Глава 2. Плетение и закидывание сетей и неводов

2.1. Предобработка данных

Среди m признаков (которые иначе могут называться *переменными*) могут быть признаки, измеряемые в количественной, номинальной или порядковой шкалах. В самом простом случае все признаки измеряются в одной и той же шкале, но в реальных ситуациях, как правило, используются несколько типов шкал измерения признаков.

Перед применением к данным различных алгоритмов анализа их структуры всегда возникает необходимость применения различных способов предобработки, которая в задаче визуализации данных заключается в *оцифровке*, *нормировке* и *выборе метрики*.

2.1.1. Обозначения

Для того, чтобы последующее изложение было более ясным, введем систему обозначений, которой будем придерживаться на протяжении всей последующей главы.

X_i, Y_i – совокупность координат i-ой точки данных (радиус-вектор);

$X_i \cdot Y_i$ или (X_i, Y_i)– скалярное произведение вектора X_i и Y_j;

$(X_i)^2$ – «квадрат вектора» – сумма квадратов его координат (число);

x_{ij} – значение j-ой координаты i-ой точки объекта (число);

ξ_i, η_i – обозначения i-ой координаты пространства данных (как меняющейся величины);

m – размерность пространства данных;

$|X|$, N – число объектов;

δ_{ij} – «дельта-символ» Кронеккера: $\begin{cases} \delta_{ij} = 1, & i = j \\ \delta_{ij} = 0, & i \neq j \end{cases}$.

2.1.2. Оцифровка дискретных шкал

Под оцифровкой, как правило, подразумевается приведение всех типов признаков к одной количественной шкале.

В самом простом случае дихотомической шкалы, то есть когда признак может принимать значения «да» или «нет», нет большой разницы какие числа будут приписаны положительному или отрицательному ответу. Самые распространенные варианты: ответу «да» приписывают число 1, ответу «нет» – либо –1, либо 0.

В случае порядковых шкал, как правило, порядок следования градаций признака отражает степень усиления или ослабления того или иного качества. Числовые метки признака в этом случае присваиваются таким

образом, чтобы расстояния между двумя отметками интуитивно соответствовали разнице между соответствующими градациями (например, если признак имеет шкалу «плохо-никак-хорошо», то логично приписать градациям метки –1;0;1, а вот в случае шкалы «малый-средний-крупный-сверхкрупный» более уместным может оказаться использовать логарифмические метки, т.е. 0.1;1;10;100). Не играет большого значения выбор «начала отсчета» шкалы признака, но психологически удобнее измерять в числовой шкале с отрицательными и положительными числовыми метками качество, меняющееся от противоположности к противоположности (как «плохо-хорошо»), а в шкалах с «абсолютным нулем» измерять постепенное нарастание какого-либо качества (как «отсутствие-частичное присутствие-полное присутствие»).

Большей свободой и математическим осмыслением обладает процедура оцифровки номинальных шкал. В это случае, как правило, не играет роли порядок следования и расстояния между градациями признаков.

✂ Хотя возможны и исключения – для примера возьмем признак «Отрасль промышленности». Значения признака «цветная металлургия» и «черная металлургия» психологически воспринимаются ближе, чем «нефтяная промышленность» и «пищевая промышленность», хотя признак, обозначающий принадлежность предприятия к определенной отрасли нельзя назвать порядковым. Для упорядочивания значений шкалы признака можно пользоваться методами многомерного шкалирования и в конечном счете возможно, что признаку будут отвечать не одна, а несколько числовых меток, если построенное «пространство восприятия» окажется эффективно двумерным. ✂

Свобода в выборе числовых меток для номинальных шкал дает возможность искусственно упростить структуру набора данных, например, добиться того, чтобы шкалы признаков были максимально скоррелированы друг с другом. Популярным методом является максимизация функционала Q, где

$$Q^2 = \sum_{i<j}^{m} r_{ij}^2 ,$$

где r_{ij} – коэффициент линейной корреляции между i-ым и j-ым признаком, m – число признаков, среди которых есть как номинальные, так и количественные. Допустим, что номинальным признаков уже были каким-то образом присвоены числовые метки. Тогда

$$Q^2 = Q_1^2 + Q_{1,2}^2 + Q_2^2 ,$$

где в Q_1 входят коэффициенты корреляции между номинальными признаками, подлежащими оцифровке, в $Q_{1,2}$ – коэффициенты корреляции между номинальными и количественными признаками, а Q_2 – часть функ-

ционала, состоящая из коэффициентов корреляции между количественными признаками. Последняя не зависит от оцифровки и оптимизации на самом деле подлежит функционал $\tilde{Q}$, где

$$\tilde{Q}^2 = Q_1^2 + Q_{1,2}^2$$

Пусть $X^{(1)}$ – набор номинальных признаков, подлежащих оцифровке, $X^{(2)}$ – набор количественных признаков, c_k^i – числовая метка, присвоенная k-ой категории i-го номинального признака. $N(i,j)$ – матрица сопряженности между i-ым и j-ым номинальными признаками, в качестве оценки которой можно взять числа $n_{kl}(i,j)$ одновременного появления для i-го признака категории k, а для j-го признака – категории l, $p_{i,k}$ – частота появления k-ой градации признака i, который имеет l_i градаций, $P_i=\mathrm{diag}(p_{i,1};p_{i,2}..,p_{i,li})$; наконец, $\bar{c}_k^{ij}$ – среднее значение количественного признака j ($j\in X^{(2)}$) на тех объектах, у которых i-ый номинальный признак имеет категорию k.

Тогда градиент функционала

$$\frac{\partial \tilde{Q}}{\partial c_k^i} = \sum_{j=1}^{l_i} a_{kj}^i c_j^i$$

, где

$$a_{kj}^i = \sum_{\substack{m\in X^{(1)} \\ m\neq i}} \sum_{i_1=1}^{l_m}\sum_{i_2=1}^{l_m} n_{ki_1}(i,m) c_{i_1}^m c_{i_2}^m n_{i_2 j}(m,i) + \sum_{m\in X^{(2)}} \bar{c}_k^{im}\bar{c}_j^{im} p_j p_k \quad , \quad i\in X^{(1)},$$

$k=1..l_i$.

При этом предполагается, что все данные (включая номинальные признаки) предварительно нормированы на единичную дисперсию и центрированы. Для номинальных признаков это означает, что

$$\sum_{i=1}^{N} c_k^j(i) = 0 \quad , \quad \frac{1}{N}\sum_{i=1}^{N}(c_k^j(i))^2 = 1$$

для всех j, k; N – число объектов; $c_k^j(i)$ – значение числовой метки для i-го объекта. Выполнения этих условий всегда можно добиться линейным преобразованием.

Зная градиент функционала $\tilde{Q}$ можно воспользоваться любым из методов градиентной оптимизации. Для количественных шкал значения признаков в результате не меняются, а значения меток номинальных признаков назначаются категориям таким образом, чтобы они были максимально скореллированы друг с другом и с количественными признаками, что приводит к снижению эффективной размерности пространства данных (часть признаков становится статистически линейно зависима от других).

2.1.3. Нормировка данных

После того, как все признаки оказываются описанными в количественной шкале, их обычно *центрируют* и *нормируют*.

Первым шагом в статистической обработке данных, как правило, является нахождение точки среднего значения всех признаков – геометрического центра многомерного облака точек данных. Обычно удобно сдвинуть все точки данных на один и тот же вектор таким образом, чтобы центр облака оказался в начале координат.

Далее следует нормировка – то есть деление всех значений признаков на определенное число таким образом, чтобы значения признаков попадали в сопоставимые по величине интервалы. В качестве такого числа обычно выбирается один из *характерных масштабов*.

В многомерном облаке данных существует несколько масштабов. Во-первых, это квадратный корень из общей дисперсии облака данных, называемый среднеквадратичным отклонением:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2}, \quad \bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i .$$

Напомним, что здесь и далее большими буквами X_i обозначаются вектора данных, а маленькими – x_{ij}: j-ая координата i-го вектора.

В случае, если выборка может считаться полученной из нормального распределения, то в шаре с центром в $\bar{x}$ радиусом σ находится около двух третей от числа точек данных.

Существует масштаб, характеризующий максимальный разброс в облаке данных

$$R = \max_{i=1..N} \|X_i - \bar{X}\| .$$

Нормировка всех признаков на R приводит к тому, что все облако данных оказывается заключено в шар единичного радиуса.

Если в качестве масштаба выбраны σ или R, то соответствующие формулы предобработки (нормировки на «единичную дисперсию» и «на единичный шар») имеют вид:

$$\tilde{X}_i = \frac{X_i - \bar{X}}{\sigma}, \quad \tilde{X}_i = \frac{X_i - \bar{X}}{R},$$

где $\tilde{X}_i, X_i$ – новые и старые значения векторов признаков.

С помощью масштаба σ, как правило, определяются понятия кластера и сгущения в облаке данных. Приведем эти определения (см. [4]).

Кластер – группа точек G такая, что средний квадрат внутригруппового расстояния до центра группы меньше среднего расстояния до общего центра в исходном наборе объектов, т.е. $\bar{d}_G^2 < \sigma^2$, где

$$\bar{d}_G^2 = \frac{1}{N}\sum_{X_i \in G}(X_i - \bar{X}_G)^2, \quad \bar{X}_G = \frac{1}{N}\sum_{x_i \in G} X_i .$$

Сгущение – группа точек G такая, что максимальный квадрат расстояния точек из G до центра группы меньше σ^2, т.е. $\bar{d}^2_{G,\max} < \sigma^2$, где $\bar{X}_G = \frac{1}{N}\sum_{x_i \in G} X_i$.

✂ К сожалению, такие определения не всегда соответствуют интуитивным представлениям о кластерах и сгущениях – например, когда сгущение представляет из себя сильно вытянутое облако точек. ✂

Кроме того, если диапазоны значений для разных признаков очень сильно отличаются друг от друга, то разумно для каждого из признаков применять собственный масштаб. Для каждого из признаков можно ввести свое среднеквадратичное отклонение и разброс:

$$\sigma_j = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_{ij} - \bar{x}_j)^2}, \quad \bar{x}_j = \frac{1}{N}\sum_{i=1}^{N} x_{ij}, \quad R_j = \max_{i=1..N}\left\|x_{ij} - \bar{x}_j\right\|,$$

где x_{ij} – значение j-го признака на i-ом объекте. Как результат, получаем формулы для нормировки на «единичную дисперсию для каждого из признаков» и «на единичный куб»:

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j}, \quad \tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_j}{R_j}.$$

Эти нормировки не являются «изотропными», то есть они сжимают облако данных в некоторых направлениях сильнее, в некоторых – меньше, что в некоторых случаях является желательным, а в некоторых – нарушает структуру данных (взаимных расстояний). Такая нормировка фактически эквивалентна выбору взвешенной евклидовой метрики, о которой речь пойдет ниже.

Наконец, следует упомянуть, что с каждым из количественных признаков могут быть связаны еще два масштаба – это точность и допуск (см раздел 1.1), с помощью которых также можно «обезразмерить» значения этих признаков.

2.1.4. Выбор метрики для пространства данных

Любая нормировка данных приводит к тому, что изменяются взаимные расстояния между точками данных. Это можно истолковать как выбор метрики иной по сравнению с обычной евклидовой. Выбор метрики является важным моментом в любой методике анализа структуры данных.

Сначала введем понятия *матрицы связи* признаков и *матрицы расстояний* между объектами.

Матрица связи – квадратная симметрическая матрица размерами $m \times m$ типа «признак-признак», где на пересечении i-ой строки j-ого столбца сто-

ит мера «взаимосвязанности» *i*-го и *j*-го признака. Самой популярной мерой связи количественных признаков является коэффициент корреляции Пирсона, который вычисляется по формуле

$$r_{kj} = \frac{s_{kj}}{\sqrt{s_{kk} s_{jj}}}, \text{ где } \quad s_{kj} = \frac{1}{N-1} \sum_{i=1}^{N} (x_{ik} - \bar{x}_k)(x_{ij} - \bar{x}_j).$$

В результате матрицей связи становится *корреляционная матрица*

$$R = \begin{bmatrix} r_{11} & \dots & r_{1m} \\ \dots & \dots & \dots \\ r_{m1} & \dots & r_{mm} \end{bmatrix}.$$

✂ В случае применения порядковых и номинальных шкал могут применяться другие меры связи такие как коэффициент ранговой корреляции для вычисления связи между порядковыми признаками или бисериальный коэффициент корреляции, применяемый для измерения связи между порядковым и количественным признаком. Подробный анализ этих коэффициентов приведен в [28] ✂

Матрица расстояний – квадратная матрица размерами $N \times N$ типа «объект-объект», где на пересечении *i*-ой строки *j*-ого столбца стоит мера удаленности между *i*-ым и *j*-ым объектом:

$$D = \begin{bmatrix} d_{11} & \dots & d_{1N} \\ \dots & d_{ij} & \dots \\ d_{N1} & \dots & d_{NN} \end{bmatrix}.$$

Для того, чтобы величины d_{ij} имели смысл расстояний между объектами в многомерном пространстве, необходимо, чтобы для всех $i, j = 1 \dots N$ выполнялись требования:

1. Максимальное сходство объекта с самим собой: $d_{ii} = \min_{j=1..N} d_{ij}$.
2. Требование симметрии: $d_{ij} = d_{ji}$.
3. Выполнение неравенства треугольника: $d_{ij} \le d_{ik} + d_{kj}$.

Если введенная мера удаленности между объектами такова, что выполняется это условия, то будем называть *D* матрицей расстояний.

✂ Данные могут быть исходно заданы в виде матрицы связи или удаленностей. Тогда возникает задача по заданной матрице восстановить в каком-либо смысле исходное множество точек данных таким образом, чтобы для него матрица связанности

или удаленностей имели заданный вид (точно или приближенно). ✂

Правило вычисления расстояния между объектами может сильно видоизменяться в зависимости от специфики задачи. Если такое правило задано, то говорят, что в пространстве признаков введена метрика. Рассмотрим различные правила измерения расстояний:

I. *Квадратичные метрики*:

Для класса квадратичных метрик квадрат расстояния между объектами является квадратичной формой от разностей значений их координат:

$$d_{ij} = \sqrt{(X_i - X_j)^T G (X_i - X_j)},$$

где G – симметричная положительно определенная матрица.

В качестве матрицы G размерами $m \times m$ можно выбрать

а) единичную матрицу $G = E$, в результате чего получаем обычное евклидово расстояние

$$d_{ij} = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2};$$

б) диагональную матрицу $G = \mathrm{diag}(g_1, g_2 \ldots g_m)$, в результате получим взвешенную евлидову метрику

$$d_{ij} = \sqrt{\sum_{k=1}^{m} g_k (x_{ik} - x_{jk})^2};$$

в) матрицу, обратную *ковариационной матрице* $G = S^{-1}$:

$$S = \begin{bmatrix} s_{11} & \ldots & s_{1N} \\ \ldots & s_{ij} & \ldots \\ s_{N1} & \ldots & s_{NN} \end{bmatrix}, \text{ где } s_{ij} = \frac{1}{N-1} \sum_{k=1}^{N} (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j),$$

что дает *махаланобисову метрику*.

С ковариационной матрицей связано понятие *эллипсоида рассеяния* облака точек. Осями эллипсоида рассеяния являются направления собственных векторов S (поскольку S-симметричная матрица, то собственные вектора образуют полную ортогональную систему векторов), длины осей выбираются равными значениям соответствующих собственных чисел.

Особенностью махаланобисовой метрики является то, что в ней эллипсоид рассеяния точек данных является шаром с единичным радиусом.

Основным преимуществом использования квадратичных метрик является тот факт, что производная от квадрата расстояния, измеренного в такой метрике является линейной функцией от координат объектов, что может быть использовано при решении различных задач (например, задач оптимизации).

II. *Специальные виды метрики*:

а) *городская метрика* или расстояние Минковского:

$$d_{ij} = \sum_{k=1}^{m} I_k(X_i, X_j),$$

применяется для измерения расстояний между объектами, описываемыми в порядковой шкале; $I_k(X_i,X_j)$ – разница в номерах градаций по k-ому признаку у сравниваемых объектов с векторами X_i и X_j;

б) Расстояние Хэмминга

$$d_{ij} = \sum_{k=1}^{m} \left| x_{ik} - x_{jk} \right|$$

чаще всего применяется для измерения расстояния между объектами, описываемыми в дихотомической шкале. Тогда расстояние Хэмминга – число несовпадающих значений признаков в рассматриваемом i-ом и j-ом объектах.

в) расстояние Колмогорова

$$d_{ij} = \left(\sum_{k=1}^{m} \left| x_{ik} - x_{jk} \right|^{p} \right)^{1/p}$$

является обобщением евклидовой метрики. Так, при $p = 1$ получаем метрику Хэмминга, при $p = 2$ – евклидову метрику, при $p = \infty$ – метрику «по максимуму модуля»:

$$d_{ij} = \max_{\{k=1..m\}} \left| x_{ik} - x_{jk} \right|$$

Все эти метрики допускают тривиальное обобщение, если производить суммирование с весами и тогда получаем взвешенную городскую, взвешенную Хэммингову и взвешенную метрику Колмогорова. Веса признаков подбираются или с помощью простых эвристических методов, или настраиваются с помощью специальных процедур (см., например, [4]).

III. *Риманова метрика*

Риманова метрика является обобщением квадратичной метрики в случае точечного пространства. В ней задается квадратичное расстояние между бесконечно близкими точками

$$ds = \sqrt{\sum_{i,j} g_{ij} d\xi_i d\xi_j},$$

где $d\xi_i$, $d\xi_j$ – бесконечно малые приращения координат, g_{ij} – метрический тензор. Значения g_{ij} зависят от точки пространства, в которой измеряется расстояние между бесконечно близкими точками. Тогда расстояние

между объектами измеряется с помощью интеграла по траектории и, вообще говоря, зависит от выбора траектории, соединяющей два объекта

$$d_{ij}(L) = \oint_L ds .$$

Если мы предполагаем, что расстояния измеряются по кратчайшему пути, то среди всех траекторий L выбирается та, при использовании которой расстояние оказывается наименьшим

$$d_{ij} = \inf_L d_{ij}(L) .$$

Такая траектория называется *геодезической* и играет роль, аналогичную прямой в евклидовом пространстве.

Приведем несколько простых видов римановых метрик

а) $g_{ij} = diag(f_1(\xi_1), f_2(\xi_2),..., f_m(\xi_m))$, где $f_1(\xi)$, $f_2(\xi)$, $f_3(\xi)$ – монотонные функции одного аргумента. По существу такая метрика может быть преобразована в евклидову нелинейным преобразованием координат $\xi_i' = f_i^{-1}(\xi_i)$;

б) конформно-плоская метрика

$$g_{ij} = \gamma(\xi_1, \xi_2 ... \xi_m)\delta_{ij},$$

где δ_{ij} – символ Кронеккера.

В общем случае эта метрика не может быть превращена в евклидову сразу во всем пространстве никаким нелинейным преобразованием координат. Ее смысл состоит в том, что масштаб, с помощью которого измеряются расстояния, меняется от точки к точке пространства.

2.1.5. Настройка метрики

При использовании взвешенных метрик остаются неопределенными веса признаков. Иногда условия задачи позволяют выделить те признаки, которые являются «более значимыми» при измерении расстояний и назначить для этих признаков значения весов. Если никаких дополнительных соображений нет, то для настройки весов могут быть использованы некоторые специальные приемы. Приведем примеры:

1) Адаптивные квадратичные метрики:

Рассмотрим метрику

$$d_{ij} = \sqrt{(X_i - X_j)^T G (X_i - X_j)},$$

d_{ij} – расстояние между i-ым и j-ым объектом.

Положим, что det G = 1. Такой выбор не делает результаты менее общими, но его использование позволяет не рассматривать решения, отличающиеся друг от друга только преобразованием гомотетии (равномерным растяжением по всем осям). Допустим, что на наборе объектов уже суще-

ствует определенная система отношений – объекты разбиты на k непересекающихся классов. Введем матрицу внутриклассового разброса W:

$$W = \frac{1}{N}\sum_{i=1}^{k}\sum_{X_l, X_m \in K_i}(X_l - X_m)(X_l - X_m)^T,$$

где T – обозначение операции транспонирования, K_i – обозначение множества объектов, принадлежащих i-ому классу.

Если выбрать $G = \alpha W^{-1}$, α – числовой множитель, то минимальной (среди всех квадратичных метрик) оказывается величина внутриклассового разброса

$$w = \sum_{i=1}^{k}\sum_{X_l, X_m \in K_i} d^2(X_l, X_m),$$

где $d^2(X_l, X_m)$ – квадрат расстояния между l-ым и m-ым объектом,

что приводит к тому, что классы оказываются максимально компактными [4].

Если разбиение на классы не заданно изначально, то возможна такая настройка метрики, при которой данные будут разбиты на k кластеров «наиболее контрастно». Итерационный алгоритм состоит из двух фаз:

Фаза 1. При фиксированной метрике $G^{(t)}$ производится разбиение множества данных на k кластеров тем или иным способом (см, например, [1,30]). Число кластеров задается в начале работы и далее не меняется.

Фаза 2. По полученной классификации строится матрица внутриклассового разброса W и вводится метрика $G^{(t+1)} = (W^{(t+1)})^{-1}$.

Шаги алгоритма повторяются до тех пор, пока относительные изменения значений элементов G не станут меньше заданного числа ε.

При использовании взвешенной евклидовой метрики вычисление весов может быть упрощено. В качестве их значений можно выбрать $g_k = \alpha w_{kk}^{-1}$, где w_{kk} – k-ый диагональный элемент матрицы W, α – нормирующий множитель (например, при $\alpha = \prod_k w_{kk}$ получаем равенство $\det G^{(t)} = 1$).

2) Использование частично-обучающих выборок

Частично обучающая выборка (ЧОВ) – множество пар объектов, относительно которых известно, что они принадлежат одному классу [4]. Естественно считать такие объекты близкими.

Введем аналог матрицы внутриклассового разброса для ЧОВ:

$$W = \frac{1}{n_{чов}} \sum_{i=1}^{n_{чов}} (X_{1i} - X_{2i})(X_{1i} - X_{2i})^T$$, (X_{1i}, X_{2i} – i-ая пара из ЧОВ).

Если W невырождена, то оптимальную метрику получим, выбирая $G = \alpha W^{-1}$, где α – нормирующий множитель.

Задачу можно упростить, используя взвешенную евклидову метрику и выбирая

$$G = \alpha \cdot diag(w_{11}^{-1}, w_{22}^{-1}, ..., w_{mm}^{-1}), \quad \alpha = \prod_i w_{ii}.$$

3) Максимизация корреляций.

Рассмотрим риманову метрику. Можно попытаться подобрать вид функций $f_i(x)$ таким образом, чтобы признаки оказались максимально скоррелированы. В качестве критерия, по которому ищется преобразование, можно использовать величину

$$Q^2 = \sum_{i<j}^{m} r_{ij}^2$$, r_{ij} – коэффициент корреляции между i-ым и j-ым признаком.

Функции могут быть выбраны из некоторого семейства монотонных преобразований, например, преобразования Бокса-Кокса [4]:

$$\begin{cases} f_i(x) = (x^{\alpha_i} - 1)/\alpha_i, & \alpha_i \neq 0 \\ f_i(x) = \ln x, & \alpha_i = 0 \end{cases}$$

или более общее двухпараметрическое семейство

$$\begin{cases} f_i(x) = (x^{\alpha_i} - \beta_i)/\alpha_i, & \alpha_i \neq 0 \\ f_i(x) = \ln(x - \beta_i), & \alpha_i = 0 \end{cases}.$$

Тогда $Q = Q(\alpha_1, \beta_1, \alpha_2, \beta_2, ..., \alpha_m, \beta_m)$ и критерий Q может быть максимизирован по параметрам α_k, β_k, k=1…m.

4) Контрастирование структуры данных

Рассмотрим такой вариант римановой метрики, в котором плотность данных оказалась бы почти равномерной или, наоборот, структура сгущений оказалась бы более контрастной.

Элемент объема риманова пространства вычисляется по формуле $dV = \sqrt{|g|}dx^1 dx^2 \ldots dx^n$, где $|g|$ – определитель матрицы метрического тензора.

Оценим нормированную на одну точку плотность распределения данных в исходном пространстве с помощью какой-либо непараметрической оценки. Например, пусть

$$\rho(x) = \frac{1}{|X|\sigma^n} \sum_{i=1}^{N} \prod_{j=1}^{n} K\left(\frac{x_j - x_j^i}{\sigma}\right),$$

где $K(x)$ – некоторая функция, удовлетворяющая условию $\int_{-\infty}^{\infty} K(x)dx = 1$, например $K(x) = \frac{1}{\sqrt{\pi}} \exp(-x^2)$, σ - радиус «области влияния», который является свободным параметром или может быть оценен методами непараметрической статистики.

Допустим, в новом пространстве плотность распределения будет постоянна: $\rho = \rho_0$. Например, можно выбрать $\rho_0 = 1/V$, где V – исходный объем фазового пространства данных (который, очевидно, конечен).

Положим, что метрика пространства имеет конформно-плоский вид:

$$ds^2 = \gamma(x) \sum_{ij} \delta_{ij} dx^i dx^j,$$

где $\gamma(x)$ - конформный множитель, зависящий от точки ***x***. Для того, чтобы плотность данных в новом пространстве стала постоянной (может быть за исключением граничных областей), необходимо выбрать

$$\gamma(x) = \left(\frac{\rho_0}{\rho(x)}\right)^{2/n}.$$

С другой стороны, выбирая

$$\gamma(x) = \left(\frac{\rho(x)}{\rho_0}\right)^{\alpha}, \alpha > 0$$

получаем метрику, в которой сгущения данных выглядят тем более контрастно, чем больше параметр α. Малые расстояния между точками данных в такой метрике становятся еще меньше, большие – больше.

2.1.6. Вычисление расстояний для данных с пробелами

Отдельные значения координат точек данных могут оказаться либо недостоверными, либо вообще неизвестными, и тогда объект в многомерном пространстве надо представлять не точкой, а прямой или гиперплоскостью, параллельной координатным осям. Как измерять расстояния между объектами в таком случае?

Будем вычислять такие расстояния как расстояния между соответствующими геометрическими образами (между точкой и прямой, точкой и плоскостью, прямой и прямой и т.д.) В результате получим очень простое правило вычисления расстояний между объектами с пропущенными значениями признаков – *расстояния вычисляются в том подпространстве, в котором значения координат у объектов известны полностью*. Иными словами, при подсчете сумм в формулах вычисления расстояний те слагаемые, которые не могут быть вычислены из-за того, что отдельные значения признаков неизвестны, просто пропускаются.

Покажем, что это так. Допустим, объект имеет следующие значения признаков

$X = (\xi_1, \xi_2, .., \xi_{k-1}, @, \xi_{k+1}, .., \xi_m)$.

Значком @ мы обозначили неизвестное значение признака ξ_k. Пусть $X(k)$ обозначает, что у объекта X неизвестно значение k-ого признака, а $X^0(k)$ обозначает следующий набор признаков:

$X^0(k) = (\xi_1, \xi_2, .., \xi_{k-1}, 0, \xi_{k+1}, .., \xi_m)$,

то есть k-ое значение признака заменено нулем.

Тогда геометрический образ, который можно сопоставить объекту $X(k)$ – прямая

$X = X^0(k) + e_k t$, где e_k – единичный орт k-ой координатной оси.

Пусть значения признаков объекта Y известны полностью. Тогда найдем кратчайшее расстояние между X и $Y = (\eta_1, \eta_2 ... \eta_m)$:

$$\frac{d}{dt}\left((X-Y)^2\right) = 0 \Rightarrow (X-Y)e_k = 0 \Rightarrow t = Ye_k = \eta_k .$$

Это означает, что при вычислении расстояния неизвестное значение k-го признака у X необходимо заменить значением k-го признака Y. Но тогда

$$(X-Y)^2 = (X^0(k) - Y)^2 - \eta_k = (X^0(k) - Y^0(k))^2,$$

то есть при вычислении расстояний можно просто приравнять к нулю значение k-го признака объектов X и Y. Тогда, например, в случае евклидова расстояния получаем формулу для вычисления расстояния

$$d(X,Y) = \sqrt{\sum_{\substack{i=1 \\ \xi_i \neq @}}^{m} (\xi_i - \eta_i)^2} .$$

Рассмотрим, что будет, если у объекта Y неизвестно значение l-ого признака. Тогда

$Y = Y^0(l) + e_l s$,

$$\begin{cases} \frac{d}{dt}\left((X-Y)^2\right) = 0 \\ \frac{d}{ds}\left((X-Y)^2\right) = 0 \end{cases} \Rightarrow \begin{cases} (X-Y)e_k = 0 \\ (X-Y)e_l = 0 \end{cases} \Rightarrow \begin{cases} -Y^0(l)e_k + t - \delta_{kl} s = 0 \\ X^0(k)e_l + \delta_{kl} t - s = 0 \end{cases} .$$

Если $k \neq l$, то $t = \eta_k$, $s = \xi_l$ и приходим к той же ситуации, что и в случае с точкой и прямой. Если $k = l$, то $t = s$ и $(X - Y)^2 = \left(X^0(k) - Y^0(k)\right)^2$ и вновь мы просто пропускаем неизвестное значение признака.

Совершенно аналогично обстоит дело в общем случае, когда и *X*, и *Y* содержат произвольное число пропущенных признаков.

Теперь вернемся к случаю, когда у *Y* известны все значения признаков, а для *X* неизвестно значение *k*-го признака, но известно, что оно лежит в диапазоне $[a_k,b_k]$ (объект представляется отрезком прямой). Тогда правило вычисления расстояния между *X* и *Y* окажется следующим: если $\eta_k \in [a_k, b_k]$, то как и прежде пропускаем значение признака, иначе, если $\eta_k < a_k$, то полагаем $\xi_k = a_k$, а если $\eta_k > a_k$, то $\xi_k = b_k$ и считаем расстояние.

2.1.6. Гравитирующие данные

Процедуры предобработки данных могут заключаться не только в нормировке данных и выборе метрики, но и в целенаправленном изменении расстояний между точками для подчеркивания определенных особенностей структуры облака точек.

Рассмотрим предложенный в главе I вариант преобразования данных как облака гравитирующих точек. Припишем каждой точке X_i «массу» m_i. «Типичные представители» классов, например, могут иметь большую массу по сравнению с другими. В самом простом случае все массы равны.

Будем использовать евклидову метрику для измерения расстояний между точками, тогда

$$\left\|X_i - X_j\right\| = \sqrt{(X_i - X_j)^2}$$

Введем потенциал взаимодействия между точками X_i и X_j. Рассмотрим простой случай центральных сил и предположим, что потенциал не зависит от номеров точек *i* и *j*:

$$\varphi(X_i, X_j) = \varphi\left(\left\|X_i - X_j\right\|\right),$$

$\varphi(r)$ – потенциальная функция, зависящая только от расстояния между точками. Тогда энергия взаимодействия пары точек

$$U_{ij} = m_i m_j \varphi\left(\left\|X_i - X_j\right\|\right)$$

и суммарная энергия

$$U = \frac{1}{2}\sum_{i \neq j} m_i m_j \varphi\left(\left\|X_i - X_j\right\|\right).$$

Тогда

$$\frac{\partial U}{\partial X_k} = \frac{1}{2}\sum_{i\neq j} m_i m_j \frac{\varphi'\left(\left\|X_i - X_j\right\|\right)}{\left\|X_i - X_j\right\|}\left(\delta_{ik}(X_i - X_j) - \delta_{jk}(X_i - X_j)\right) =$$

$$= m_k \sum_{i\neq k} m_i \frac{\varphi'\left(\left\|X_i - X_j\right\|\right)}{\left\|X_i - X_j\right\|}(X_k - X_i)$$

.

Будем считать, что легкая частица движется в вязкой среде, так что инерция движения гасится средой. Тогда уравнение движения оказывается первого порядка:

$$\frac{\partial X_i}{\partial t} = -\frac{\partial U}{\partial X_k}.$$

Проще всего решать это уравнение по схеме Эйлера, что дает следующую итерационную формулу:

$$X_i^{(t+1)} = X_i^{(t)} - \frac{\partial U}{\partial X_i}\Delta t$$

.

Шаг Δt должен быть достаточно мал, чтобы обеспечить адекватность решения.

Теперь вспомним, что наша Вселенная данных должна раздуваться, чтобы обеспечить постоянную плотность данных. Этого эффекта можно достигнуть простой перенормировкой данных на первоначальный объем. Для этого нужно вычислить новый фазовый объем данных $V^{(t+1)}$ – объем прямоугольного параллелепипеда со сторонами, равными диапазонам значений признаков. Тогда получаем окончательный итерационный алгоритм:

Шаг 1. Движение частиц

$$X_i^{(t+1)} = X_i^{(t)} - \frac{\partial U}{\partial X_i}\Delta t$$

Шаг 2. Перенормировка.

На этом шаге можно предложить два варианта:

1) изотропный (расширение происходит одинаково во всех направлениях):

$$\left(X_i^{(t+1)}\right)' = \left(\frac{V^{(0)}}{V^{(t+1)}}\right)^{1/m} \cdot X_i^{(t+1)}$$

, где $V^{(0)}$ – начальный объем данных.

В этом варианте возможно «схлапывание» Вселенной данных вдоль какого-то направления (параллелепипед может начать неограниченно вытягиваться).

2) неизотропный (расширение происходит так, чтобы сохранять форму исходного фазового объема)

$$\left(x_{ik}^{(t+1)}\right)' = \frac{\Delta_k^{(0)}}{\Delta_k^{(t+1)}} x_{ik}^{(t+1)}$$

,

где $\Delta_k^{(t)}$ – диапазон изменения значений k-го признака на шаге t.

Теперь рассмотрим некоторые варианты потенциальных функции $\varphi(r)$:

1) ньютоновский потенциал без сингулярности

$$\varphi(r) = \frac{\alpha}{r + \varepsilon^2}$$

Если $\alpha < 0$, то данные притягиваются и их кластерная структура становится более контрастной, при $\alpha > 0$ – отталкиваются и распределение становится более равномерным (но отношения соседства при этом нарушаются не слишком сильно) . Регуляризующая постоянная ε^2 нужна для того, чтобы точки данных не испытывали слишком сильных перемещений (за времена порядка Δt) вблизи $r = 0$.

2) потенциал ядерных сил

$$\varphi(r) = \frac{\alpha \exp(-r / r_0)}{r + \varepsilon^2} .$$

Смысл постоянной r_0 – эффективный радиус взаимодействия: на расстояниях больше r_0 точки данных практически «не замечают» друг друга.

3) использование других степенных зависимостей

$$\varphi(r) = \frac{\alpha}{r^{\beta} + \varepsilon^2} .$$

Так, для многомерного ньютоновского потенциала $\beta = \frac{m-1}{2}$.

2.1.8. Локальные статистики

Рассмотрим вкратце идею локальных статистик, которая тоже может быть использована при обработке данных перед применением процедур визуализации.

Выберем k-ый объект, который будет играть роль базового. Перейдем к новой векторной переменной $r' = r - r_k$, или, что тоже самое – отцентрируем данные так, чтобы k-ый объект оказался в начале координат:

$X'_p = X_p - X_k$, $p = 1 \ldots N$.

Теперь введем квадратичную метрику

$$d_k^2(X'_p, X'_j) = (X'_p - X'_j) G_k (X'_p - X'_j)^T ,$$

G_k – положительно определенная симметричная матрица. Ее коэффициенты $g_{ij}^{(k)}$ могут быть настроены из различных соображений. Приведем два самых простых:

1) Оптимизация разделения классов. Коэффициенты $g_{ij}^{(k)}$ настраиваются из условия минимума функционала

$$J = \frac{\sum\limits_{X'_i \in w_k} d_k(0, X'_i)}{\sum\limits_{X'_i \notin w_k} d_k(0, X'_i)} \to \min$$

, где w_k – класс, к которому изначально принадлежал объект X_k.

Таким образом, объекты того же класса, что и X_k, оказываются в результате ближе к началу координат (где находится X_k), чем все остальные объекты.

2) Нормализация распределения

Матрица $G_k = W_k^{-1}$, где

$$W_k = \frac{1}{N}\sum_{i=1}^{m} X'_i (X'_i)^T$$

.

Если матрица W_k невырождена, то G_k существует и в результате распределение векторов X'_i «с точки зрения» X'_k будет выглядеть похожим на нормальное.

Для решения упомянутых задач можно упростить G_k, выбрав ее диагональной (метрика окажется взвешенной евклидовой):

$$d_k^2(X'_p, X'_j) = \sum_i w_{ki}(x'_{pi} - x'_{ji})^2$$

, $w_{kp} \geq 0$.

Тогда, например, для нормализации распределения можно выбрать $w_{ki} = (W_k)_{ii}^{-1}$, где $(W_k)_{ii}$ – i-ый диагональный элемент W_k.

Построенное пространство признаков можно анализировать (и визуализировать) любыми методами анализа многомерных данных. В результате мы получим описание набора данных «с точки зрения» объекта X_k. Строя N таких пространств (и получая N описаний) и сравнивая их между собой, можно сделать полезные выводы об структуре данных. Разумеется, для эффективного сравнения N результатов обработки необходимо каким-то образом автоматизировать этот процесс, выдавая уже «сухой остаток» результатов сравнения.

Другой способ состоит в конструировании нового пространства, которое определенным образом «обобщает» все построенные пространства. Допустим, построены N метрик $d_k(X'_i, X'_j)$, $k = 1 \ldots N$. Тогда можно записать такую матрицу удаленностей:

$$D = \begin{Bmatrix} d_1(X_1, X_1) & \ldots & d_1(X_1, X_N) \\ \ldots & & \ldots \\ d_k(X_k, X_1) & \ldots & d_k(X_k, X_N) \\ \ldots & & \ldots \\ d_N(X_N, X_1) & \ldots & d_N(X_N, X_N) \end{Bmatrix},$$

где на пересечении *i*-ой строки и *j*-го столбца стоит расстояние между *i*-ым и *j*-ым объектом в метрике, построенной для *i*-го объекта. Поскольку для различных объектов будут построены различные метрики, то для элементов матрицы *D* могут не выполняться условия симметричности ($d_i(X_i, X_j) \neq d_j(X_j, X_i)$) и неравенства треугольника ($d_i(X_i, X_j)$ может быть больше $d_i(X_i, X_k) + d_k(X_k, X_j)$). Поэтому такая матрица не может напрямую служить матрицей расстояний.

Попробуем устранить указанные нарушения, вводя новый класс так называемых $d^{(s)}$-метрик:

$$d^{(s)}(X_i, X_j) = a \cdot s[\varphi(d_{ik}), \varphi(d_{jk})] + b, \; k = 1 \ldots N,$$

где d_{ik}, d_{jk} – элементы *i*-ой и *j*-ой строк матрицы *D*; $\varphi(x)$ – монотонная функция, например, $\varphi(x) = x$ или $\varphi(x) = \mathrm{rank}(x)$ – преобразование к порядковой шкале; s[…,…] – мера подобия двух последовательностей; *a*, *b* – константы, значения которых подбираются с целью масштабирования и выполнения метрической аксиомы неравенства треугольника. Тогда расстояние между объектами в $d^{(s)}$ метрике имеет ясный смысл – это различие двух последовательностей чисел – всех расстояний до объекта X_i в метрике d_i и всех расстояний до объекта X_j в метрике d_j. Другими словами – это мера сходства двух представлений данных – с точки зрения объекта X_i и X_j.

В качестве конкретных вариантов $d^{(s)}$-метрик могут быть использованы:

$d^{(d)}$-метрика:

$$d^{(d)}(X_i, X_j) = \sum_{k=1}^{N} \left[\varphi(d_{ik}) - \varphi(d_{jk})\right]^2$$

основана на простой евклидовой формуле для сравнения последовательностей. Она автоматически обеспечивает выполнение условия симметричности и неравенства треугольника, однако нивелирует некоторые важные особенности рядов $\varphi(d_{ik})$ и $\varphi(d_{jk})$.

$d^{(s)}$-метрика, основанная на стандартных мерах связи:

$$d^{(s)}(X_i, X_j) = \frac{1 - s_{ij}}{2},$$

где в качестве s_{ij}, можно выбрать коэффициент корреляции Пирсона (что обеспечивает условие симметрии, но возможны незначительные нарушения неравенства треугольника) или коэффициент связи τ-Кендалла между ранговыми признаками (обеспечивает выполнение всех условий).

После того, как исследователь остановится на том или ином варианте $d^{(s)}$ метрики, матрица *D* преобразуется таким образом, чтобы она могла служить матрицей расстояний для некоторого распределения точек данных. После этого можно применять те методы анализа данных, в которых

исходной информацией служит матрица расстояний (например, методы метрического шкалирования).

2.2. Линейный анализ данных

Теперь перейдем к краткому описанию традиционных линейных методов анализа, которые так или иначе можно использовать для визуализации структуры данных.

Сначала выпишем формулы, которыми описывается случайная величина (вектор), подчиненная многомерному нормальному закону распределения. Плотность вероятности в этом случае равна

$$\rho(X) = C\exp\left(-\frac{1}{2}(X-MX)^T\Sigma^{-1}(X-MX)\right),\quad C=\frac{1}{(2\pi)^{m/2}\sqrt{\det\Sigma}}.$$

Здесь MX – математическое ожидание X, Σ – ковариационная матрица

$$\Sigma = M((X-MX)(X-MX)^T).$$

Величина MX и ковариационная матрица оцениваются с помощью данных выборки:

$$MX \approx \frac{1}{N}\sum_i X_i,\quad \Sigma \approx S = \frac{1}{N-1}\sum_{i=1}^{N}(X_i-MX)(X_i-MX)^T.$$

2.2.1. Метод главных компонент

Как уже упоминалось, цель анализа данных – извлечение содержащихся в них информации. Задача снижения размерности набора данных – описание каждой точки данных с помощью величин, число которых меньше размерности пространства и которые являются функциями исходных координат

$$\eta_k = F_k(\xi_1,\xi_2,\ldots,\xi_m),\ k = 1\ldots m',\ m' < m.$$

Функции F_k задают отображение F из исходного пространства R^m в пространство $R^{m'}$. Это отображение должно выбираться таким образом, чтобы на наборе данных X максимизировать определенный критерий, както отражающий количество сохраняемой при этом преобразовании информации. Выбирая отображение F из определенного класса отображений и критерий сохранения информации J, можно получать различные методы сокращения размерности пространства признаков.

В методе главных компонент F – некоторое линейное ортогональное нормированное отображение, т.е.

$F_k(\xi_1, \xi_2, ..., \xi_m) = c_{1k}(\xi_1 - \mu_1) + ... + c_{mk}(\xi_m - \mu_m)$, где $\mu_j = \frac{1}{N}\sum_{i=1}^{N} x_{ij}$ – средние по набору данных значения признаков, а на коэффициенты c_{ij} накладываются условия

$\sum_{k=1}^{m} c_{ik}^2 = 1$, $\sum_{k=1}^{m} c_{ik} c_{jk} = 0$, $i, j = 1 \ldots m$, $i \neq j$.

Вид критерия J:

$$J = \frac{D\eta_1 + ... + D\eta_{m'}}{D\xi_1 + ... + D\xi_m},$$

где D – вычисление дисперсии случайной величины.

Согласно этому критерию, количество сохраненной информации равно доле «объясненной» с помощью новых признаков $\eta_1 \ldots \eta_m$ дисперсии исходных признаков.

Первой главной компонентой называют такую нормированно-центрированную линейную комбинацию исходных признаков, которая среди всех прочих нормированно-центрированных линейных комбинаций обладает на данном наборе данных наибольшей дисперсией.

Решим задачу нахождения первой главной компоненты. Для этого необходимо решить задачу

$$D(l_1 X) \to \max_{l_1},$$

где l_1 – вектор-строка размерности m, при условии нормировки $l_1 l_1^T = 1$. Вектор l_1 можно представлять как единичный вектор пространства данных, тогда $(l_1, X_i) l_1$ – точка проекции вектора X_i на вектор l_1.

Положим, что система векторов данных является центрированной, т.е. $E(X) \equiv \overline{X} = 0$ Тогда

$$D(l_1 X) = E(l_1 X)^2 = E(l_1 X X^T l_1^T) = l_1 E(X X^T) l_1^T = l_1 S l_1^T,$$

где S – ковариационная матрица набора данных X.

Введем функцию Лангранжа $\varphi(l_1, \lambda) = l_1 S l_1^T - \lambda(l_1 l_1^T - 1)$, тогда

$$\frac{\partial \varphi}{\partial l_1^T} = 2 S l_1^T - 2\lambda l_1^T = 0,$$

и

$$(S - \lambda I) l_1^T = 0,$$

то есть l_1^T – собственный вектор ковариационной матрицы. Но $D(l_1 X) = l_1 S l_1^T = \lambda$, значит для того, чтобы $D(l_1 X)$ достигало максимума, нужно выбрать максимальное собственное значение. Нормированный соб-

ственный вектор, отвечающий этому значению и задаст направление первой главной компоненты в пространстве.

k-ой главной компонентой ($k = 2…m$) называется такая нормированно-центрированная линейная комбинация исходных признаков, которая не коррелирована с k-1 предыдущими главными компонентами и среди всех прочих нормированно-центрированных линейных комбинаций, не коррелированных с предыдущими k-1 главными компонентами обладает на данном наборе данных наибольшей дисперсией.

Можно показать, что k-ая главная компонента задается собственным вектором ковариационной матрицы данных, который соответствует k-ому по величине собственному значению.

Заметим, что решение задачи нахождения главных компонент не является инвариантным относительно смены масштабов у разных признаков. Поэтому перед применением метода данные нормируются так, чтобы все признаки были измерены в сопоставимых масштабах.

На m' главных компонент можно натянуть подпространство размерности m'. Легко понять, что сумма квадратов расстояний от точек данных до этого подпространства равна умноженной на N (число точек) остаточной дисперсии, «не объясненной» с помощью m' главных компонент, то есть $N(D\xi_{m'+1} + ... + D\xi_m) = N(\lambda_{m'+1} + ... + \lambda_m)$, где $\lambda_{m'+1},…,\lambda_m$ – наименьшие по величине собственные значения. Отсюда становится понятным важное экстремальное свойство указанного подпространства:

Свойство 1. Сумма квадратов расстояний от исходных точек-наблюдений $X_1,…,X_N$ до пространства, натянутого на m' главных компонент наименьшая среди всех других подпространств размерности m', полученных с помощью произвольной линейно-независимой системы из m' векторов.

Укажем еще два экстремальных свойства подпространства главных компонент.

Зададим правило перехода к меньшему числу переменных с помощью линейного преобразования:

$$z_{ij} = \sum_{k=1}^{m} c_{jk} x_{ik}$$, $j = 1…m'$, $i = 1…N$, или

$Z = CX$,

здесь x_{ik} – k-ая координата вектора данных X_i, z_{ij} – j-ая координата i-ой точки данных в некотором подпространстве меньшей размерности $R^{m'}$. Можно рассматривать эти формулы как проекцию точек данных из исходного пространства в $R^{m'}$.

Рассмотрим величины

$$M = \sum_{i=1}^{N} \sum_{j=1}^{N} (X_i - X_j)^2$$,

$$M(C) = \sum_{i=1}^{N} \sum_{j=1}^{N} (Z_i - Z_j)^2 .$$

Их смысл – сумма квадратов расстояний между всевозможными парами объектов в исходном пространстве и в $R^{m'}$. Введем в качестве меры искажения суммы квадратов попарных расстояний между точками данных величину M-$M(C)$. Можно показать [4], что

$$M - M(L) = \min_{C} \{M - M(C)\} = N^2 (\lambda_{m'+1} + ... + \lambda_m),$$

где L – матрица, задающая проекцию точек данных в подпространство, натянутое на m' главных компонент. Отсюда сдедует

Свойство 2. Среди всех подпространств размерности m' , полученных из исходного пространства данных с помощью произвольного линейного преобразования исходных координат, в подпространстве, натянутом на первые m' главных компонент наименее искажается сумма квадратов расстояний между всевозможными парами рассматриваемых точек.

Наконец, введем меру искажения расстояний до начала координат и углов между прямыми, соединяющими всевозможные пары точек с началом координат. Обозначим ее $\|H - H(C)\|$, где

$$H = \{h_{ij}\}, \quad h_{ij} = (X_i, X_j),$$

$$H(C) = \{h_{ij}(C)\}, \quad h_{ij}(C) = (Z_i, Z_j),$$

а под $\|A\|$ – евклидова норма матрицы A. Оказывается, что

$$\|H - H(L)\| = \min_{C} \|H - H(C)\| = N^2 (\lambda_{m'+1}^2 + ... + \lambda_m^2) .$$

То есть справедливо

Свойство 3. Среди всех подпространств размерности m' , полученных из исходного пространства данных с помощью произвольного линейного преобразования исходных координат, в подпространстве, натянутом на первые m' главных компонент наименее искажаются расстояния от точек до начала координат, а также углы между прямыми, соединяющими всевозможные пары точек с началом координат.

Наконец, отметим, что указанное вначале требование «центрированности» данных не является принципиальным. Если в данных нет пробелов, то геометрический центр облака точек определен однозначно. Отличия в формулировках свойств 1-3 будет лишь в том, что вместо линейного подпространства, натянутого на главные компоненты надо рассматривать линейное многообразие, построенное на первых главных компонентах и проходящее через точку геометрического центра.

2.2.2. Итерационный алгоритм нахождения главных компонент

Используя экстремальное Свойство 1 подпространств, натянутых на главные компоненты, можно предложить итерационный алгоритм нахождения первой главной компоненты (ср. с [41,53]). Будем искать прямую в пространстве данных, заданную параметрическим уравнением

$$\mathbf{y} = \mathbf{a}t + \mathbf{b},$$

такую, что сумма квадратов расстояний от точек данных до этой прямой минимальна. Эта сумма, равная

$$Q = \sum_{i=1}^{N} (X_i - \mathbf{a}t_i - \mathbf{b})^2$$

является критерием, который можно минимизировать с помощью следующей простой процедуры:

Зададимся произвольными векторами **a** и **b**. Далее итерация алгоритма состоит из двух шагов:

Шаг 1. При заданных векторах **a** и **b** определяется набор $\{t_i\}$, $i = 1 \ldots N$:

$$\frac{\partial Q}{\partial t_i} = -2(X_i - \mathbf{a}t_i - \mathbf{b})\mathbf{a} = -2(X_i - \mathbf{b})\mathbf{a} - 2\mathbf{a}^2 t_i = 0,$$

$$t_i = \frac{(X_i - \mathbf{b})\mathbf{a}}{\mathbf{a}^2}.$$

Шаг 2. При заданном наборе $\{t_i\}$ определяются новые координаты векторов **a** и **b**:

$$\begin{cases} \dfrac{\partial Q}{\partial \mathbf{a}} = -2\sum\limits_{i=1}^{N} (X_i - \mathbf{a}t_i - \mathbf{b})t_i = 0 \\ \dfrac{\partial Q}{\partial \mathbf{b}} = -2\sum\limits_{i=1}^{N} (X_i - \mathbf{a}t_i - \mathbf{b}) = 0 \end{cases},$$

$$\begin{cases} \mathbf{a}\sum\limits_{i=1}^{N} t_i^2 + \mathbf{b}\sum\limits_{i=1}^{N} t_i = \sum\limits_{i=1}^{N} X_i t_i \\ \mathbf{a}\sum\limits_{i=1}^{N} t_i + \mathbf{b}N = \sum\limits_{i=1}^{N} X_i \end{cases},$$

что дает m систем линейных уравнений 2x2 для определения всех компонент векторов **a** и **b.**

Проверка на останов. Алгоритм останавливается, когда $\dfrac{\Delta Q}{Q} < \varepsilon$, где ΔQ – изменение величины Q за итерацию, а ε – малая величина.

Преимущество такого способа нахождения первой главной компоненты состоит в том, что он легко обобщается на случай, когда некоторые данные содержат неполные значения. Рецепт прост – если в соответствующей сумме встречается неизвестное значение, то такое слагаемое пропускается. Тогда, если неполных данных нет, то **b** дает вектор среднего

$\mathbf{b} = \frac{1}{N}\sum_{i=1}^{N} X_i$ значения всех координат: , иначе – некоторый «эффективный» вектор среднего. Вектор **a** в случае полных данных задает направление первой главной компоненты, в случае неполных – «эффективную» первую главную компоненту.

Для того, чтобы найти вторую главную компоненту, поступают следующим образом:

1. Рассчитывается множество векторов первых остатков X': $X'_i = X_i - \mathbf{a}t_i - \mathbf{b}$. Это множество лежит в пространстве, ортогональном первой главной компоненте, размерностью на единицу меньше размерности исходного пространства данных.

2. Для нового множества векторов рассчитывается первая главная компонента. Она и будет второй главной компонентой исходного набора данных.

Для нахождения третьей главной компоненты ищется множество вторых остатков и для него определяется первая главная компонента, и т.д.

2.2.3. Модели линейного факторного анализа

Напомним, что метод главных компонент может быть сформулирован как задача оптимизации функционала *J* качества «сохранения» информации при заданном отображении *F* из исходного пространства в пространство меньшей размерности. В методе главных компонент в качестве функционала *J* выступает доля «объясненной» с помощью новых координат дисперсии.

В модели факторного анализа каждому вектору данных X_i сопоставляется набор из *m'* значений факторов $y_i^1,\ldots,y_i^{m'}$:

$$x_{ij} - \mu_j = \sum_{k=1}^{m'} q_{jk} y_i^k + u^j \quad , j = 1\ldots m',\ \text{или}$$

$$X_i - \overline{X} = QY_i + U,$$

где μ_j – среднее значение *j*-го признака, x_{ij} – значение *j*-го признака для *i*-го объекта, q_{ij} – «нагрузки» факторов, u^j – остаточная случайная компонента. При этом выполняются условия:

$$Ey^k = 0,\ Eu^k = 0,\ Dy^k = 1,$$

и $y_i^1,\ldots,y_i^{m'}$, $u^1,\ldots,u^{m'}$ попарно некоррелированы.

В качестве *F* выбирается линейное преобразование координат такое, чтобы выполнялись эти условия, и достигал максимума функционал

$$J(F) = 1 - \left\| R_X - R_{\hat{X}} \right\|^2,$$

где R_X – корреляционная матрица исходных признаков, $R_{\hat{X}}$ – корреляционная матрица «проекций» в пространство факторов $\hat{X}_i = QY_i$, $\|A\|$ – евклидова норма матрицы *A*.

Поясним выкладки. Матрица нагрузок *Q* размерами *m'×m* осуществляет линейное отображение из пространства факторов (размерности *m'*) в исходное пространство (размерности *m*). В результате получается множество данных $\hat{X}_i = QY_i$, которое совпадает со множеством исходных данных с точностью до случайной компоненты *U*. Это множество лежит в некотором подпространстве размерности *m'*, натянутом на *m'* столбцов матрицы *Q*. Матрица *Q* и исходный набор факторов $y_i^1,\ldots,y_i^{m'}$, $i = 1\ldots N$ выбираются таким образом, чтобы корреляционная матрица набора данных $\hat{X}$ максимально точно воспроизводила корреляционную матрицу исходного набора данных. Таким образом, критерием количества «сохраненной» информации является здесь объяснение не дисперсии признаков, а их взаимной скоррелированности.

Что касается «шумовой» компоненты *U*, то обычно полагают, что она не зависит от распределения данных и подчинена *m*-мерному нормальному распределению с нулевым средним значением. Тогда ковариационная матрица распределения *U* имеет диагональный вид:

$V = E(UU^T)$, $V = diag(v_{11}...v_{mm})$, $v_{ii} = Du^i$.

Если исходное распределение данных предполагается центрированным, то его ковариационная матрица

$S = QQ^T + V$

Решением задачи факторного анализа называют пару матриц (*Q*,*V*), удовлетворяющую этому условию. Очевидно, что если одно такое решение существует, то одновременно решением является (*QC*,*V*), где *C* – произвольное ортогональное преобразование (поворот векторов-столбцов матрицы *Q*). По этой и другим причинам решение задачи факторного анализа является неоднозначным, поэтому для ее решения необходимо выбрать какие-либо дополнительные предположения о свойствах матрицы *Q*. Независимо от этих предположений итерационный метод решения задачи выглядит следующим образом.

Вначале задается нулевое приближение матрицы $V=V^{(0)}$.

Шаг 1. Получаем нулевое приближение матрицы $\Psi = QQ^T$, т.е. $\Psi^{(0)} = S - V^{(0)}$.

Шаг 2. С помощью $\Psi^{(0)}$ определяем нулевое приближение матрицы *Q*.

Алгоритм продолжается до получения необходимой точности.

Дополнительные предположения о структуре матрицы *Q* используются для реализации Шага 2 алгоритма. Рассмотрим два условия:

1) Q^TQ – диагональная матрица, причем диагональные элементы различны и упорядочены в порядке убывания.

Тогда

$$A = Q^T Q = diag(\lambda_1 ... \lambda_{m'}),\ \lambda_1 > \lambda_2 > .. > \lambda_m,$$

$$\Psi Q = QQ^T Q = AQ,$$

то есть m' столбцов $q_1 \ldots q_{m'}$ матрицы Q удовлетворяют уравнениям $\Psi q_i - \lambda_i q_i = 0$ на собственные значения матрицы Ψ. Собственные вектора матрицы Ψ, отвечающие первым m' по величине собственным значениям и составят очередное приближение матрицы нагрузок Q.

2) Q^TVQ – диагональная матрица, причем диагональные элементы различны и упорядочены в порядке убывания.

Тогда

$$A = Q^T VQ = diag(\lambda_1 ... \lambda_{m'}),\ \lambda_1 > \lambda_2 > .. > \lambda_m,$$

$$\Psi Q = V^{-1} QQ^T VQ = AV^{-1} Q,$$

то есть m' столбцов $q_1 \ldots q_{m'}$ матрицы Q удовлетворяют уравнениям $\Psi q_i - \lambda_i v_{ii}^{-1} q_i = 0$ на обобщенные собственные значения матрицы Ψ. Обобщенные собственные вектора матрицы Ψ, отвечающие первым m' по величине обобщенным собственным значениям составляют очередное приближение матрицы нагрузок Q.

Подведем некоторые итоги. По данному набору данных, используя метод главных компонент или методы факторного анализа, можно построить линейные модели данных. Фактически эти методы строят специальное линейное многообразие меньшей размерности, на которое проецируются исходные данные. Это подпространство оказывается в некотором смысле оптимальным среди всех других линейных многообразий той же размерности. В случае метода главных компонент оптимальность заключается в том, что проекции данных максимально воспроизводят дисперсию исходных данных. В случае методов факторного анализа значения признаков проекций максимально похожи на исходные значения признаков в смысле взаимной корреляции. Следует заметить, что в случае, когда остаточные дисперсии (суммарные расстояния до построенного подпространства) невелики, оба метода дают сходные результаты (это становится особенно понятно, если рассмотреть условие 1 на структуру матрицы Q).