции исходных признаков новый признак P_1 может быть выбран в качестве их фактора.

Заметим, что разбиение признаков на группы означает разбиение матрицы данных на вертикальные полосы, то есть перегруппировку столбцов матрицы так, что внутри одной группы признаки сильно связаны между собой и слабо связаны с любым признаком из другой группы.

МАТРИЦА ОБЪЕКТ-ОБЪЕКТ И ПРИЗНАК-ПРИЗНАК. РАССТОЯНИЕ И БЛИЗОСТЬ

Пусть имеется матрица данных $\mathbf{X}(N \times n)$. Если рассматривать строки данной матрицы как N векторов $\mathbf{x}_i$ в пространстве n признаков, то естественно рассмотреть расстояние между двумя некоторыми векторами. Расстояния между всевозможными парами векторов дают матрицу $\mathbf{R}(N \times N)$ расстояний типа объект - объект.

Напомним, что расстоянием между векторами в пространстве признаков называется некоторая положительная величина d, удовлетворяющая следующим трем аксиомам метрики:

1. $d(\mathbf{x}_1, \mathbf{x}_2) > 0, \quad d(\mathbf{x}_1, \mathbf{x}_1) = 0;$
2. $d(\mathbf{x}_1, \mathbf{x}_2) = d(\mathbf{x}_2, \mathbf{x}_1);$
3. $d(\mathbf{x}_1, \mathbf{x}_2) + d(\mathbf{x}_2, \mathbf{x}_3) \geq d(\mathbf{x}_1, \mathbf{x}_3)$ (неравенство треугольника).

Таким образом, матрица расстояний является симметричной с нулевой главной диагональю. Существуют различные метрики, но наиболее известной вообще и наиболее применяемой в обработке данных, в частности, является евклидова метрика

$$d(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2}.$$

Часто используется линейная метрика вида

$$d(\mathbf{x}_1, \mathbf{x}_2) = \sum_{i=1}^{n} |x_{1i} - x_{2i}|.$$

Применение линейной метрики оправдано, когда расстояние определяется как расстояние между домами в городе по кварталам, а не напрямик. Возможны и другие виды расстояний.

Часто рассматривается величина, обратная в некотором смысле расстоянию - близость. На практике часто используют функции близости вида

$$\mu(\mathbf{x}_1, \mathbf{x}_2) = \exp\left[-\alpha d^2(\mathbf{x}_1, \mathbf{x}_2)\right] \text{ или } \mu(\mathbf{x}_1, \mathbf{x}_2) = \frac{1}{1 + \alpha d(\mathbf{x}_1, \mathbf{x}_2)},$$

где α определяет крутизну функции близости. Очевидно, что матрица близостей также является симметричной с единичной главной диагональю, так как $\mu(\mathbf{x}_1, \mathbf{x}_1) = 1.$

Если рассмотреть признаки как n векторов в N-мерном пространстве объектов, то получим другое преобразование матрицы данных в матрицу $\mathbf{R}(n \times n)$ типа признак - признак. Элементом r_{ij} такой матрицы является значение расстояния или близости между признаками X_i и X_j. Наиболее распространено представление в виде матрицы близостей между признаками, где под близостью понимается, например, корреляция соответствующих признаков.

Легко заметить, что содержательные задачи на матрице данных $\mathbf{X}(N \times n)$ интерпретируются на квадратных матрицах $\mathbf{R}(N \times N)$ и $\mathbf{R}(n \times n)$ как выделение блочно - диагональной структуры путем одновременной перегруппировки строк и столбцов. Тогда в каждом диагональном блоке группируются элементы, близкие в соответствующем пространстве и далекие от элементов других блоков. Такая задача группировки известна как задача диагонализации матрицы связей (Рис. 1.8). Задача о диагонализации матрицы связей является наиболее общей для матриц связей произвольной природы. Особенно интересным является случай, когда матрица связей является корреляционной матри-

цей. Именно для этого случая разработаны и широко применяются на практике специальные алгоритмы, известные как алгоритмы экстремальной группировки признаков (параметров).

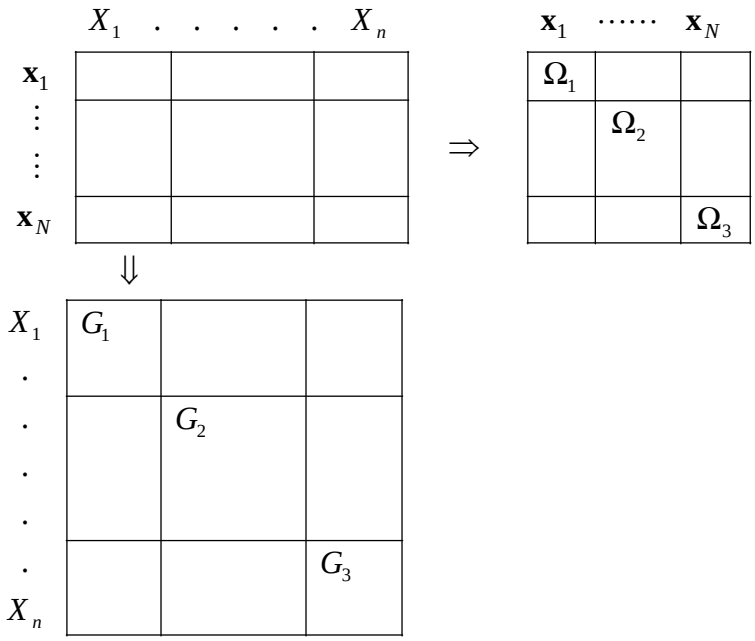

Рис. 1.8. Диагонализация матрицы связей.

Задание на работу

1. Выбрать матрицу данных в одном из публичных репозиториев данных:

- http://polygon.machinelearning.ru - репозиторий данных и алгоритмов «Полигон» ВЦ РАН

- http://archive.ics.uci.edu/ml/– репозиторий данных Центра машинного обучения и интеллектуальных систем (университет Калифорнии, Ирвайн)

2. Рассмотреть содержательную задачу обработки выбранных данных, изучить описание данных

3. Составить матрицу количественных данных вида объект-признак
4. Привести матрицу данных к стандартизированному виду

Содержание отчета

Номер и название лабораторной работы;
Цель лабораторной работы;
Доклад к презентации.
Выводы.

Контрольные вопросы

1. Как вычислить коэффициент корреляции
2. Что характеризует коэффициент корреляции
3. Для чего выполняется стандартизация данных
4. В чем заключаются свойства расстояния

Лабораторная работа №2

ПОСТРОЕНИЕ ГЛАВНЫХ КОМПОНЕНТ

Цель и задача работы

Изучение основных методов поиска собственных векторов и собственных чисел корреляционной матрицы

Теоретические положения

КОРРЕЛЯЦИОННАЯ МАТРИЦА И ЕЕ СВОЙСТВА

При анализе связей важное значение имеет структура взаимосвязей между признаками. Как известно, измерителем линейной связи между признаками служит коэффициент корреляции или, в более общем случае, коэффициент ковариации. С другой стороны, вектор средних и матрица ковариаций являются исчерпывающими характеристиками нормального закона распределения. Поэтому остановимся более подробно на свойствах корреляционной матрицы.

Корреляционная матрица $\mathbf{R}(n\times n)$ является симметричной, с единичной главной диагональю, положительно полуопределенной матрицей. Напомним из линейной алгебры, что квадратная матрица, не обязательно симметричная, называется положительно полуопределенной, если для любого вектора $\mathbf{y}=(y_1,\dots y_n)^T$ квадратичная форма $\mathbf{y}^T\mathbf{R}\mathbf{y}\ge 0$ не отрицательна. Квадратная матрица **R** положительно определена, если для любых **y** квадратичная форма $\mathbf{y}^T\mathbf{R}\mathbf{y}>0$ строго положительна. В данном свойстве матрицы **R** легко убедиться:

$$\mathbf{y}^T\mathbf{R}\mathbf{y}=\sum_{i=1}^{n}y_i\left(\sum_{j=1}^{n}r_{ij}y_j\right)=\sum_{i=1}^{n}\sum_{j=1}^{n}r_{ij}y_iy_j=\frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{N}x_{ki}x_{kj}y_iy_j=$$

$$\frac{1}{N}\sum_{k=1}^{N}\sum_{i=1}^{n}x_{ki}y_i\sum_{j=1}^{n}x_{kj}y_j=\frac{1}{N}\sum_{k=1}^{N}\left(\sum_{i=1}^{n}x_{ki}y_i\right)^2\ge 0,\ \text{где}\ \frac{1}{N}\sum_{k=1}^{N}x_{ki}x_{kj}=\frac{1}{N}X_i^TX_j=r_{ij},$$

r_{ij} -коэффициент корреляции, вычисленный как скалярное произведение признаков X_i и X_j в стандартной матрице **X**.

Заметим, что при ненулевом векторе **y** квадратичная форма $\mathbf{y}^T\mathbf{R}\mathbf{y}$ может обратиться в нуль, только если признаки $X_i=(x_{1i},\dots x_{Ni})^T$, $i=1,\dots n$ линейно зависимы между собой.

Действительно, пусть все признаки X_i линейно зависимы между собой. Тогда матрица **R**=(r_{ij} = 1), i =1,... n, j = 1,... n состоит из единиц, если линейная связь, например, положительна. Тогда для некоторого вектора **y** получим

$$\mathbf{y}^T\mathbf{R}\mathbf{y}=(y_1,\dots y_n)\begin{pmatrix}1&\cdots&1\\ \vdots&\vdots&\vdots\\ 1&\cdots&1\end{pmatrix}\begin{pmatrix}y_1\\ \vdots\\ y_n\end{pmatrix}=\left(\sum_{i=1}^{n}y_i,\dots\sum_{i=1}^{n}y_i\right)\begin{pmatrix}y_1\\ \vdots\\ y_n\end{pmatrix}=\sum_{i=1}^{n}y_i\sum_{j=1}^{n}y_j=\sum_{i=1}^{n}\sum_{j=1}^{n}y_iy_j=0$$

Очевидно, что данное число представляет собой сумму всевозможных комбинаций попарных произведений координат вектора **y**. Все попарные произведения координат данного вектора можно представить в виде квадратной матрицы размером $n\times n$:

$$\begin{pmatrix} y_1^2 & y_1y_2 & \cdots & y_1y_n \\ y_2y_1 & y_2^2 & \cdots & y_2y_n \\ \vdots & \vdots & \vdots & \vdots \\ y_ny_1 & y_ny_2 & \cdots & y_n^2 \end{pmatrix} = \mathbf{y}\mathbf{y}^T.$$

Матрица $\mathbf{y}\mathbf{y}^T$ является симметричной, а сумма ее диагональных элементов представляет собой квадрат длины вектора **y** и всегда положительна для ненулевого **y**. Следовательно, равенство $\mathbf{y}^T\mathbf{R}\mathbf{y} = 0$ выполняется только, когда сумма диагональных элементов равна по модулю и противоположна по знаку сумме недиагональных элементов $\sum_{i=1}^{n} y_i^2 + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} y_i y_j = 0$.

Для случая $n = 2$ получим: $y_1^2 + y_2^2 + 2y_1y_2 = 0$. Решив данное квадратное уравнение относительно y_1, получим, что $\mathbf{y}^T\mathbf{R}\mathbf{y} = 0$ при $y_1 = -y_2$.

Признаки X_i представляют собой результаты измерений, где часто число объектов N много больше числа признаков n. Поэтому, в силу возможных ошибок и неточностей измерений, не говоря уже о случайных помехах, линейная зависимость признаков X_i маловероятна. Поэтому, как правило, данная квадратичная форма оказывается строго положительной при любом ненулевом векторе **y**.

СОБСТВЕННЫЕ ВЕКТОРЫ ЧИСЛА КОРРЕЛЯЦИОННОЙ МАТРИЦЫ

Собственным вектором корреляционной матрицы **R**, соответствующим собственному числу λ, называется ненулевой вектор $\mathbf{x} = (x_1, \ldots x_n)^T$, удовлетворяющий уравнению $\mathbf{R}\mathbf{x} = \lambda\,\mathbf{x}$.

Как известно из линейной алгебры, матрица **R** рассматривается в данном случае как матрица линейного преобразования вектора **x** в вектор $\lambda\mathbf{x}$. Это означает, что для данного линейного преобразования **R** в n-мерном пространстве существует такое направление, что преобразование **R** только растягивает вектор **x** в λ раз, сохраняя его ориентацию.

Векторное уравнение можно переписать в виде однородного уравнения относительно **x**: $(\mathbf{R} - \lambda\ \mathbf{E})\mathbf{x} = 0$. Данное уравнение имеет ненулевое (нетривиальное) решение только тогда, когда определитель $\det(\mathbf{R} - \lambda\mathbf{E})$ равен нулю. Данный определитель представляет собой уравнение относительно λ и является полиномом n степени вида $(-1)^n\lambda^n + (-1)^{n-1}p_1\lambda^{n-1} + \ldots + p_n = 0$.

Данный полином называется характеристическим полиномом (многочленом), а уравнение $\det(\mathbf{R} - \lambda\,\mathbf{E}) = 0$ - характеристическим уравнением. Характеристическое уравнение имеет n, вообще говоря, различных корней. При этом его корни λ_i являются собственными числами преобразования **R**. В качестве собственных векторов $\mathbf{x}_i, i = 1, \ldots n$ линейного преобразования **R**, соответствующих собственным числам $\lambda_i, i = 1, \ldots n$, берутся векторы единичной длины $\sum_{j=1}^{n} x_{ij}^2 = 1;\ i = 1, \ldots n$, каждый из которых удовлетворяет соответствующему характеристическому уравнению $\det(\mathbf{R} - \lambda_i\mathbf{E}) = 0$. Рассмотрим случай n=2. Тогда получим

$$\mathbf{R} = \begin{pmatrix} 1 & r_{12} \\ r_{21} & 1 \end{pmatrix} = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix};\quad \mathbf{R} - \lambda\mathbf{E} = \begin{pmatrix} \mathbf{1} - \lambda & r \\ r & 1 - \lambda \end{pmatrix};\quad \det(\mathbf{R} - \lambda\mathbf{E}) = (1 - \lambda)^2 - r^2 = 0.$$

Решением квадратного уравнения $\lambda^2 - 2\lambda + 1 - r^2 = 0$ относительно λ являются корни

$\lambda_1 = 1 + r$ и $\lambda_2 = 1 - r$.

Отметим следующие свойства собственных чисел.

1) $\lambda_1 > \lambda_2 > 0$. Так как корреляционная матрица **R** практически положительно определена, то при произвольном *n* все ее собственные числа являются действительными и строго положительными $\lambda_1 > \lambda_2 > \ldots > \lambda_n > 0$.

2) $\lambda_1 + \lambda_2 = 2$. Вычислим след матрицы **R** как сумму ее диагональных элементов $\mathrm{tr}\mathbf{R} = r_{11} + r_{22} = 1 + 1 = 2$. Следовательно, $\mathrm{tr}\mathbf{R} = \lambda_1 + \lambda_2$, то есть сумма собственных чисел корреляционной матрицы равна ее следу. При произвольном *n* получим $\sum_{i=1}^{n} \lambda_i = \mathrm{tr}\mathbf{R}$.

3) $\lambda_1 \lambda_2 = 1 - r^2$. Определитель корреляционной матрицы равен $\det \mathbf{R} = 1 - r^2$. Следовательно, $\det \mathbf{R} = \lambda_1 \lambda_2$. При произвольном *n* получим $\prod_{i=1}^{n} \lambda_i = (-1)^n \det \mathbf{R} = \det \mathbf{R}$. Следовательно, произведение собственных чисел равно определителю корреляционной матрицы, взятому со знаком плюс, так как все собственные числа положительны.

Найдем собственные векторы $\mathbf{x}_1$ и $\mathbf{x}_2$, соответствующие собственным числам λ_1 и λ_2. Из характеристического уравнения следует, что первый вектор найдется из уравнения

$$\begin{pmatrix} 1-\lambda_1 & r \\ r & 1-\lambda_1 \end{pmatrix} \begin{pmatrix} x_{11} \\ x_{12} \end{pmatrix} = \begin{pmatrix} -r & r \\ r & -r \end{pmatrix} \begin{pmatrix} x_{11} \\ x_{12} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Согласно определению $x_{11}^2 + x_{12}^2 = 1$. Тогда получим систему уравнений

$$\begin{cases} -rx_{11} + rx_{12} = 0 \\ rx_{11} - rx_{12} = 0 \\ x_{11}^2 + x_{12}^2 = 1. \end{cases}$$

Из решения данной системы следует, что $x_{11} = x_{12} = \pm\sqrt{2}/2 = \pm 0.707$. Два решения указывают на противоположные направления вдоль диагонали первого и третьего квадрантов плоскости координат:

$$\begin{cases} x_{11} = 0.707 \\ x_{12} = 0.707 \end{cases} \qquad \begin{cases} x_{11} = -0.707 \\ x_{12} = -0.707 \end{cases}.$$

Второй вектор найдется из уравнения

$$\begin{pmatrix} 1-\lambda_2 & r \\ r & 1-\lambda_2 \end{pmatrix} \begin{pmatrix} x_{21} \\ x_{22} \end{pmatrix} = \begin{pmatrix} r & r \\ r & r \end{pmatrix} \begin{pmatrix} x_{21} \\ x_{22} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

В результате получим два решения, указывающие на противоположные направления вдоль диагонали второго и четвертого квадрантов плоскости координат

$$\begin{cases} x_{21} = 0.707 \\ x_{22} = -0.707 \end{cases} \begin{cases} x_{21} = -0.707 \\ x_{22} = 0.707 \end{cases}.$$

Как сразу нетрудно заметить, собственные векторы матрицы **R**, то есть вещественной симметричной матрицы, соответствующие различным собственным числам, ортогональны между собой. Покажем это для произвольного *n*. Рассмотрим уравнения

$\mathbf{R}\mathbf{x}_1 = \lambda_1\mathbf{x}_1$ и $\mathbf{R}\mathbf{x}_2 = \lambda_2\mathbf{x}_2$, где $\lambda_1 \neq \lambda_2$. Домножим каждое из уравнений на собственный вектор другого уравнения и получим $\mathbf{x}_2^T\mathbf{R}\mathbf{x}_1 = \lambda_1\mathbf{x}_2^T\mathbf{x}_1$ и $\mathbf{x}_1^T\mathbf{R}\mathbf{x}_2 = \lambda_2\mathbf{x}_1^T\mathbf{x}_2$. Так как

$$\mathbf{x}_2^T\mathbf{R}\mathbf{x}_1 = \mathbf{x}_1^T(\mathbf{x}_2^T\mathbf{R})^T = \mathbf{x}_1^T\mathbf{R}^T\mathbf{x}_2 = \mathbf{x}_1^T\mathbf{R}\mathbf{x}_2,$$

то, вычтя одно уравнение из другого, получим

$$0 = \lambda_1\mathbf{x}_2^T\mathbf{x}_1 - \lambda_2\mathbf{x}_1^T\mathbf{x}_2 = \lambda_1\mathbf{x}_2^T\mathbf{x}_1 - \lambda_2\mathbf{x}_2^T\mathbf{x}_1 = (\lambda_1 - \lambda_2)\mathbf{x}_2^T\mathbf{x}_1.$$

Отсюда следует, что $\mathbf{x}_2^T\mathbf{x}_1 = 0$. Следовательно, собственные векторы линейного преобразования **R** образуют ортонормированный базис в *n*-мерном пространстве. Такие векторы называются главными компонентами корреляционной матрицы.

Главные компоненты корреляционной матрицы обладают весьма важными свойствами, которые имеют содержательный смысл в обработке данных и поэтому широко используются. Ниже мы покажем геометрический смысл главных компонент на плоскости.

ПРИВЕДЕНИЕ КОРРЕЛЯЦИОННОЙ МАТРИЦЫ К ДИАГОНАЛЬНОЙ ФОРМЕ

Преобразование корреляционной матрицы к диагональной форме основано на следующем свойстве вещественной (действительной) симметричной матрицы. Пусть **R** - невырожденная корреляционная матрица и имеет *n* различных собственных чисел $\lambda_i, i = 1,\dots n$. Пусть $\mathbf{a}_i, i = 1,\dots n$ - соответствующие собственные векторы, выбранные из пар собственных векторов, соответствующих каждому собственному числу, составляющие ортонормированный базис в *n*-мерном пространстве. Пусть $\mathbf{A} = (\mathbf{a}_1,\dots \mathbf{a}_n)$ - матрица, столбцами которой являются собственные векторы $\mathbf{a}_i$. Рассмотрим матрицу

$$\mathbf{A}^T\mathbf{A} = \begin{pmatrix}\mathbf{a}_1^T\\ \vdots \\ \mathbf{a}_n^T\end{pmatrix}(\mathbf{a}_1 \dots \mathbf{a}_n) = \begin{pmatrix}\mathbf{a}_1^T\mathbf{a}_1 \dots \mathbf{a}_1^T\mathbf{a}_n\\ \vdots \quad \cdots \quad \vdots \\ \mathbf{a}_n^T\mathbf{a}_1 \dots \mathbf{a}_n^T\mathbf{a}_n\end{pmatrix} = \begin{pmatrix}1 \ \dots \ 0\\ \ddots \\ 0 \ \dots \ 1\end{pmatrix} = \mathbf{E},$$

где **E** - единичная матрица. Следовательно, матрица **A** является ортогональной.

Напомним, что некоторая матрица **A** ортогональна, если $\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}^T\mathbf{A} = \mathbf{E}$. По уравнению $\mathbf{R}\mathbf{a} = \lambda\mathbf{a}$ получим $\mathbf{R}\mathbf{A} = (\lambda_1\mathbf{a}_1 \dots \lambda_n\mathbf{a}_n)$, где столбцами матрицы в правой части являются векторы $\lambda_i\mathbf{a}_i$. Учитывая, что векторы $\mathbf{a}_i$ ортогональны, получим

$$\mathbf{A}^T\mathbf{R}\mathbf{A} = \begin{pmatrix}\mathbf{a}_1^T\\ \vdots \\ \mathbf{a}_n^T\end{pmatrix}(\lambda_1\mathbf{a}_1 \dots \lambda_n\mathbf{a}_n) = \begin{pmatrix}\lambda_1\mathbf{a}_1^T\mathbf{a}_1 \dots \lambda_n\mathbf{a}_1^T\mathbf{a}_n\\ \vdots \quad \cdots \quad \vdots \\ \lambda_1\mathbf{a}_n^T\mathbf{a}_1 \dots \lambda_n\mathbf{a}_n^T\mathbf{a}_n\end{pmatrix} = \begin{pmatrix}\lambda_1 \ \dots \ 0\\ \ddots \\ 0 \ \cdots \ \lambda_n\end{pmatrix} = \Lambda.$$

Матрица $\mathbf{A}^T\mathbf{R}\mathbf{A}$ диагональна, и ее диагональные элементы являются собственными числами. Из условия $\mathbf{A}^T\mathbf{R}\mathbf{A} = \mathbf{\Lambda}$ следует $\mathbf{A}\mathbf{A}^T\mathbf{R}\mathbf{A}\mathbf{A}^T = \mathbf{A}\mathbf{\Lambda}\mathbf{A}^T$ и $\mathbf{R} = \mathbf{A}\mathbf{\Lambda}\mathbf{A}^T$, так как $\mathbf{A}\mathbf{A}^T = \mathbf{A}^T\mathbf{A} = \mathbf{E}$. Следовательно, невырожденная корреляционная матрица **R** может быть приведена к диагональной форме путем ортогонального преобразования $\mathbf{A}^T\mathbf{R}\mathbf{A}$.

Пусть $\mathbf{x} = (x_1,\dots x_n)^T$ - некоторый вектор, заданный своими проекциями на осях координат $X_i, i = 1,\dots n$. Рассмотрим вектор $\mathbf{y} = (y_1,\dots y_n)^T$, где $\mathbf{y} = \mathbf{A}^T\mathbf{x}$, а строками матрицы $\mathbf{A}^T$ являются собственные векторы $\mathbf{a}_i^T$ линейного преобразования **R**. Тогда

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \mathbf{a}_1^T\mathbf{x} \\ \vdots \\ \mathbf{a}_n^T\mathbf{x} \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^{n} a_{1j}x_j \\ \vdots \\ \sum_{j=1}^{n} a_{nj}x_j \end{pmatrix}.$$

Следовательно, компонента y_i вектора $\mathbf{y}$ - это скалярное произведение собственного вектора $\mathbf{a}_i$ и вектора $\mathbf{x}$. С другой стороны, скалярное произведение - это произведение модулей векторов $\mathbf{a}_i$ и $\mathbf{x}$ на косинус угла между ними. Так как $\|\mathbf{a}_i\| = 1$, то это есть произведение $\|\mathbf{x}\|$ на косинус угла между $\mathbf{a}_i$ и $\mathbf{x}$ - проекция вектора $\mathbf{x}$ на $\mathbf{a}_i$. Поэтому вектор $\mathbf{y}$ представлен своими проекциями на ортонормированный базис собственных векторов корреляционной матрицы $\mathbf{R}$. Можно считать, что новый базис $\mathbf{a}_i, i = 1,\dots n$ образует новое n-мерное пространство признаков $Y_i = (y_1,\dots y_N)^T$, $i = 1,\dots n$, принимающих свои значения на N объектах.

Значения n признаков Y_i, как бы измеренных на N объектах, образуют новую матрицу данных $\mathbf{Y} = \mathbf{XA}$, полученную из матрицы $\mathbf{X}$ ортогональным преобразованием $\mathbf{A}$:

$$\mathbf{Y} = \begin{pmatrix} \mathbf{y}_1^T \\ \vdots \\ \mathbf{y}_N^T \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1^T\mathbf{a}_1 & \cdots & \mathbf{x}_1^T\mathbf{a}_n \\ \vdots & \cdots & \vdots \\ \mathbf{x}_N^T\mathbf{a}_1 & \cdots & \mathbf{x}_N^T\mathbf{a}_n \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_N^T \end{pmatrix}(\mathbf{a}_1 \cdots \mathbf{a}_n) = \mathbf{XA}.$$

Корреляционная матрица $\mathbf{R}$, вычисленная по матрице $\mathbf{X}$, представляет собой матрицу

$$\mathbf{R} = \frac{1}{N}\begin{pmatrix} X_1^T X_1 & \cdots & X_1^T X_n \\ \vdots & \cdots & \vdots \\ X_n^T X_1 & \cdots & X_n^T X_n \end{pmatrix} = \frac{1}{N}\begin{pmatrix} X_1^T \\ \vdots \\ X_n^T \end{pmatrix}(X_1 \cdots X_n) = \frac{1}{N}\mathbf{X}^T\mathbf{X}.$$

Вычислим среднее признака Y_j

$$\overline{y}_j = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i^T\mathbf{a}_j = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{n} x_{ik}a_{kj} = \frac{1}{N}\sum_{k=1}^{n} a_{kj}\sum_{i=1}^{N} x_{ik} = \sum_{k=1}^{n} a_{kj}\overline{x}_k = 0,$$

так как матрица $\mathbf{X}$ стандартизована. Вычислим величину

$$\frac{1}{N}\mathbf{Y}^T\mathbf{Y} = \frac{1}{N}(\mathbf{XA})^T(\mathbf{XA}) = \frac{1}{N}\mathbf{A}^T\mathbf{X}^T\mathbf{XA} = \mathbf{A}^T\mathbf{RA} = \mathbf{\Lambda} = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ & \ddots & \\ 0 & \cdots & \lambda_n \end{pmatrix}.$$

Тогда матрица $\mathbf{\Lambda}$ является ковариационной матрицей, вычисленной по матрице $\mathbf{Y}$. Диагональная структура матрицы $\mathbf{\Lambda}$ показывает, как и следовало ожидать, независимость признаков $Y_i, i = 1,\dots n$. Собственные числа λ_i являются дисперсиями этих признаков, то есть $\lambda_i = \sigma_i^2$. Если разделить значения компонент каждого признака Y_i на величину $\sigma_i = \sqrt{\lambda_i}$, то матрица $\mathbf{Y}$ будет приведена к стандартизованному виду. Тогда преобразование $\mathbf{Y} = \mathbf{XA\Lambda}^{-1/2}$ даст стандартизованную матрицу данных $\mathbf{Y}$ с единичной корреляционной матрицей:

$$\frac{1}{N}\mathbf{Y}^T\mathbf{Y} = \frac{1}{N}\left(\mathbf{XA\Lambda}^{-1/2}\right)^T\left(\mathbf{XA\Lambda}^{-1/2}\right) = \frac{1}{N}\mathbf{\Lambda}^{-1/2}\mathbf{A}^T\mathbf{X}^T\mathbf{XA\Lambda}^{-1/2} =$$
$$= \mathbf{\Lambda}^{-1/2}\mathbf{A}^T\mathbf{RA\Lambda}^{-1/2} = \mathbf{\Lambda}^{-1/2}\mathbf{\Lambda\Lambda}^{-1/2} = \mathbf{\Lambda}^{1/2}\mathbf{\Lambda}^{-1/2} = \mathbf{E}.$$

ВЫЧИСЛЕНИЕ СОБСТВЕННЫХ ЧИСЕЛ И ВЕКТОРОВ КОРРЕЛЯЦИОННОЙ МАТРИЦЫ

В задачах обработки часто возникает необходимость в определении собственных векторов корреляционной матрицы, соответствующих тем или иным собственным числам. Как было показано, для нахождения собственных чисел и векторов следует найти корни характеристического полинома порядка *n* относительно λ. Затем для каждого λ_i, $i=1,\dots n$ следует найти свой собственный вектор, который мы обозначим как $\mathbf{a}_i = \left(a_{1i},\dots a_{ni}\right)^T$, как решение однородной системы линейных уравнений относительно этого собственного вектора при ограничении на его длину $\|\mathbf{a}_i\| = \sum_{j=1}^{n} a_{ji}^2 = 1$.

Известно, что точные методы поиска корней полинома и корней системы линейных уравнений представляют собой громоздкие процедуры при больших *n*, практически начиная с $n \geq 3$. Поэтому данная задача часто решается итерационными методами вычислительной математики. Итерационные методы для одновременного поиска всех собственных чисел и векторов представляют собой методы преобразования симметричной матрицы в диагональную форму. Часто требуется вычислить только максимальное собственное число и соответствующий ему собственный вектор. Рассмотрим известный итерационный метод приближенного вычисления максимального собственного числа и соответствующего собственного вектора.

Пусть все собственные числа различны и упорядочены $\lambda_1 > \lambda_2 > \dots > \lambda_n > 0$. Пусть $\mathbf{x} = \left(x_1,\dots x_n\right)^T$ - некоторый вектор. Совокупность собственных векторов $\mathbf{a}_i$, $i=1,\dots n$ корреляционной матрицы **R** образует ортонормированный базис, в пространстве которого вектор **x** преобразуется в вектор $\mathbf{y} = \mathbf{A}^T\mathbf{x}$. Т.к. матрица **A** ортогональна, то $\mathbf{y} = \mathbf{A}^{-1}\mathbf{x}$ и $\mathbf{x} = \mathbf{A}\mathbf{y}$, где вектор **x** представлен своим разложением по базису собственных векторов:

$$\mathbf{x} = \left(\mathbf{a}_1,\dots \mathbf{a}_n\right)\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = y_1\mathbf{a}_1 + \dots + y_n\mathbf{a}_n = \sum_{i=1}^{n} y_i\mathbf{a}_i \ .$$

Тогда $\mathbf{R}\mathbf{x} = \mathbf{R}\left(\sum_{i=1}^{n} y_i\mathbf{a}_i\right) = \sum_{i=1}^{n} y_i\mathbf{R}\mathbf{a}_i = \sum_{i=1}^{n} y_i\lambda_i\mathbf{a}_i$.

Выделим первое слагаемое

$$\mathbf{R}\mathbf{x} = y_1\lambda_1\mathbf{a}_1 + \sum_{i=2}^{n} y_i\lambda_i\mathbf{a}_i = \lambda_1\left(y_1\mathbf{a}_1 + \sum_{i=2}^{n} y_i\frac{\lambda_i}{\lambda_1}\mathbf{a}_i\right).$$

Умножим это равенство еще раз слева на **R**:

$$\mathbf{R}\mathbf{R}\mathbf{x} = \mathbf{R}^2\mathbf{x} = \mathbf{R}\lambda_1\left(y_1\mathbf{a}_1 + \sum_{i=2}^{n} y_i\frac{\lambda_i}{\lambda_1}\mathbf{a}_i\right) = \lambda_1\left(y_1\mathbf{R}\mathbf{a}_1 + \sum_{i=2}^{n} y_i\frac{\lambda_i}{\lambda_1}\mathbf{R}\mathbf{a}_i\right) =$$

$$\lambda_1\left(y_1\lambda_1\mathbf{a}_1 + \sum_{i=2}^{n} y_i\frac{\lambda_i^2}{\lambda_1}\mathbf{a}_i\right) = \lambda_1^2\left(y_1\mathbf{a}_1 + \sum_{i=2}^{n} y_i\left(\frac{\lambda_i}{\lambda_1}\right)^2\mathbf{a}_i\right).$$

Тогда для некоторого s получим

$$\mathbf{R}^s\mathbf{x} = \lambda_1^s\left(y_1\mathbf{a}_1 + \sum_{i=2}^{n} y_i \left(\frac{\lambda_i}{\lambda_1}\right)^s \mathbf{a}_i \right).$$

Так как $\lambda_1 > \lambda_2 > \ldots > \lambda_n > 0$ и $0 \le \frac{\lambda_i}{\lambda_1} < 1$, то $\lim\limits_{s\to\infty}\left(\frac{\lambda_i}{\lambda_1}\right)^s = 0$ и $\lim\limits_{s\to\infty}\mathbf{R}^s\mathbf{x} = \lambda_1^s y_1\mathbf{a}_1$.

Тогда при $y_1 \neq 0$ первый собственный вектор определяется достаточно далеким членом последовательности $\mathbf{x}, \mathbf{Rx}, \mathbf{R}^2\mathbf{x}, \ldots \mathbf{R}^s\mathbf{x}, \ldots$. Но при $\lambda_1 > 1$ получим, что $\lim\limits_{s\to\infty}\lambda_1^s y_1\mathbf{a}_1 = \infty$, а при $\lambda_1 < 1$ получим, что $\lim\limits_{s\to\infty}\lambda_1^s y_1\mathbf{a}_1 = 0$. Следовательно, вектор $\mathbf{R}^s\mathbf{x}$ стремится по направлению к вектору $\mathbf{a}_1$, но его длина значительно отличается от единичной.

Поэтому строят две другие последовательности $\mathbf{x}_0, \mathbf{x}_1, \ldots \mathbf{x}_s, \ldots$ и $\mathbf{z}_1, \ldots \mathbf{z}_s, \ldots$, где $\mathbf{z}_s = \mathbf{R}\mathbf{x}_{s-1}$, $\mathbf{x}_s = \mathbf{z}_s / \| \mathbf{z}_s \|$, начиная с некоторого вектора $\mathbf{x}_0$ единичной длины. Следовательно, $\| \mathbf{x}_s \| = 1$ при любом s, а предел последовательности $\{\mathbf{x}_s\}$ стремится по направлению к вектору $\mathbf{a}_1$. Следовательно, $\lim\limits_{s\to\infty}\mathbf{x}_s = \mathbf{a}_1$. Тогда $\mathbf{z}_{s+1} = \mathbf{R}\mathbf{x}_s \approx \lambda_1\mathbf{a}_1$ и $\|\mathbf{z}_{s+1}\| = \|\lambda_1\mathbf{a}_1\| = \lambda_1\sqrt{\sum_{i=1}^{n} a_{i1}^2} = \lambda_1$.

Задание на работу

Ознакомиться с теоретической справкой к данной лабораторной работе. Найти собственные векторы и собственный числа корреляционной матрицы.

Содержание отчета

Номер и название лабораторной работы;
Цель лабораторной работы;
Пояснительная записка к проекту;
Выводы.

Контрольные вопросы

1. Основные свойства корреляционной матрицы
2. Собственные векторы и собственный числа квадратной матрицы
3. Интерпретация главных компонент на плоскости
4. Алгоритм приближенного вычисления максимального собственного числа

Лабораторная работа №3

ВИЗУАЛИЗАЦИЯ ДАННЫХ В ПРОСТРАНСТВЕ ГЛАВНЫХ КОМПОНЕНТ

Цель и задача работы

2D и 3D представление данных в пространстве первых главных компонент

Теоретические положения

ПРЕДСТАВЛЕНИЕ ГЛАВНЫХ КОМПОНЕНТ НА ПЛОСКОСТИ

Пусть в соответствии со статистической гипотезой порождения матрицы данных **X** в *n*-мерном пространстве признаков существует многомерное нормальное распределение с плотностью вероятности $f(\mathbf{x}/\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Для стандартизованной матрицы **X** мы полагаем, что

$$f(\mathbf{x}/\mathbf{0}, \mathbf{R}) = \frac{1}{\sqrt{(2\pi)^n \det \mathbf{R}}} \exp\left(-\frac{1}{2}\mathbf{x}^T \mathbf{R}^{-1}\mathbf{x}\right).$$

Проведем ортогональное преобразование матрицы данных **X** в новую матрицу данных $\mathbf{Y} = \mathbf{XA}$, где **A** - матрица, столбцами которой являются собственные векторы корреляционной матрицы **R**. Тем самым мы перешли в новое признаковое пространство, образованное ортонормированным базисом линейного преобразования **R**. Очевидно, что в новом признаковом пространстве задано нормальное распределение с плотностью вероятности

$$f(\mathbf{y}/\mathbf{0}, \boldsymbol{\Lambda}) = \frac{1}{\sqrt{(2\pi)^n \det \boldsymbol{\Lambda}}} \exp\left(-\frac{1}{2}\mathbf{y}^T \boldsymbol{\Lambda}^{-1}\mathbf{y}\right).$$

Так как $\boldsymbol{\Lambda} = \mathbf{A}^T\mathbf{RA}$, то $\boldsymbol{\Lambda}^{-1} = \left(\mathbf{A}^T\mathbf{RA}\right)^{-1} = \mathbf{A}^{-1}\mathbf{R}^{-1}\left(\mathbf{A}^{\mathbf{T}}\right)^{-1} = \mathbf{A}^T\mathbf{R}^{-1}\mathbf{A}$,

$$\det \boldsymbol{\Lambda} = \det\begin{pmatrix} \lambda_1 \cdots 0 \\ \ddots \\ 0 \cdots \lambda_n \end{pmatrix} = \prod_{i=1}^{n} \lambda_i; \quad \boldsymbol{\Lambda}^{-1} = \begin{pmatrix} \frac{1}{\lambda_1} \cdots & 0 \\ & \ddots \\ 0 & \cdots \frac{1}{\lambda_n} \end{pmatrix}.$$

Тогда $f(\mathbf{y}/\mathbf{0}, \boldsymbol{\Lambda}) = \frac{1}{\sqrt{(2\pi)^n} \prod_{i=1}^{n} \sqrt{\lambda_i}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} \frac{y_i^2}{\lambda_i}\right)$.

Пусть $n = 2$, тогда двухмерное нормальное распределение имеет вид

$$f(\mathbf{y}/\mathbf{0}, \boldsymbol{\Lambda}) = \frac{1}{2\pi\sqrt{\lambda_1\lambda_2}} \exp\left[-\frac{1}{2}\left(\frac{y_1^2}{\lambda_1} + \frac{y_2^2}{\lambda_2}\right)\right].$$

Рассмотрим уравнение $\frac{y_1^2}{\lambda_1} + \frac{y_2^2}{\lambda_2} = p, \quad p > 0$. Из курса аналитической геометрии известно, что это уравнение линии второго порядка. При заданном p и найденных λ_1 и λ_2 данная линия является линией постоянного значения плотности вероятности

$\frac{1}{2\pi\sqrt{\lambda_1\lambda_2}}e^{-p/2}$. Преобразуем данное уравнение линии второго порядка к каноническому виду $\frac{y_1^2}{p\lambda_1}+\frac{y_2^2}{p\lambda_2}=1$. Так как $\lambda_1>\lambda_2>0$, то данное уравнение является каноническим уравнением эллипса в системе координат, образованной собственными векторами, которые соответствуют собственным числам λ_1 и λ_2.

Если $r>0$, то $\lambda_1=1+r$ и $\lambda_2=1-r$ и система главных компонент y_10y_2 повернута на 45^0 относительно исходной системы координат x_10x_2. Если $r<0$, то $\lambda_1=1-r$ и $\lambda_2=1+r$ и система главных компонент y_10y_2 повернута на 135^0 относительно x_10x_2. Если $r=0$, то $\lambda_1=\lambda_2=1$. Тогда уравнение эллипса представляет собой уравнение окружности $y_1^2+y_2^2=p$ радиуса $\sqrt{p}$. В этом случае система главных компонент y_10y_2 может быть ориентирована в любом направлении, то есть любое направление является главным для такого линейного преобразования **R**. Если $r=1$, то $\lambda_1=2,\lambda_2=0$. Тогда уравнение эллипса для линии постоянного значения плотности вероятности вырождается в уравнение для двух точек, расположенных на оси $0y_1$, вида $y_1=\pm\sqrt{p(1+r)}=\pm\sqrt{2p},\ y_2=0$ (Рис. 2.1).

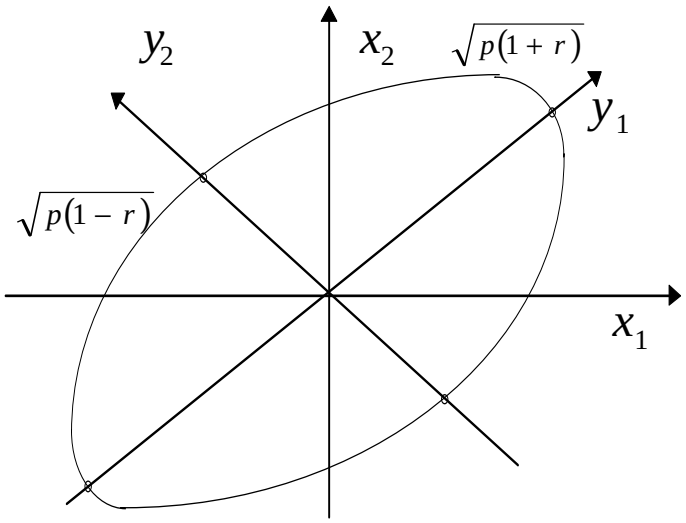

Рис.2.1. Главные компоненты

Определим уравнение максимального эллипса в соответствии с правилом “трех сигм”, согласно которому 99.73% всех наблюдений сосредоточено внутри него. Согласно свойствам канонического уравнения эллипса его главная ось совпадает с направлением первой главной компоненты $0y_1$. Длина главной полуоси составляет величину $\sqrt{p\lambda_1}$. В то же время максимальное положительное случайное отклонение величины y_1 на оси $0y_1$ от центра координат с вероятностью 0.9973 не превышает величины $3\sigma_1=3\sqrt{\lambda_1}$. Следовательно, $\sqrt{p\lambda_1}=3\sqrt{\lambda_1}$, откуда $p=9$. Проведя те же рассуждения для второй оси максимального эллипса, получим, что уравнение имеет вид $\frac{y_1^2}{9\lambda_1}+\frac{y_2^2}{9\lambda_2}=1$ и описывает линию постоянного значения плотности вероятности на уровне $f(p)=\frac{1}{2\pi\sqrt{\lambda_1\lambda_2}}e^{-p/2}\approx\frac{1}{6.28\sqrt{1-r^2}}e^{-4.5}=\frac{0.004}{\sqrt{1-r^2}}$.

Так как длина главной полуоси равна $\sqrt{p\lambda_1} = \sqrt{p(1+r)}$, то при увеличении значения r длина главной полуоси увеличивается. В то же время длина второй полуоси эллипса $\sqrt{p\lambda_2} = \sqrt{p(1-r)}$ уменьшается при увеличении r. Следовательно, чем сильнее связаны признаки X_1 и X_2 корреляционной зависимостью, тем больше дисперсия $\sigma_1^2 = \lambda_1$ признака Y_1 и меньше дисперсия $\sigma_2^2 = \lambda_2$ признака Y_2 при неизменной суммарной дисперсии $\sigma_1^2 + \sigma_2^2 = \lambda_1 + \lambda_2 = 2$.

МОДЕЛЬ ГЛАВНЫХ КОМПОНЕНТ

Пусть новая матрица данных получена путем ортогонального преобразования $\mathbf{Y} = \mathbf{X}\mathbf{A}\mathbf{\Lambda}^{-1/2}$. Такая матрица является стандартизованной, то есть $\frac{1}{N}\mathbf{Y}^T\mathbf{Y} = \mathbf{E}$. Тогда преобразование некоторого вектора $\mathbf{x}$ к вектору $\mathbf{y}$ выполняется как $\mathbf{y}^T = \mathbf{x}^T\mathbf{A}\mathbf{\Lambda}^{-1/2}$ или $\mathbf{y} = \mathbf{\Lambda}^{-1/2}\mathbf{A}^T\mathbf{x}$. Выполним обратное преобразование $\mathbf{X} = \mathbf{Y}\mathbf{\Lambda}^{1/2}\mathbf{A}^T$. Тем самым мы выразили матрицу исходных данных через матрицу $\mathbf{Y}$.

Согласно гипотезе скрытых факторов, значение каждого исходного признака, измеренного на некотором объекте, зависит от влияния некоторых “скрытых” неизмеряемых факторов и определяется совокупностью их вкладов, пропорциональных силе влияния. Тогда матрицу $\mathbf{Y}$ будем считать матрицей n скрытых факторов, а матрицу $\mathbf{U}^T = \mathbf{\Lambda}^{1/2}\mathbf{A}^T$ - матрицей факторных нагрузок. Тогда каждая компонента некоторого вектора $\mathbf{x}$, измеренного на некотором объекте, представляется как совокупность значений факторов на этом объекте $\mathbf{x}^T = \mathbf{y}^T\mathbf{\Lambda}^{1/2}\mathbf{A}^T$ или $\mathbf{x} = \mathbf{A}\mathbf{\Lambda}^{1/2}\mathbf{y} = \mathbf{U}\mathbf{y}$:

$$\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \left(\sqrt{\lambda_1}\mathbf{a}_1 \cdots \sqrt{\lambda_n}\mathbf{a}_n\right)\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \sum_{i=1}^{n} y_i\sqrt{\lambda_i}\mathbf{a}_i = \begin{pmatrix} \sum_{i=1}^{n}\sqrt{\lambda_i}a_{1i}y_i \\ \vdots \\ \sum_{i=1}^{n}\sqrt{\lambda_i}a_{ni}y_i \end{pmatrix}.$$

Тогда корреляционная матрица имеет вид

$$\mathbf{R} = \frac{1}{N}\mathbf{X}^T\mathbf{X} = \frac{1}{N}\left(\mathbf{Y}\mathbf{\Lambda}^{1/2}\mathbf{A}^T\right)^T\left(\mathbf{Y}\mathbf{\Lambda}^{1/2}\mathbf{A}^T\right) = \frac{1}{N}\mathbf{A}\mathbf{\Lambda}^{1/2}\mathbf{Y}^T\mathbf{Y}\mathbf{\Lambda}^{1/2}\mathbf{A}^T =$$

$$= \left(\mathbf{A}\mathbf{\Lambda}^{1/2}\right)\left(\mathbf{\Lambda}^{1/2}\mathbf{A}^T\right) = \mathbf{U}\mathbf{U}^T = \mathbf{A}\mathbf{\Lambda}\mathbf{A}^T.$$

Рассмотрим взаимные корреляции между признаками из $\mathbf{X}$ и факторами из матрицы $\mathbf{Y}$

$$\frac{1}{N}\mathbf{X}^T\mathbf{Y} = \frac{1}{N}\left(\mathbf{Y}\mathbf{\Lambda}^{1/2}\mathbf{A}^T\right)^T\mathbf{Y} = \frac{1}{N}\mathbf{A}\mathbf{\Lambda}^{1/2}\mathbf{Y}^T\mathbf{Y} = \mathbf{A}\mathbf{\Lambda}^{1/2} = \mathbf{U}.$$

Следовательно, матрица $\mathbf{U}$ факторных нагрузок является матрицей взаимных корреляций между исходными признаками и скрытыми факторами, где элемент $u_{ij} = a_{ij}\sqrt{\lambda_i}$ равен величине взаимной корреляции между признаком X_i и фактором Y_j. Рассмотрим структуру корреляционной матрицы

$$\mathbf{R} = \mathbf{U}\mathbf{U}^T = \begin{pmatrix} \sum_{i=1}^{n} \lambda_i a_{1i}^2 & \cdots & \sum_{i=1}^{n} \lambda_i a_{1i} a_{ni} \\ \vdots & \cdots & \vdots \\ \sum_{i=1}^{n} \lambda_i a_{ni} a_{1i} & \cdots & \sum_{i=1}^{n} \lambda_i a_{ni}^2 \end{pmatrix}.$$

Дисперсия $\sigma_k^2 = r_{kk} = \sum_{i=1}^{n} \lambda_i a_{ki}^2 = \sum_{i=1}^{n} u_{ki}^2 = 1$ некоторого признака X_k есть величина, состоящая из вкладов соответствующих главных компонент. Полный вклад всех главных компонент в дисперсии всех признаков составляет величину

$$\sum_{k=1}^{n} \sigma_k^2 = \sum_{k=1}^{n} \left(\sum_{i=1}^{n} \lambda_i a_{ki}^2 \right) = \mathrm{tr}\mathbf{U}\mathbf{U}^T = \mathrm{tr}\mathbf{R} = \mathrm{tr}\mathbf{\Lambda} = n.$$

При преобразовании к главным компонентам вместо n исходных признаков получается такое же число факторов. Но вклад довольно большой части главных компонент в суммарную дисперсию признаков является небольшим. Поэтому часто целесообразно исключить те главные компоненты, вклад которых невелик. При этом оказывается, что при помощи m первых наиболее весомых главных компонент, где $m<n$, можно объяснить основную долю суммарной дисперсии признаков. Эта доля называется объясняемой долей дисперсии $h^2 = \sum_{i=1}^{m} \sigma_i^2 = \sum_{i=1}^{m} \lambda_i < n$, где обычно $h^2 / n \geq 0.8$.

Задание на работу

Ознакомиться с теоретической справкой к данной лабораторной работе. Найти собственные векторы и собственный числа корреляционной матрицы.

Выполнить проекцию объектов на первые два и первые три собственных вектора, отобразить полученные векторы на плоскости (2D) и в ортогональной проекции трех координат (3D).

Содержание отчета

Номер и название лабораторной работы;
Цель лабораторной работы;
Пояснительная записка к проекту;
Выводы.

Контрольные вопросы

1. Основные свойства корреляционной матрицы
2. Собственные векторы и собственный числа квадратной матрицы
3. Интерпретация главных компонент на плоскости
4. Алгоритм приближенного вычисления максимального собственного числа
5. Методы визуализации данных
6. Задача Карунена-Лоэва