Конспект лекции
«Уменьшение размерности описания данных: метод главных компонент»
по курсу «Математические основы теории прогнозирования» 2011

Проблема анализа многомерных данных

При решении различных задач распознавания предполагается, что в наличии имеется некоторая выборка объектов, и для каждого объекта вычислен один и тот же набор признаков. На практике объекты могут быть представлены сложными многомерными данными, например, изображениями, набором кривых, текстом, ДНК-микрочипами и т.д. (см. рис. 1). Поэтому возникает задача извлечения из входных многомерных данных набора признаков, информативных с точки зрения дальнейшего решения задачи распознавания.

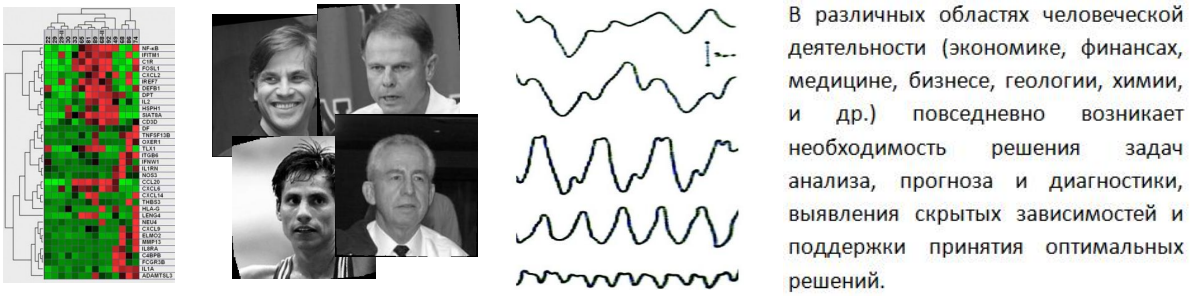


Рис. 1: Примеры многомерных данных.

Любые многомерные данные всегда можно представить в виде вектора чисел. В случае изображений достаточно развернуть матрицу пикселей в вектор. Для текстов можно вычислить количество раз, которое встречается каждое слово в тексте, и сформировать вектор чисел, длина которого определяется общим числом слов в словаре. Подобные вектора чисел имеют, как правило, большую длину, а содержащиеся в них признаки, как правило, малоинформативны. Поэтому рассматривается задача сокращения размерности описания данных с целью получения относительно компактного множества информативных признаков.

Рассмотрим задачу классификации изображений рукописных цифр MNIST[1]. Здесь имеется некоторое количество черно-белых изображений, на каждом из которых представлена одна цифра (см. рис. 2). Задача состоит в автоматическом определении цифры для входного изображения (задача классификации на 10 классов).

Для того, чтобы применить методы распознавания в данной задаче, необходимо предварительно выбрать пространство признаков, характеризующее изображения цифр. В простейшей случае в качестве признаков можно взять исходные интенсивности пикселов изображения. Тогда для изображения размера 28×28 получаем 784 признака. Такой способ

[1]Исходные данные можно скачать по адресу http://yann.lecun.com/exdb/mnist/

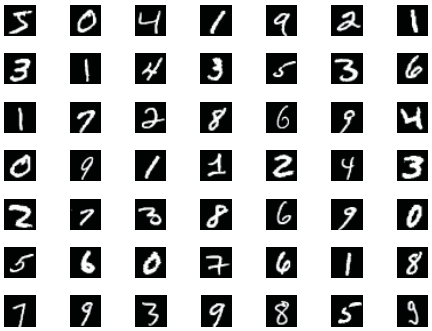
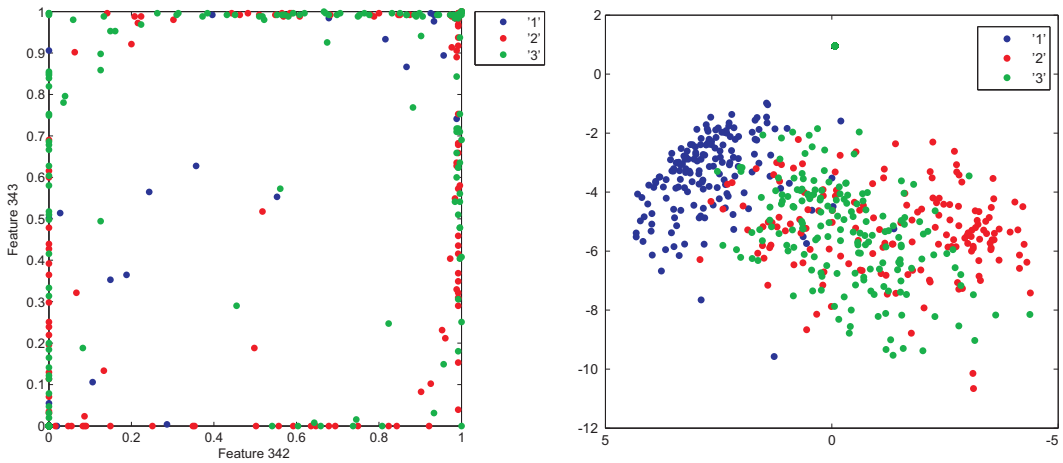


Рис. 2: Примеры изображений рукописных цифр из базы данных MNIST.

формирования признакового пространства обладает рядом существенных недостатков. Во-первых, получается большое количество признаков. Например, для относительно небольших изображений размера 300×200 получается 60000 признаков. Большое количество признаков приводит к высоким временным затратам на обработку данных, большим объемам памяти, требуемой для хранения информации, а также к необходимости сбора большого числа прецедентов для уверенного восстановления скрытых зависимостей в существенно многомерном пространстве. Другим серьезным недостатком полученного признакового пространства является тот факт, что близкие в пространстве признаков объекты не соответствуют одним и тем же классам (см. рис. 3a). Выполнение гипотезы компактности является одним из основных требований для большинства методов распознавания. Методы уменьшения размерности в данных позволяют получать представление выборок в маломерных пространствах, обладающих рядом хороших свойств. В частности, для изображений рукописных цифр метод главных компонент позволяет получить существенно более качественное признаковое пространство (см. рис. 3b).



(a) (b)

Рис. 3: Проекция выборки изображений цифр ’1’, ’2’ и ’3’ на два признака, соответствующих интенсивностям пикселов (a) и на два признака, полученных с помощью метода главных компонент (b).

Пусть имеется некоторая выборка объектов $X = \{\boldsymbol{x}_n\}_{n=1}^N$, $\boldsymbol{x}_n \in \mathbb{R}^D$. Задача уменьшения размерности состоит в получении представления этой выборки в пространстве меньшей

размерности $T = \{\boldsymbol{t}_n\}_{n=1}^N, \boldsymbol{t}_n \in \mathbb{R}^d$. Здесь $d \ll D$, но в частных случаях d может и совпадать с D. Уменьшение размерности в описании данных может преследовать множество целей:

- Сокращение вычислительных затрат при обработке данных;
- Борьба с переобучением. Чем меньше количество признаков, тем меньше требуется объектов для уверенного восстановления скрытых зависимостей в данных и тем выше качество восстановления подобных зависимостей;
- Сжатие данных для более эффективного хранения информации. В этом случае помимо преобразования $X \to T$ требуется иметь возможность осуществлять также обратное преобразование $T \to X$;
- Визуализация данных. Проектирование выборки на двух-/трехмерное пространство позволяет графически представить выборку;
- Извлечение новых признаков. Новые признаки, полученные в результате преобразования $X \to T$, могут оказывать значимый вклад при последующем решении задач распознавания (например, как метод главных компонент в случае рис. 3b);
- и др.

Заметим, что описанные далее методы уменьшения размерности относятся к классу методов обучения без учителя, т.е. в качестве исходной информации выступает только признаковое описание объектов X. В частности, в задаче классификации рукописных цифр результат, показанный на рис. 3b, был получен без использования информации о цифрах (метках класса).

Метод главных компонент

Метод главных компонент (разложение Карунена-Лоева, principal component analysis, PCA) является простейшим методом уменьшения размерности в данных. Идея метода заключается в поиске в исходном пространстве гиперплоскости заданной размерности с последующим проектированием выборки на данную гиперплоскость. При этом выбирается та гиперплоскость, ошибка проектирования данных на которую является минимальной в смысле суммы квадратов отклонений.

Пусть $D = 2, d = 1$, т.е. задача состоит в проектировании двухмерных данных на прямую. Предположим далее, что прямая проходит через начало координат, а ее направление задается единичным вектором $\boldsymbol{u}$, $\|\boldsymbol{u}\| = 1$ (см. рис. 4a). Тогда величина проекции вектора $\boldsymbol{x}$ на эту прямую составляет $\|\boldsymbol{x}_{pr}\| = \boldsymbol{u}^T\boldsymbol{x}$, а сам вектор проекции определяется как $\boldsymbol{x}_{pr} = \|\boldsymbol{x}_{pr}\|\boldsymbol{u}$. По теореме Пифагора квадрат ошибки проектирования вычисляется как

$$\|\boldsymbol{x}_{err}\|^2 = \|\boldsymbol{x}\|^2 - \|\boldsymbol{x}_{pr}\|^2 = \boldsymbol{x}^T\boldsymbol{x} - \boldsymbol{u}^T\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{u}.$$

Таким образом, критерий средней ошибки проектирования выборки X на прямую, задаваемую единичным вектором $\boldsymbol{u}$, может быть записан как

$$J = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n^T\boldsymbol{x}_n - \boldsymbol{u}^T\boldsymbol{x}_n\boldsymbol{x}_n^T\boldsymbol{u}) = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n^T\boldsymbol{x}_n - \boldsymbol{u}^T\left[\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n\boldsymbol{x}_n^T\right]\boldsymbol{u}.$$

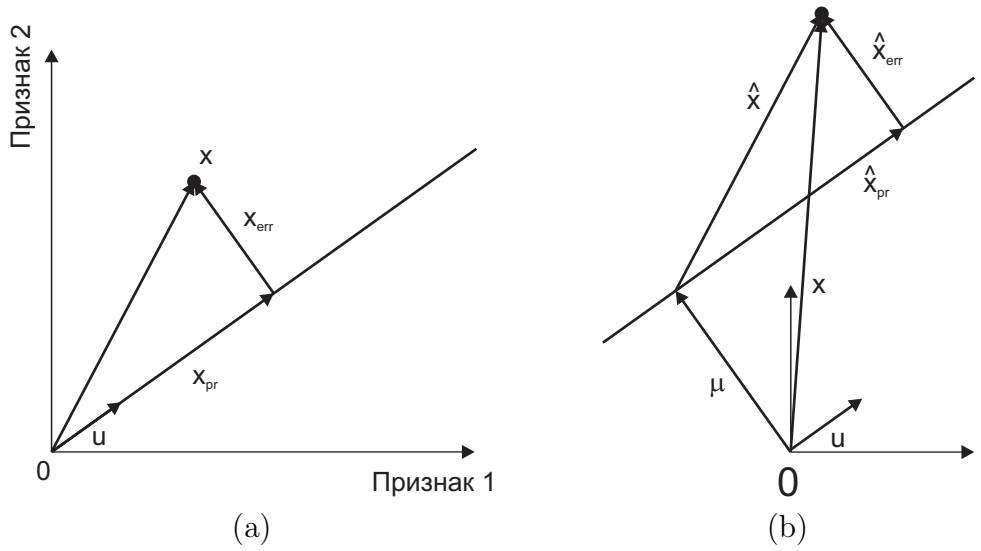


Рис. 4: Проекция объекта $\boldsymbol{x}$ на прямую, задаваемую направляющим вектором $\boldsymbol{u}$ и вектором сдвига $\boldsymbol{\mu}$.

Обозначим через S матрицу $(1/N)\sum_n \boldsymbol{x}_n\boldsymbol{x}_n^T$. Минимизация критерия J по $\boldsymbol{u}$ эквивалентна следующей задаче условной максимизации:

$$\begin{aligned} &\boldsymbol{u}^T S\boldsymbol{u} \to \max_{\boldsymbol{u}}, \\ &\boldsymbol{u}^T\boldsymbol{u} = 1. \end{aligned} \tag{1}$$

Записывая функцию Лагранжа $L(\boldsymbol{u},\lambda)$ и приравнивая к нулю все ее производные, получим необходимое условие экстремума:

$$\begin{aligned} &L(\boldsymbol{u},\lambda) = \boldsymbol{u}^T S\boldsymbol{u} + \lambda(1-\boldsymbol{u}^T\boldsymbol{u}), \\ &\nabla_{\boldsymbol{u}} L(\boldsymbol{u},\lambda) = 2S\boldsymbol{u} - 2\lambda\boldsymbol{u} = 0, && \Rightarrow\ S\boldsymbol{u} = \lambda\boldsymbol{u}, && (2) \\ &\frac{\partial}{\partial\lambda} L(\boldsymbol{u},\lambda) = 1 - \boldsymbol{u}^T\boldsymbol{u} = 0, && \Rightarrow\ \boldsymbol{u}^T\boldsymbol{u} = 1. && (3) \end{aligned}$$

Условие (2) означает, что оптимальный вектор $\boldsymbol{u}$ является собственным вектором матрицы S, отвечающим некоторому собственному значению λ. Заметим, что собственный вектор всегда определен с точностью до нормы, поэтому условию (3) всегда можно удовлетворить.

Условия (2),(3) являются необходимыми условиями экстремума. Для того, чтобы найти точку глобального условного максимума критерия (1), подставим условие (2) в критерий (1):

$$J = \boldsymbol{u}^T S\boldsymbol{u} = \lambda\boldsymbol{u}^T\boldsymbol{u} = \lambda.$$

Таким образом, оптимальный вектор $\boldsymbol{u}$ является собственным вектором матрицы S, отвечающим ее максимальному собственному значению λ_{max}.

Теперь рассмотрим случай, когда прямая не проходит через начало координат (см. рис. 4b). Вектор сдвига прямой $\boldsymbol{\mu}$ относительно начала координат всегда можно выбрать перпендикулярно прямой, т.е. $\boldsymbol{\mu}^T\boldsymbol{u} = 0$. Задача поиска прямой с минимальной ошибкой проектирования может быть сведена к рассмотренной выше задаче путем перехода от объектов

$\boldsymbol{x}_n$ к объектам $\hat{\boldsymbol{x}}_n = \boldsymbol{x}_n - \boldsymbol{\mu}$:

$$J = \frac{1}{N}\sum_{n=1}^{N}((\boldsymbol{x}_n - \boldsymbol{\mu})^T(\boldsymbol{x}_n - \boldsymbol{\mu}) - \boldsymbol{u}^T(\boldsymbol{x}_n - \boldsymbol{\mu})(\boldsymbol{x}_n - \boldsymbol{\mu})^T\boldsymbol{u}) =$$
$$\frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n^T\boldsymbol{x}_n - 2\boldsymbol{\mu}^T\boldsymbol{x}_n + \boldsymbol{\mu}^T\boldsymbol{\mu} - \boldsymbol{u}^T\boldsymbol{x}_n\boldsymbol{x}_n^T\boldsymbol{u}) = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n^T\boldsymbol{x}_n - 2\boldsymbol{\mu}^T\left[\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n\right] + \boldsymbol{\mu}^T\boldsymbol{\mu} - \boldsymbol{u}^T S\boldsymbol{u}.$$

Приравнивая к нулю градиент J по $\boldsymbol{\mu}$, получаем:

$$\nabla_{\boldsymbol{\mu}} J = -2\left[\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n\right] + 2\boldsymbol{\mu} = 0 \;\Rightarrow\; \boldsymbol{\mu}_{opt} = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n.$$

После того, как $\boldsymbol{\mu}$ и $\boldsymbol{u}$ найдены, осталось спроектировать выборку X на найденную прямую:

$$t_n = \boldsymbol{u}^T(\boldsymbol{x}_n - \boldsymbol{\mu}). \tag{4}$$

Сами объекты выборки в исходном пространстве после проектирования можно вычислить следующим образом:

$$\boldsymbol{x}_{n,pr} = t_n\boldsymbol{u} + \boldsymbol{\mu}. \tag{5}$$

Теперь пусть d является произвольным. Следовательно, нам требуется найти вектор сдвига $\boldsymbol{\mu}$ и направляющие векторы гиперплоскости $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d$ такие, чтобы ошибка проектирования выборки на эту гиперплоскость была бы минимальна. Рассуждая аналогично случаю $d = 1$, легко показать, что

$$\begin{aligned} &\boldsymbol{\mu} = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n, \\ &S\boldsymbol{u}_i = \lambda_i\boldsymbol{u}_i,\ i = 1, \ldots, d, \\ &\lambda_1 \geq \lambda_2 \geq \ldots \lambda_D. \end{aligned} \tag{6}$$

Таким образом, оптимальные $\boldsymbol{u}_i$ являются собственными векторами матрицы S, отвечающие ее d наибольшим собственным значениям. Обозначим через U матрицу $[\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d]$. Тогда редукция X при проектировании на оптимальную гиперплоскость вычисляется как

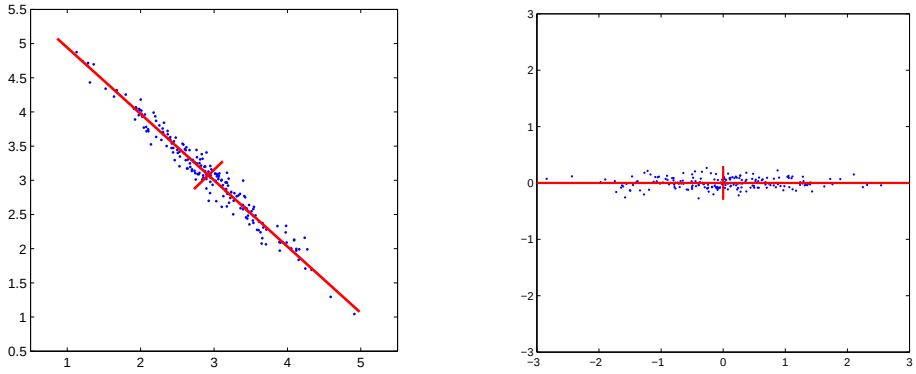
$$\boldsymbol{t}_n = (\boldsymbol{x}_n - \boldsymbol{\mu})^T U,$$

а сами точки проекции определяются как

$$\boldsymbol{x}_{n,pr} = \boldsymbol{t}_n^T U + \boldsymbol{\mu}.$$

Метод главных компонент через критерий максимизации разброса в данных

Наряду с критерием минимизации ошибки проектирования можно рассмотреть альтернативный критерий поиска гиперплоскости, связанный с максимизацией разброса спроектированных точек выборки. Рассмотрим снова простейшую ситуацию $D = 2, d = 1$. Обозначим через t^i случайную величину, соответствующую значению i-ого признака



(a) (b)

Рис. 5: Пример применения метода главных компонент. На рис. a показана исходная выборка в двухмерном пространстве вместе с направлениеми, определяемыми собственными векторами выборочной матрицы ковариации. На рис. b показан переход к некоррелированным признакам.

в редуцированном пространстве $\mathbb{R}^d$. Характеристикой разброса данных в одномерном пространстве является выборочная дисперсия. Предположим, что наша выборка является центрированной, т.е. $\sum_{n=1}^{N} \boldsymbol{x}_n = 0$. Тогда

$$\mathbb{E}t^1 = \frac{1}{N}\sum_{n=1}^{N} t_n = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{u}^T\boldsymbol{x}_n = \boldsymbol{u}^T\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n = 0,$$

$$\mathbb{D}t^1 = \frac{1}{N}\sum_{n=1}^{N}(t_n - \mathbb{E}t^1)^2 = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{u}^T\boldsymbol{x}_n\boldsymbol{x}_n^T\boldsymbol{u} = \boldsymbol{u}^T S\boldsymbol{u}.$$

Таким образом, задача максимизации дисперсии данных на прямой совпадает с задачей оптимизации (1). Следовательно, оптимальная прямая определяется собственным вектором матрицы S, отвечающим наибольшему собственному значению λ_{max}.

Рассмотрим случай произвольного значения d. Характеристикой разброса данных в многомерном пространстве является выборочная матрица ковариации. В качестве скалярного критерия разброса выберем след выборочной матрицы ковариации, что эквивалентно сумме выборочных дисперсий по ортогональным направлениям $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d$. Тогда можно показать, что максимизация следа выборочной матрицы ковариации приводит к решению (6). При этом значение критерия составляет

$$J = \sum_{i=1}^{d}\mathbb{D}t^i = \sum_{i=1}^{d}\lambda_i.$$

Здесь $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_D$.

Итак, метод главных компонент предполагает переход от исходного базиса к базису из собственных векторов матрицы ковариации S с дальнейшим отбрасыванием проекций выборки на собственные вектора, отвечающие $D - d$ наименьшим собственным значениям. В базисе из собственных векторов матрица ковариации S имеет диагональный вид $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_D)$. Таким образом, признаки, получаемые с помощью метода главных компонент, являются некоррелированными. Переход к некоррелированным признакам часто является разумным

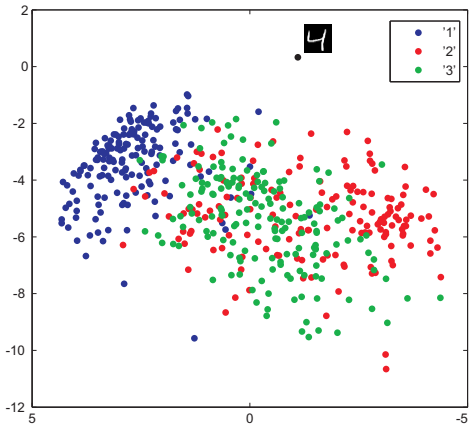


Рис. 6: Метод главных компонент может быть использован для решения задачи идентификации.

методом предобработки исходных данных. Поэтому метод главных компонент применяется и в случае $d = D$.

Рассмотрим простой модельный пример применения метода главных компонент. Пусть исходная выборка представляет собой данные в двухмерном пространстве (см. рис. 5a). При использовании метода главных компонент центр координат нового пространства переносится в центр выборки, а оси определяются собственными векторами выборочной матрицы ковариации (см. рис. 5b). Таким образом, новые признаки являются некоррелированными. В том случае, если $d = 1$, то дополнительно осуществляется проекция выборки на направление, соответствующее наибольшему собственному значению (направление с наибольшей дисперсией). Для данного примера это координата x.

Решение задачи идентификации

Интерпретация метода главных компонент с помощью разброса данных позволяет использовать его для решения задачи идентификации. Мы знаем, что d наибольших собственных значений λ_i выборочной матрицы ковариации определяют дисперсии выборки $\mathbb{D}t^i$ вдоль направлений $\boldsymbol{u}_i$. Из неравенства Чебышева следует, что вероятность отклонения случайной величины от своего математического ожидания на k стандартных отклонения не превышает $1/k^2$. Следовательно, с вероятностью не выше $1/k^2$ значение i-ого признака для n-ого объекта в редуцированном пространстве находится в доверительном интервале

$$-k\sqrt{\lambda_i} \leq t_{ni} \leq k\sqrt{\lambda_i}.$$

Если дополнительно известно, что случайная величина t^i является нормальной или приближенно нормальной, то доверительный интервал значительно сокращается. В частности, вероятность отклонения на 3 стандартных отклонения составляет всего 0.3%. Этот результат известен как «правило трех сигма».

Таким образом, если оказалось, что тестовый объект после проектирования на $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d$ имеет хотя бы одно значение проекции, которое не укладывается в доверительный интервал, то это является поводом признать этот объект не соответствующим генеральной совокупности объектов, представленной в обучающей выборке. На рис. 6 показан пример идентификации с помощью метода главных компонент. Рассматривается задача распознавания рукописных

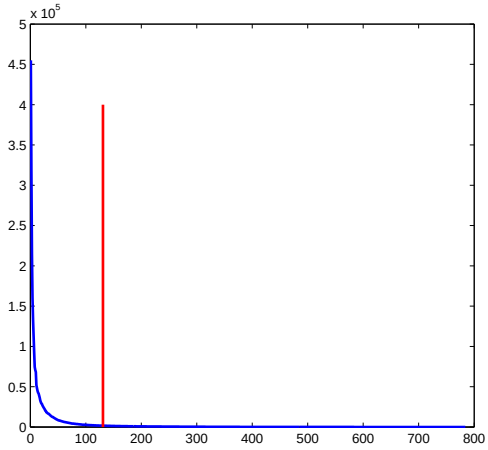


Рис. 7: Схема выбора размерности редуцированного пространства для метода главных компонент.

цифр с классами '1', '2', '3'. Изображение, соответствующее цифре '4', не укладывается в доверительный интервал проекций по первым двум собственным векторам.

Выбор размерности редуцированного пространства d

До сих пор предполагалось, что размерность редуцированного пространства d задается пользователем заранее. Это значение легко выбрать в том случае, если стоит задача визуализации данных ($d = 2$ или $d = 3$) или задача вложения выборки в заданный объем памяти. Однако, во многих других случаях выбор d является далеко не очевидным из априорных предположений. Для метода главных компонент существует простой эвристический прием выбора величины d. Одной из особенностью метода является тот факт, что все редуцированные пространства для $d = 1, 2, \ldots, D$ являются вложенными друг в друга. В частности, однократное вычисление всех собственных векторов и собственных значений выборочной матрицы ковариации S позволяет получить редуцированное пространство для любого значения d. При этом ошибка проектирования данных на соответствующую гиперплоскость определяется величиной $\sum_{i=d+1}^{D} \lambda_i$. Поэтому для выбора значения d можно отобразить на графике собственные значения в порядке убывания (см. рис. 7) и выбрать порог отсечения таким образом, чтобы справа остались значения, незначимо отличные от нуля. Другой способ предполагает выбор порога так, чтобы справа оставался определенный процент от общей площади под кривой (например, 5% или 1%), т.е.

$$d : \frac{\sum_{i=d+1}^{D} \lambda_i}{\sum_{i=1}^{D} \lambda_i} < \eta.$$

Площадь под кривой определяется значением $\mathrm{tr}(S)$ и соответствует величине разброса в данных.

Эффективные вычисления при $D > N$

Алгоритм 1 Схема метода главных компонент

Вход: $X \in \mathbb{R}^{N \times D}$ – исходная выборка данных, d – размерность редуцированного пространства
Выход: $T \in \mathbb{R}^{N \times d}$ – представление выборки в редуцированном пространстве

$\bar{\boldsymbol{x}} = \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{x}_n$; *// Вычисляем выборочное среднее*

$\boldsymbol{x}_n \leftarrow \boldsymbol{x}_n - \bar{\boldsymbol{x}}$; *// Переносим начало координат в центр выборки*

если $N > D$ **то**

$S = \frac{1}{N} X^T X$; *// Вычисляем выборочную матрицу ковариации*

$S = Q \Lambda Q^T, \Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_D), Q^T Q = I, Q = (\boldsymbol{q}_1 | \ldots | \boldsymbol{q}_D)$; *// Находим собственные вектора и собственные значения матрицы ковариации*

Выбираем d наибольших собственных значений $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_d$ и соответствующие им собственные вектора $W = (\boldsymbol{q}_1 | \ldots | \boldsymbol{q}_d)$;

иначе

$S = \frac{1}{N} X X^T$;

$S = Q \Lambda Q^T, \Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_D), Q^T Q = I, Q = (\boldsymbol{q}_1 | \ldots | \boldsymbol{q}_D)$; *// Находим собственные вектора и собственные значения матрицы S*

$Q \leftarrow \frac{1}{\sqrt{N}} X^T Q \operatorname{diag}\left(\frac{1}{\sqrt{\lambda_1}}, \ldots, \frac{1}{\sqrt{\lambda_D}}\right)$; *// Переходим к нормированным собственным векторам выборочной матрицы ковариации*

Выбираем собственные вектора, соответствующие d наибольшим собственным значениям $W = (\boldsymbol{q}_1 | \ldots | \boldsymbol{q}_d)$;

$T = XW$; *// Проектируем выборку на выбранные направления*

При использовании метода главных компонент необходимо вычислять выборочную матрицу ковариации, которая имеет размер $D \times D$, а также ее собственные вектора и собственные значения. Сложность этих операций составляет $O(ND^2)$ и $O(D^3)$. В том случае, если $D > N$, то существует способ более экономного вычисления собственных векторов и собственных значений матрицы ковариации с помощью матрицы размера $N \times N$ и сложностью, соответственно, $O(DN^2)$ и $O(N^3)$. Действительно, в пространстве размерности D множество из N точек порождает линейное многообразие максимальной размерности $N-1$. Поэтому не имеет смысла применять метод главных компонент для $d > N-1$. С точки зрения матрицы ковариации это означает, что только $N-1$ собственных значений отличны от нуля. Все остальные собственные вектора не имеет смысла вычислять, т.к. дисперсия выборки вдоль этих направлений заведомо равна нулю.

Пусть $X \in \mathbb{R}^{N \times D}$ – исходная выборка с нулевым центром, т.е. $\frac{1}{N} \sum_{n=1}^{N} \boldsymbol{x}_n = \mathbf{0}$. Тогда выборочная матрица ковариации $S = \frac{1}{N} X^T X$. Рассмотрим собственные вектора и собственные значения матрицы S:

$$\frac{1}{N} X^T X \boldsymbol{q}_i = \lambda_i \boldsymbol{q}_i.$$

Домножим обе части этого уравнения на X слева:

$$\frac{1}{N} X X^T (X \boldsymbol{q}_i) = \lambda_i (X \boldsymbol{q}_i). \tag{7}$$

Обозначая $\boldsymbol{v}_i = X \boldsymbol{q}_i$, получаем

$$\frac{1}{N} X X^T \boldsymbol{v}_i = \lambda_i \boldsymbol{v}_i. \tag{8}$$

Таким образом, матрица $\frac{1}{N} X X^T$ размера $N \times N$ имеет те же собственные значения, что и выборочная матрица ковариации S (у которой, в свою очередь, есть $D - N$ дополнительных

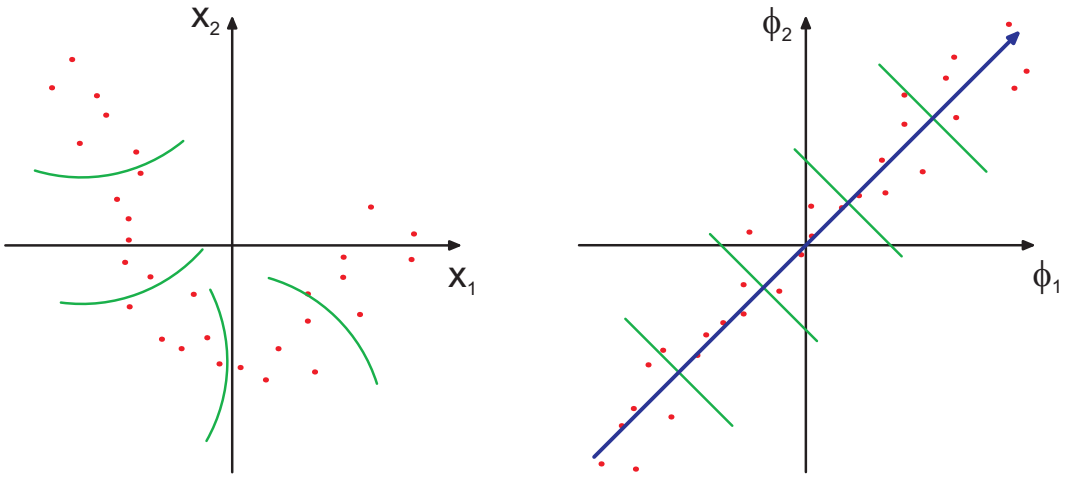


Рис. 8: Иллюстрация ядрового метода главных компонент.

нулевых собственных значений). Сложность поиска собственных значений и собственных векторов матрицы $\frac{1}{N}XX^T$ составляет $O(N^3)$, что может давать значительную выгоду по сравнению с $O(D^3)$ при $D > N$. Для получения собственных векторов матрицы S домножим обе части последнего уравнения на X^T:

$$\frac{1}{N}X^TX(X^T\boldsymbol{v}_i) = \lambda_i(X^T\boldsymbol{v}_i).$$

Таким образом, $X^T\boldsymbol{v}_i$ является собственным вектором матрицы S, отвечающим собственному значению λ_i. Однако, в том случае, если исходные вектора $\boldsymbol{v}_i$ являются нормированными, т.е. $\|\boldsymbol{v}_i\| = 1$, то вектора $X^T\boldsymbol{v}_i$ нормированными уже не являются. Нормированные вектора можно получить с помощью следующего преобразования:

$$\boldsymbol{q}_i = \frac{1}{\sqrt{N\lambda_i}}X^T\boldsymbol{v}_i.$$

Теперь, объединяя все вышесказанное, можно составить схему метода главных компонент, представленную в алгоритме 1.

Ядровой метод главных компонент

Метод главных компонент является линейным методом уменьшения размерности, т.к. преобразование (4) от $\boldsymbol{x}_n$ к $\boldsymbol{t}_n$, а также обратное преобразование (5) от $\boldsymbol{t}_n$ к $\boldsymbol{x}_{n,pr}$ являются линейными. В том случае, если выборка данных образует в многомерном пространстве нелинейное многообразие, то применение метода главных компонент будет приводить к большой ошибке проектирования.

Обобщение метод главных компонент на нелинейный случай возможно с помощью ядрового перехода. Рассмотрим некоторое нелинейное преобразование $\phi : \mathbb{R}^D \to H$ такое, что в новом пространстве H нелинейное многообразие выборки переходит в гиперплоскость (см. рис. 8). Например, квадратичное многообразие в двухмерном пространстве (x, y) вида

$$a_{11}x^2 + a_{12}xy + a_{22}y^2 + a_1x + a_2y + a_3 = 0$$

является гиперплоскостью в пятимерном пространстве (x^2, xy, y^2, x, y). Пусть далее известно, что скалярное произведение в пространстве H может быть вычислено с помощью функции объектов в исходном пространстве

$$\langle \boldsymbol{\phi}(\boldsymbol{x}), \boldsymbol{\phi}(\boldsymbol{y}) \rangle = \boldsymbol{\phi}^T(\boldsymbol{x})\boldsymbol{\phi}(\boldsymbol{y}) = K(\boldsymbol{x}, \boldsymbol{y}).$$

Такая функция K называется ядровой функцией. Если представить схему метода главных компонент таким образом, чтобы она зависела от выборки X только посредством скалярных произведений объектов $\boldsymbol{x}_n^T\boldsymbol{x}_m$, то тогда поиск оптимальной гиперплоскости проектирования в пространстве H можно осуществлять без рассмотрения преобразования ϕ. Преобразуем схему метода главных компонент в соответствии с этим требованием.

Предположим сначала, что выборка является центрированной, т.е. $\sum_{n=1}^N \boldsymbol{x}_n = 0$. Матрица всех скалярных произведений объектов выборки X вычисляется как XX^T. Условие (7) показывает, что матрица скалярных произведений XX^T имеет те же собственные значения, что и выборочная матрица ковариации X^TX. Пусть найден собственный вектор $\boldsymbol{v}_i$ из условия (8). Тогда, как было показано выше, вектор $\boldsymbol{u}_i = X^T\boldsymbol{v}_i$ является собственным вектором выборочной матрицы ковариации S. Вычисление вектора $\boldsymbol{u}_i$ требует знания матрицы X^T, что в случае ядрового перехода соответствует знанию преобразования ϕ. Однако, при уменьшении размерности с помощью метода главных компонент нам требуется знать лишь величину проекции выборки X на собственные вектора $\boldsymbol{u}_i$. Эти проекции вычисляются как

$$t_{ni} = \boldsymbol{x}_n^T\boldsymbol{u}_i = \boldsymbol{x}_n^TX^T\boldsymbol{v}_i = (X\boldsymbol{x}_n)^T\boldsymbol{v}_i.$$

Таким образом, проекции t_{ni} зависят от выборки X только посредством скалярных произведений $X\boldsymbol{x}_n$.

Пусть теперь выборка X не является центрированной. Рассмотрим центрированную выборку $\hat{\boldsymbol{x}}_n = \boldsymbol{x}_n - \frac{1}{N}\sum_{m=1}^N \boldsymbol{x}_m$. Скалярное произведение $\hat{\boldsymbol{x}}_n^T\hat{\boldsymbol{x}}_l$ может быть вычислено как

$$\hat{\boldsymbol{x}}_n^T\hat{\boldsymbol{x}}_l = \boldsymbol{x}_n^T\boldsymbol{x}_l - \frac{1}{N}\sum_{m=1}^N \boldsymbol{x}_n^T\boldsymbol{x}_m - \frac{1}{N}\sum_{m=1}^N \boldsymbol{x}_m^T\boldsymbol{x}_l + \frac{1}{N^2}\sum_{m,k=1}^N \boldsymbol{x}_m^T\boldsymbol{x}_k.$$

Таким образом, скалярное произведение объектов центрированной выборки может быть вычислено через скалярные произведения объектов исходной выборки.