Лекция 13. Уменьшение размерности в данных. Метод главных компонент.

А. С. Конушин[1] Д. П. Ветров[2] Д. А. Кропотов[3]
В. С. Конушин[1] О. В. Баринова[1]

[1]МГУ, ВМиК, лаб. КГ

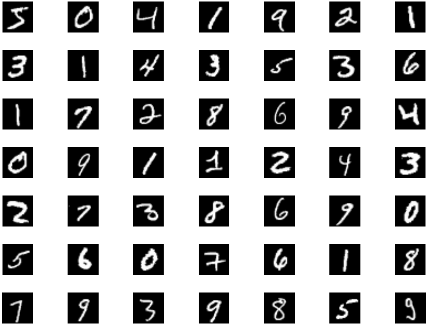
[2]МГУ, ВМиК, каф. ММП

[3]ВЦ РАН

Спецкурс «Структурные методы анализа изображений и сигналов»

План лекции

Практический пример: распознавание рукописных цифр



Задача: выбор адекватного признакового описания изображения $\boldsymbol{x} = (x_1, \ldots, x_D)$ для дальнейшего использования алгоритмов распознавания.

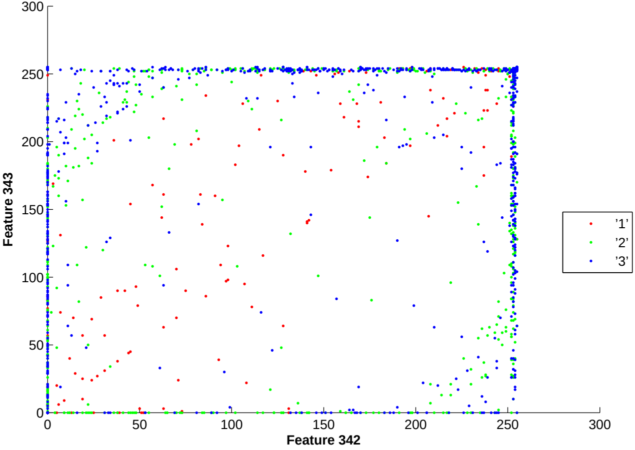
Выбор признакового пространства

Прямой способ: вытянуть матрицу интенсивностей в вектор.
Такой способ описания данных не является адекватным в силу ряда причин:

- Слишком большое число признаков (для картинки 28×28 получается 784 признака). Как следствие, большое время обработки, проблемы переобучения и т.д.
- Близкие в признаковом пространстве объекты не соответствуют одному классу. Гипотеза компактности является одним из основных предположений большинства методов распознавания.

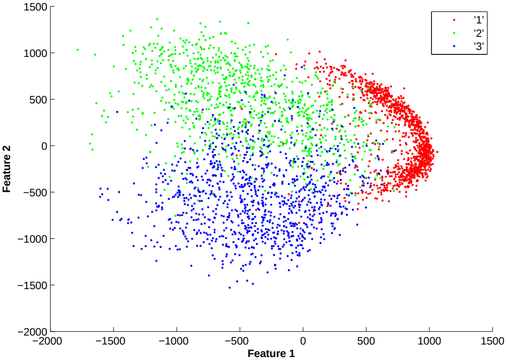
Иллюстрация

Близкие в признаковом пространстве объекты не соответствуют одному классу.



Преобразование признакового пространства

После преобразования признакового описания (с помощью метода главных компонент) количество признаков уменьшается (с 784 до нескольких десятков), причем объекты одного класса образуют компактные области в признаковом пространстве.



Сокращение размерности в данных

Пусть имеется некоторая выборка $X = \{\boldsymbol{x}_n\}_{n=1}^N$, $\boldsymbol{x}_n \in \mathbb{R}^D$.
Цель — представить выборку в пространстве меньшей размерности $d < D$, причем в новом пространстве «схожие» объекты должны образовывать компактные области.
Причины сокращения размерности:

- уменьшение вычислительных затрат при обработке данных
- борьба с переобучением
- сжатие данных для более эффективного хранения информации
- визуализация данных
- извлечение признаков
- интерпретация данных
- ...

План лекции

Идея метода

Идея метода главных компонент (разложение Карунена-Лоева, Principal Component Analysis, PCA) — проекция данных на гиперплоскость с наименьшей ошибкой проектирования. Эквивалентная формулировка: поиск проекции на гиперплоскость с сохранением большей части дисперсии в данных.

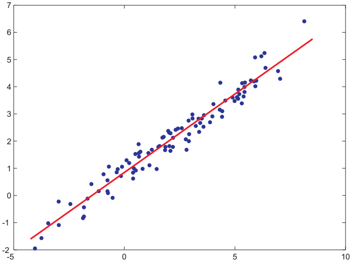
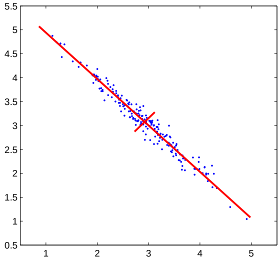


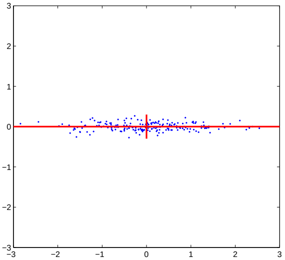
Иллюстрация решения PCA

Выборка в исходном пространстве



$\longrightarrow$

Выборка в пространстве декоррелированных признаков



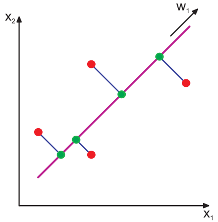
- Переносим начало координат в центр выборки
- Поворачиваем оси координат так, чтобы признаки стали декоррелированными
- Отбрасываем направления с низкой дисперсией

Линейное преобразование

Пусть D — размерность исходного пространства, d — размерность искомого пространства. Будем искать преобразование данных в семействе линейных функций:

$$\boldsymbol{x} = \boldsymbol{\mu} + t_1\boldsymbol{w}_1 + \cdots + t_d\boldsymbol{w}_d = W\boldsymbol{t} + \boldsymbol{\mu}$$

Здесь $\boldsymbol{x} \in \mathbb{R}^D$, $\boldsymbol{t} \in \mathbb{R}^d$ — новые координаты объекта, $W = (\boldsymbol{w}_1|\dots|\boldsymbol{w}_d) \in \mathbb{R}^{D\times d}$, $\boldsymbol{\mu} \in \mathbb{R}^D$.
Графическая интерпретация — проекция данных на гиперплоскость, $\boldsymbol{w}_1, \dots, \boldsymbol{w}_d$ — базис гиперплоскости.



Критерий поиска гиперплоскости

Критерий выбора гиперплоскости — минимальная ошибка проектирования в смысле суммы квадратов отклонений исходных точек и их проекций.
Пусть $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_D$ — ортонормированный базис в пространстве $\mathbb{R}^D$, первые d компонент которого — базис искомой гиперлоскости.
Тогда

$$\boldsymbol{x}_n = \sum_{i=1}^{D} (\boldsymbol{x}_n^T \boldsymbol{w}_i) \boldsymbol{w}_i \text{ — исходные точки}$$

$$\hat{\boldsymbol{x}}_n = \sum_{i=1}^{d} t_{ni} \boldsymbol{w}_i + \sum_{i=d+1}^{D} \mu_i \boldsymbol{w}_i \text{ — приближение}$$

Критерий

$$J = \frac{1}{N} \sum_{n=1}^{N} \|\boldsymbol{x}_n - \hat{\boldsymbol{x}}_n\|^2 \to \min_{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_d, \boldsymbol{t}_1, \ldots, \boldsymbol{t}_N, \boldsymbol{\mu}}$$

Решение задачи

$$J = \frac{1}{N}\sum_{n=1}^{N}\|\boldsymbol{x}_n - \hat{\boldsymbol{x}}_n\|^2 \to \min_{\boldsymbol{w}_1,\dots,\boldsymbol{w}_d,\boldsymbol{t}_1,\dots,\boldsymbol{t}_N,\boldsymbol{\mu}}$$

Можно показать, что решением задачи будет следующее:

$$\bar{\boldsymbol{x}} = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n \text{ — выборочное среднее}$$

$$S = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n - \bar{\boldsymbol{x}})(\boldsymbol{x}_n - \bar{\boldsymbol{x}})^T \text{ — выборочная матрица ковариации}$$

$\boldsymbol{w}_1,\dots,\boldsymbol{w}_d$ — ортонормированный базис из собственных векторов матрицы S, отвечающих d наибольшим собственным значениям $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d$.

$$t_{ni} = \boldsymbol{x}_n^T\boldsymbol{w}_i$$

$$\mu_i = \bar{\boldsymbol{x}}^T\boldsymbol{w}_i$$

Альтернативная интерпретация PCA

Рассмотрим в качестве критерия проектирования максимизацию дисперсии спроектированных данных. Пусть $d = 1$, т.е. необходимо найти некоторое направление $\boldsymbol{w}_1$: $\boldsymbol{w}_1^T\boldsymbol{w}_1 = 1$. Каждый объект выборки $\boldsymbol{x}_n$ проектируется в $\boldsymbol{w}_1^T\boldsymbol{x}_n$. Тогда дисперсия проекций вычисляется как

$$\frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{w}_1^T\boldsymbol{x}_n - \boldsymbol{w}_1^T\bar{\boldsymbol{x}})^2 = \boldsymbol{w}_1^T S\boldsymbol{w}_1$$

Здесь

$$\bar{\boldsymbol{x}} = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_i, \quad S = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n - \bar{\boldsymbol{x}})(\boldsymbol{x}_n - \bar{\boldsymbol{x}})^T$$

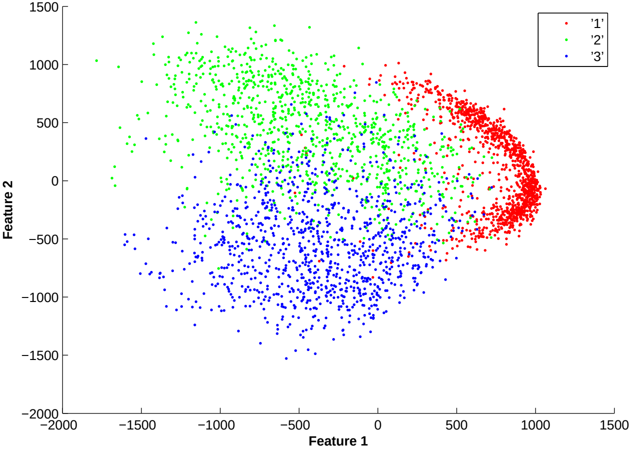
Альтернативная интерпретация II

Теперь необходимо максимизировать дисперсию проекций $\boldsymbol{w}_1^T S \boldsymbol{w}_1$ при ограничении $\boldsymbol{w}_1^T \boldsymbol{w}_1 = 1$. Таким образом, мы приходим к точно такому же критерию, который возникал в случае минимизации невязки. Максимальная дисперсия достигается для собственного вектора выборочной матрицы ковариации, отвечающего максимальному собственному значению.

Рассматривая аналогично проектирование на новые направления, ортогональные уже найденным, получим, что наилучшее d-мерное линейное подпространство определяется собственными векторами выборочной матрицы ковариации, отвечающие d максимальным собственным значениям.

Пример с распознаванием рукописных цифр

Проекция данных на первые два признака (два собственных вектора матрицы ковариации, отвечающих наибольшим собственным значениям).



Интерпретация новых признаков

По построению каждой точке $\boldsymbol{x} \in \mathbb{R}^D$ соответствует некоторая картинка. Новые признаки $\boldsymbol{t} \in \mathbb{R}^d$ — проекции на выбранные направления (собственные вектора). Изменение одного признака в новом пространстве $\mathbb{R}^d$ соответствует движению вдоль собственного вектора в исходном пространстве $\mathbb{R}^D$.

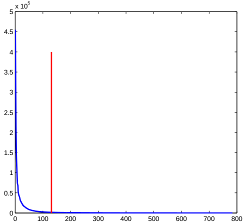
Движение вдоль первого признака:

Интерпретация: изменение ширины цифры

Движение вдоль второго признака:

Интерпретация: изменение характерного наклона цифры

Выбор размерности подпространства d в PCA



- Рассматриваем собственные значения в порядке убывания $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_D$
- Выбираем d так, что $\sum_{i=d+1}^{D} \lambda_i \ll \sum_{i=1}^{d} \lambda_i$ либо, чтобы

$$\sum_{i=1}^{d} \lambda_i \geq \gamma \sum_{i=1}^{D} \lambda_i$$

Здесь γ — доля объясняемой дисперсии, типичные значения 0.95, 0.99.

Неоднозначность решения PCA

- Если все собственные значения выборочной матрицы ковариации различны, т.е. $\lambda_1 > \lambda_2 > \cdots > \lambda_D$, то ортонормированный базис из собственных векторов определен однозначно, т.е. гиперплоскость проекции определена однозначно. Тем не менее, в полученной гиперплоскости базис может быть выбран произвольным образом.
- Если существуют одинаковые собственные вектора, то тогда гиперплоскость проекции определена не однозначно.

В PCA существует произвол в выборе координат объектов в новом пространстве

План лекции

Мотивация

PCA может быть сформулирован как вероятностная модель с латентными переменными, для оптимизации которой используется метод максимального правдоподобия.
Преимущества вероятностной постановки:

- Вычисление правдоподобия на тестовой выборке позволяет непосредственно сравнивать различные вероятностные модели
- EM-алгоритм для PCA позволяет быстро находить решения в ситуациях, когда требуется небольшое число лидирующих главных компонент, а также позволяет избежать вычисление выборочной матрицы ковариации в качестве промежуточного шага
- Возможность применения байесовского подхода для автоматического определения количества главных компонент (по аналогии с методом релевантных векторов)
- Можно работать со смесью PCA и использовать вариант EM-алгоритма для обучения в такой модели

Вероятностная модель PCA

Пусть имеется выборка $\{\boldsymbol{x}_n\}_{n=1}^N$, $\boldsymbol{x}_n \in \mathbb{R}^d$. Рассмотрим для каждого объекта $\boldsymbol{x}_n$ латентную переменную $\boldsymbol{t}_n \in \mathbb{R}^d$ как координаты объекта $\boldsymbol{x}_n$ в подпространстве, вообще говоря, меньшей размерности $d < D$. Определим априорное распределение в пространстве латентных переменных как

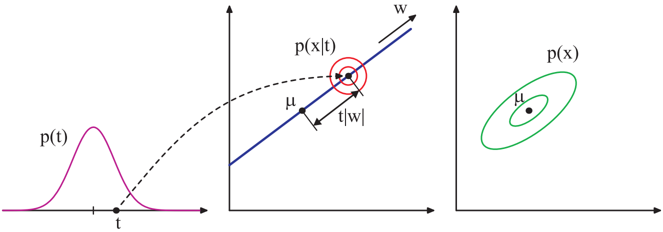
$$p(\boldsymbol{t}) = \mathcal{N}(\boldsymbol{t}|\mathbf{0}, I)$$

Модель наблюдаемых переменных $\boldsymbol{x}$ представляет собой линейное преобразование латентной переменной с добавлением гауссовского шума с единой дисперсией по всем направлениям:

$$p(\boldsymbol{x}|\boldsymbol{t}) = \mathcal{N}(\boldsymbol{x}|W\boldsymbol{t} + \boldsymbol{\mu}, \sigma^2 I)$$

Здесь $W \in \mathbb{R}^{D\times d}$, $\boldsymbol{\mu} \in \mathbb{R}^D$.

Иллюстрация вероятностной модели PCA



Функция правдоподобия

Функция правдоподобия может быть вычислена как

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x}|\boldsymbol{t})p(\boldsymbol{t})d\boldsymbol{t}$$

Этот интеграл является сверткой двух нормальных распределений и может быть вычислен аналитически. В результате получается снова нормальное распределение

$$p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, C), \quad C = WW^T + \sigma^2 I$$

Действительно,

$$\begin{aligned}
\mathbb{E}\boldsymbol{x} &= \mathbb{E}(W\boldsymbol{t} + \boldsymbol{\mu} + \boldsymbol{\varepsilon}) = \boldsymbol{\mu} \\
\mathbb{E}(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^T &= \mathbb{E}(W\boldsymbol{t} + \boldsymbol{\varepsilon})(W\boldsymbol{t} + \boldsymbol{\varepsilon})^T = \\
&= W\mathbb{E}\boldsymbol{t}\boldsymbol{t}^T W^T + \mathbb{E}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T = WW^T + \sigma^2 I
\end{aligned}$$

Инвариантность относительно поворота координат в пространстве латентных переменных

Пусть R — некоторая ортогональная матрица. Рассмотрим матрицу базисных векторов $\tilde{W} = WR$. В этом случае вероятностная модель PCA приводит к тому же распределению, что и в случае матрицы W. Действительно, т.к. $RR^T = I$, то

$$C = \tilde{W}\tilde{W}^T + \sigma^2 I = WRR^T W^T + \sigma^2 I = WW^T + \sigma^2 I$$

Решение

Логарифм правдоподобия для рассматриваемой модели выглядит как

$$\log p(X|\boldsymbol{\mu}, W, \sigma^2) = \sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\boldsymbol{\mu}, W, \sigma^2) =$$

$$= -\frac{ND}{2}\log 2\pi - \frac{N}{2}\log\det C - \frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{\mu})^T C^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu})$$

Дифференцируя правдоподобие и приравнивая производную к нулю, получаем:

$$\boldsymbol{\mu}_{ML} = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n, \qquad \boldsymbol{\sigma}^2_{ML} = \frac{1}{D-d}\sum_{i=d+1}^{D}\lambda_i$$

$$W_{ML} = U_d(L_d - \sigma^2_{ML} I)^{1/2} R$$

Здесь $U_d \in \mathbb{R}^{D\times d}$ — матрица, состоящая из d собственных векторов выборочной матрицы ковариации, отвечающие d наибольшим собственным значениям $\lambda_1, \ldots, \lambda_d$, $L_d = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$, R — произвольная ортогональная матрица.

Апостериорное распределение латентной переменной

Можно показать, что

$$p(\boldsymbol{t}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{t}|M^{-1}W^T(\boldsymbol{x}-\boldsymbol{\mu}), \sigma^2 M^{-1})$$

Здесь $M = W^T W + \sigma^2 I$. Для решения задачи визуализации данных в пространстве латентных переменных требуется мат. ожидание латентной переменной по своему апостериорному распределению:

$$\mathbb{E}(\boldsymbol{t}) = M^{-1}W_{ML}(\boldsymbol{x}-\overline{\boldsymbol{x}})$$

EM-алгоритм в общем виде

Пусть имеется вероятностная модель, в которой:

- часть переменных X — известна
- часть переменных T — не известна
- имеется некоторый набор параметров Θ

Требуется оценить набор параметров Θ с помощью метода максимального правдоподобия:

$$p(X|\Theta) = \int p(X, T|\Theta) dT \to \max_{\Theta}$$

Схема EM-алгоритма

Требуется найти максимум правдоподобия в вероятностной модели со скрытыми переменными:

$$p(X|\Theta) = \int p(X,T|\Theta)dT \to \max_{\Theta} \Leftrightarrow \log\left(\int p(X,T|\Theta)dT\right) \to \max_{\Theta}$$

- E-шаг. Фиксируется значение параметров Θ_{old}. Оценивается апостериорное распределение на скрытые переменные $p(T|X,\Theta_{old})$, и полное правдоподобие усредняется по полученному распределению:

$$\mathbb{E}_{T|X,\Theta_{old}} \log p(X,T|\Theta) = \int \log p(X,T|\Theta) p(T|X,\Theta_{old}) dT$$

- M-шаг. Фиксируется апостериорное распределение $p(T|X,\Theta_{old})$, и производится поиск новых значений параметров Θ_{new}:

$$\Theta_{new} = \arg\max_{\Theta} \mathbb{E}_{T|X,\Theta_{old}} \log p(X,T|\Theta)$$

- Шаги E и M повторяются до сходимости.

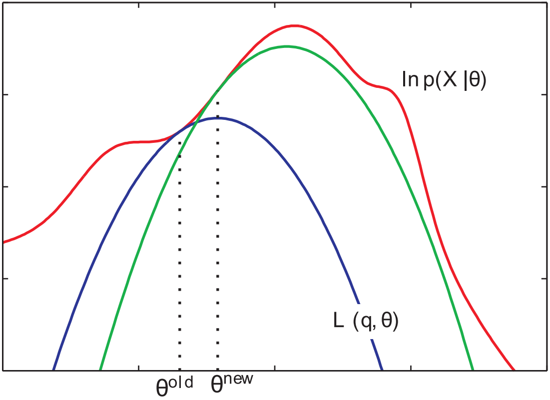
Нижняя оценка для логарифма правдоподобия

Можно показать, что

$$\log p(X|\Theta) \geq \mathbb{E}_{T|X,\Theta_{old}} \log p(X,T|\Theta) + Const \text{ и } \Longleftrightarrow \Theta = \Theta_{old}$$

Здесь $Const$ - некоторая константа, не зависящая от Θ.

Иллюстрация EM-алгоритма



EM-алгоритм для PCA

$$\mathbb{E}\boldsymbol{t}_n = M^{-1}W^T(\boldsymbol{x} - \overline{\boldsymbol{x}})$$

$$\mathbb{E}\boldsymbol{t}_n\boldsymbol{t}_n^T = \sigma^2 M^{-1} + \mathbb{E}\boldsymbol{t}_n\mathbb{E}\boldsymbol{t}_n^T$$

$$W_{new} = \left[\sum_{n=1}^N (\boldsymbol{x}_n - \overline{\boldsymbol{x}})\mathbb{E}\boldsymbol{t}_n^T\right]\left[\sum_{n=1}^N \mathbb{E}\boldsymbol{t}_n\boldsymbol{t}_n^T\right]^{-1}$$

$$\sigma^2_{new} = \frac{1}{ND}\sum_{n=1}^N [\|\boldsymbol{x}_n - \overline{\boldsymbol{x}}\|^2 - 2\mathbb{E}\boldsymbol{t}_n^T W_{new}^T(\boldsymbol{x}_n - \overline{\boldsymbol{x}}) + \mathrm{tr}(\mathbb{E}\boldsymbol{t}_n\boldsymbol{t}_n^T W_{new}^T W_{new})]$$

Мотивация EM-алгоритма для PCA

Несмотря на существование решения для W и σ^2 в явном виде, использование EM-алгоритма может быть предпочтительным в ряде случаев:

- EM-алгоритм избегает вычисления выборочной матрицы ковариации (сложность $O(ND^2)$) и поиска ее собственных значений (сложность $O(D^3)$). Самые сложные операции в EM-алгоритме требуют $O(NDd)$ и $O(d^3)$, что может дать существенный выигрыш в скорости для данных больших размерностей.
- EM-алгоритм может быть применен для модели факторного анализа, для которой не существует решения в явном виде. Эта модель полностью повторяет вероятностную модель для PCA, однако различные факторы могут иметь разную дисперсию:

$$p(\boldsymbol{x}|\boldsymbol{t}) = \mathcal{N}(\boldsymbol{x}|W\boldsymbol{t} + \boldsymbol{\mu}, \Psi)$$

Здесь Ψ — диагональная матрица.

Проблема выбора количества главных компонент d

- Параметр d является типичным структурным параметром и должен задаваться до старта ЕМ-алгоритма
- Одним из возможных способов выбора d является визуальное оценивание распределения собственных значений выборочной матрицы ковариации с выделением области, в которой собственные значения не отличаются существенно от нуля. К сожалению, на практике такой порог часто провести не удается.
- Разумной альтернативой является байесовский подход

Априорное распределение на базисные вектора

Выберем в качестве априорного распределения на веса нормальное распределение с отдельными коэффициентами регуляризации для каждого базисного вектора:

$$p(W|\boldsymbol{\alpha}) = \prod_{i=1}^{D} \left(\frac{\alpha_i}{2\pi}\right)^{D/2} \exp\left(-\frac{1}{2}\alpha_i \boldsymbol{w}_i^T \boldsymbol{w}_i\right)$$

Здесь $\boldsymbol{w}_i$ — i-ая колонка матрицы W. Для подбора коэффициентов $\boldsymbol{\alpha}$ можно использовать максимизацию обоснованности:

$$p(X|\boldsymbol{\alpha}, \boldsymbol{\mu}, \sigma^2) = \int p(X|W, \boldsymbol{\mu}, \sigma^2) p(W|\boldsymbol{\alpha}) dW$$

Данный интеграл не берется аналитически, поэтому возможными путями являются приближение Лапласа и вариационный подход.

Процедура обучения

Использование приближения Лапласа приводит к следующей формуле пересчета коэффициентов регуляризации:

$$\alpha_i^{new} = \frac{D}{\boldsymbol{w}_i^T \boldsymbol{w}_i}$$

Таким образом, процедура обучения является итерационной. На каждой итерации для текущих значений $\boldsymbol{\alpha}$ с помощью ЕМ-алгоритма оцениваются $W, \boldsymbol{\mu}, \sigma^2$, а затем коэффициенты $\boldsymbol{\alpha}$ пересчитываются. В ЕМ-алгоритме формула для пересчета W выглядит следующим образом:

$$W_{new} = \left[\sum_{n=1}^{N} (\boldsymbol{x}_n - \bar{\boldsymbol{x}}) \mathbb{E}\boldsymbol{t}_n^T\right] \left[\sum_{n=1}^{N} \mathbb{E}\boldsymbol{t}_n \boldsymbol{t}_n^T + \sigma^2 A\right]^{-1}$$

Здесь $A = \mathrm{diag}(\alpha_1, \ldots, \alpha_D)$.

План лекции

Недостатки PCA

- Метод главных компонент способен находить только линейные подпространства исходного пространства, которые «объясняют» данные с высокой точностью. На практике поверхность, вдоль которой располагаются данные, может быть существенно нелинейной
- PCA является инвариантным относительно поворота координат в пространстве латентных переменных. Это означает, что восстановление значений латентных переменных является неоднозначным. В ряде случаев такая ситуация может быть неадекватной, например, в задаче разделения независимых источников, представленных линейной смесью с неизвестными коэффициентами.

Метод независимых компонент, ICA

Пусть наблюдаемые данные $\boldsymbol{x} = (x_1, \ldots, x_d)$ являются линейной комбинацией независимых источников $\boldsymbol{t} = (t_1, \ldots, t_d)$:

$$\boldsymbol{x} = W\boldsymbol{t},$$

причем размерности $\boldsymbol{x}$ и $\boldsymbol{t}$ совпадают, а матрица W неизвестна.

Если преобразование W является невырожденным, то $\boldsymbol{t} = W^T\boldsymbol{x}$. Рассмотрим задачу поиска одного источника t_1. Будем искать t_1 как линейную комбинацию наблюдаемых данных $t_1 = \boldsymbol{a}^T\boldsymbol{x}$. Задача – определить вектор $\boldsymbol{a}$.

Если $\boldsymbol{a}$ совпадает с колонкой матрицы W, то t_1 будет одним из искомых источников.

$$t_1 = \boldsymbol{a}^T\boldsymbol{x} = \boldsymbol{a}^T W\boldsymbol{t} = \{\boldsymbol{z} \triangleq W^T\boldsymbol{a}\} = \boldsymbol{z}^T\boldsymbol{t} = z_1t_1 + \cdots + z_dt_d$$

По центральной предельной теореме сумма случайных величин приближается к нормальному распределению. Следовательно, чем больше слагаемых в сумме $z_1t_1 + \cdots + z_dt_d$, тем более она похожа на гауссиану. Отсюда

$\boldsymbol{a}$ должен быть таким, чтобы $\boldsymbol{a}^T\boldsymbol{x}$ было как можно меньше похоже на гауссиану.

Критерии негауссовости

- Эксцесс.

$$\mathrm{kurt} = \mathbb{E}t^4 - 3(\mathbb{E}t^2)^2$$

  В предположении, что случайная величина t нормализована, $\mathrm{kurt} = \mathbb{E}t^4$. Для гауссианы значение эксцесса равно 0, для остальных распределений может быть как положительным, так и отрицательным. Поэтому в качестве меры негауссовости выбирают модуль или квадрат эксцесса.

- Негэнтропия.

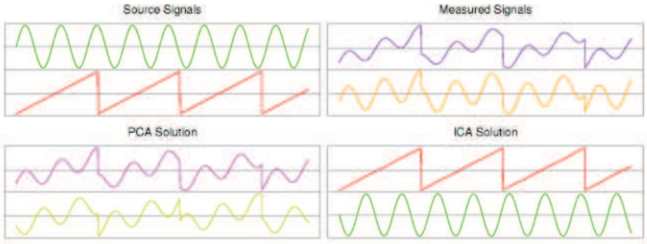
$$H(t) = -\int \log p(t)p(t)dt$$

  У гауссовского распределения энтропия $H(t)$ является максимальной среди всех распределений с одинаковой матрицей ковариации. Поэтому в качестве меры негауссовости используется негэнтропия

$$H(t_{gauss}) - H(t)$$

  Здесь t_{gauss} — гауссиана с той же матрицей ковариации, что и t. Негэнтропия является неотрицательной, и ее нужно максимизировать для поиска независимых источников.

ICA vs. PCA. Пример.



Анализ независимых факторов, Independent Factor Analysis, IFA

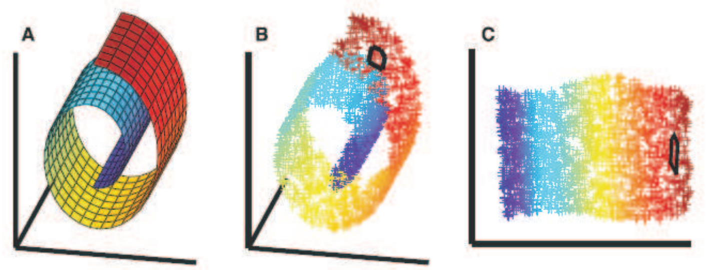
В методе PCA и факторном анализе в качестве априорного распределения латентных переменных предполагается выбирать нормальное распределение с центром в нуле и некоторой дисперсией, возможно, различной для разных направлений. В анализе независимых факторов в качестве априорного распределения предлагается выбирать независимое:

$$p(\boldsymbol{t}) = \prod_{i=1}^{d} p(\boldsymbol{t}_i), \qquad p(\boldsymbol{x}|\boldsymbol{t}) = \mathcal{N}(\boldsymbol{x}|W\boldsymbol{t} + \boldsymbol{\mu}, \Psi)$$

С одной стороны, данная модель является существенным обобщением PCA и позволяет находить, вообще говоря, нелинейные зависимости, с другой стороны, модель является достаточно простой, чтобы можно было предложить варианты EM-алгоритма для ее оптимизации.

Одним из возможных вариантов реализации данного метода является моделирование распределений отдельных факторов с помощью смеси гауссиан с разным числом компонент для разных факторов.

Локальное линейное погружение, Local Linear Embedding, LLE



Если данные располагаются вдоль некоторой поверхности, то разумно искать такое преобразование признакового пространства, при котором близкие объекты на этой поверхности были бы близки в новом пространстве. При этом близкие в евклидовом смысле объекты в результате такого преобразования могут оказаться очень далекими.

PCA не позволяет находить преобразования с подобными свойствами!

Схема LLE

Основная идея — искать преобразование с сохранением отношения соседства между объектами

- Восстановление структуры соседства на обучающей выборке:

$$\varepsilon(W) = \sum_{n=1}^{N} \|\boldsymbol{x}_n - \sum_{j \in Neighbours(\boldsymbol{x}_n)} W_{nj}\boldsymbol{x}_j\|^2 \to \min_W$$

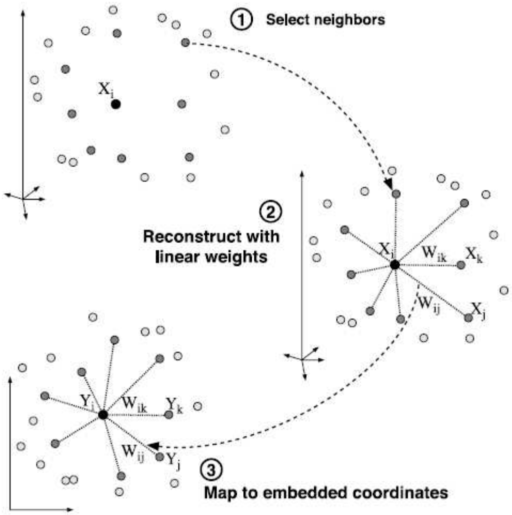
$$\sum_j W_{nj} = 0$$

- Поиск координат в новом пространстве со схожей структурой соседства:

$$\Phi(T) = \sum_{n=1}^{N} \|\boldsymbol{t}_n - \sum_{j \in Neighbours(\boldsymbol{x}_n)} W_{nj}\boldsymbol{t}_j\|^2 \to \min_T$$

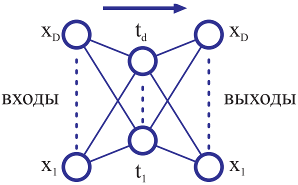
$$\sum_{n=1}^{N} \boldsymbol{t}_n = \boldsymbol{0}, \;\; \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{t}_n\boldsymbol{t}_n^T = I$$

Иллюстрация LLE



Автоассоциативные нейронные сети

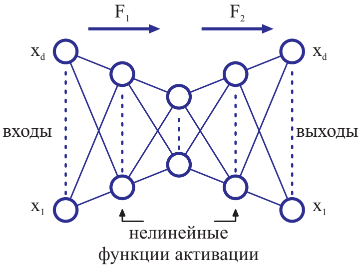
Рассмотрим многослойный персептрон следующего вида:

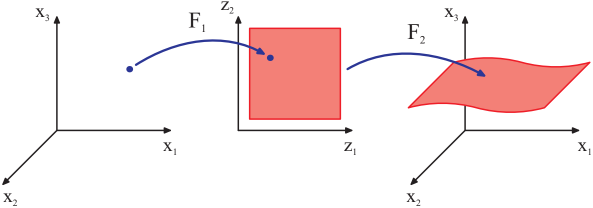


Здесь $d < D$. Задача нейронной сети — объяснить как можно точнее входные данные с помощью выходных. Таким образом, функционалом качества обучения является следующий:

$$E(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N} \|y(\boldsymbol{x}_n, \boldsymbol{w}) - \boldsymbol{x}_n\|^2$$

Если в сети только один скрытый слой, то даже с использованием нелинейной сигмоидной функцией активации результат обучения сети совпадает с результатом PCA.


F_1
F_2
x_d
x_d
входы
выходы
x_1
x_1
нелинейные
функции активации


x_3
F_1
z_2
F_2
x_3
x_1
z_1
x_1
x_2
x_2

Generative Topographic Mapping, GTM

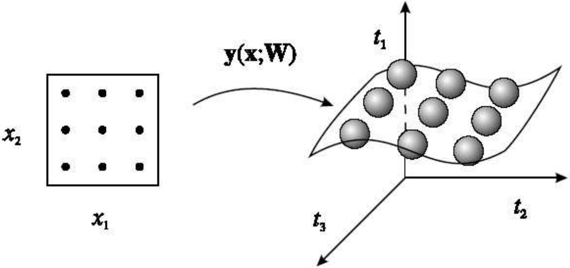
Метод GTM используется, в основном, для визуализации данных и является вероятностным обобщением самоорганизующихся сетей Кохонена.
В модели GTM предполагается, что данные образуются в результате нелинейного преобразования латентных переменных с добавлением нормального шума:

$$p(\boldsymbol{x}|\boldsymbol{t}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{y}(\boldsymbol{t}, W), \sigma^2 I)$$

Пространство латентных переменных является двухмерным, а распределение латентных переменных представляет собой сумму дельта-функций с центрами в некоторой регулярной сетке:

$$p(\boldsymbol{t}) = \frac{1}{l}\sum_{j=1}^{l} \delta(\boldsymbol{t} - \boldsymbol{t}_j)$$

Иллюстрация GTM



Функция правдоподобия выглядит следующим образом:

$$\log p(X|W,\sigma^2) = \sum_{n=1}^{N} \log \left[\frac{1}{l} \sum_{j=1}^{l} p(\boldsymbol{x}_n|\boldsymbol{t}_j, W, \sigma^2) \right]$$

Иллюстрация GTM

В качестве функции регрессии предлагается взять обобщенную линейную:

$$\boldsymbol{y}(\boldsymbol{t}, W) = W\boldsymbol{\phi}(\boldsymbol{t})$$

Здесь $\{\phi_j(\boldsymbol{t})\}_{j=1}^d$ — набор фиксированных базисных функций, в качестве которых могут выступать, например, гауссианы с центрами в регулярной сетке. Тогда можно предложить эффективный EM-алгоритм для оптимизации такой модели.