Курс: Байесовские методы машинного обучения, 2011

Байесовский метод главных компонент

Дата: 23 ноября 2011

Задача уменьшения размерности в данных

Рассмотрим задачу классификации изображений рукописных цифр MNIST[1]. Пусть имеется некоторое количество изображений в шкале серого цвета, на каждом из которых представлена одна цифра (см. рис. 1). Задача состоит в автоматическом определении цифры для входного изображения.

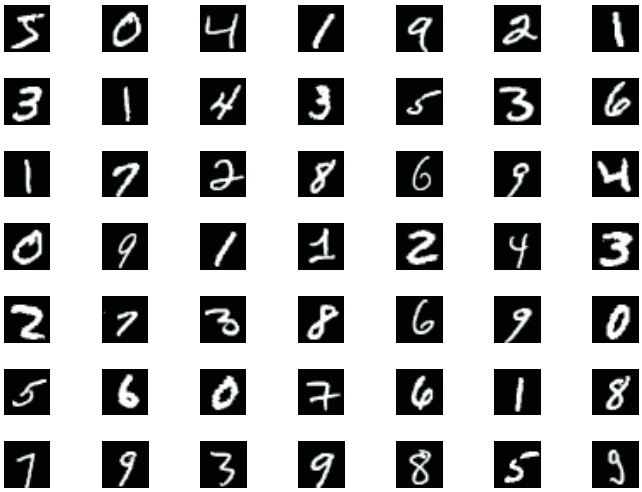
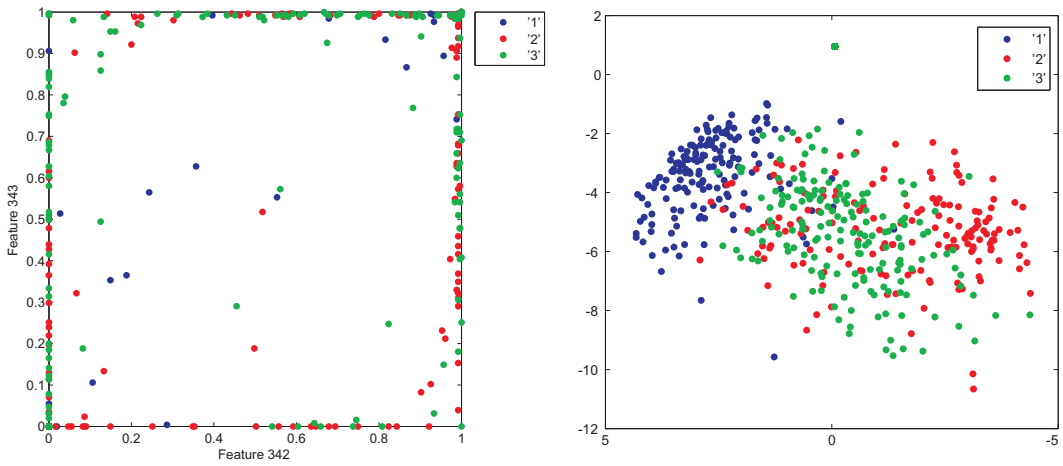


Рис. 1: Примеры изображений рукописных цифр из базы данных MNIST.

Для применения методов распознавания в данной задаче необходимо предварительно выбрать пространство признаков, характеризующее изображения цифр. В простейшем случае в качестве признаков можно взять исходные интенсивности пикселов изображения. Тогда для изображения размера 28×28 получаем 784 признака. Такой способ формирования признакового пространства обладает рядом недостатков. Во-первых, размерность получаемого признакового пространства становится слишком большой. Например, для относительно небольших изображений размера 300×200 размерность пространства составит 60000 признаков. Большое количество признаков приводит к значительным 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 недостатком полученного признакового пространства является невыполнение гипотезы компактности (близкие в пространстве признаков объекты не соответствуют одним и тем же классам), см. рис. 2а. 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

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



(a) (b)

Рис. 2: 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).

метод главных компонент позволяет получить более качественное признаковое пространство (см. рис. 2b).

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 $d < D$. 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 больше 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. 2b);
- и др.

Заметим, что все 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. 2b, был получен без использования информации о цифрах.

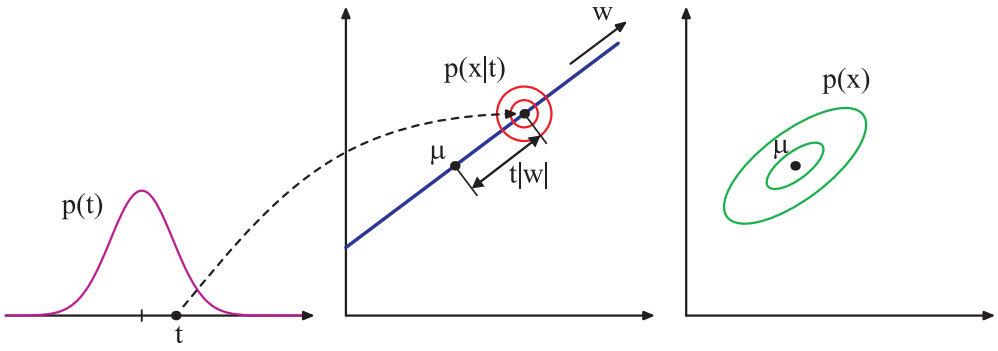


Рис. 3: Иллюстрация процесса генерации объекта в вероятностной модели PCA для $D = 2$ и $d = 1$. Наблюдаемое значение $\boldsymbol{x}$ образуется путем генерирования значения скрытой компоненты t из априорного распределения $p(t)$ и последующего генерирования значения $\boldsymbol{x}$ из изотропного нормального распределения с центром $\boldsymbol{\mu} + t\boldsymbol{w}$ и матрицей ковариации $\sigma^2 I$. Зеленые эллипсы показывают линии уровня плотности маргинального распределения $p(\boldsymbol{x})$.

Метод главных компонент

Метод главных компонент (Principal Component Analysis, PCA) является простейшим линейным методом уменьшения размерности в данных. Идея метода заключается в поиске в исходном пространстве $\mathbb{R}^D$ гиперплоскости заданной размерности d с последующим проектированием выборки на данную гиперплоскость. Вероятностную модель метода главных компонент можно сформулировать следующим образом:

$$p(\boldsymbol{x}|\boldsymbol{t}) = \mathcal{N}(\boldsymbol{x}|W\boldsymbol{t} + \boldsymbol{\mu}, \sigma^2 I), \quad p(\boldsymbol{t}) = \mathcal{N}(\boldsymbol{t}|\boldsymbol{0}, I).$$

Здесь матрица $W \in \mathbb{R}^{D\times d}$ задает направляющие вектора гиперплоскости, $\boldsymbol{\mu} \in \mathbb{R}^D$ – смещение гиперплоскости относительно начала координат, параметр $\sigma > 0$ определяет дисперсию шума в данных относительно гиперплоскости.

Процесс генерации объекта $\boldsymbol{x}$ в заданной вероятностной модели показан на рис. 3. Считая все объекты выборки X независимыми реализациями, приходим к следующему полному совместному распределению:

$$p(X, T|W, \boldsymbol{\mu}, \sigma) = \prod_{n=1}^{N} p(\boldsymbol{x}_n, \boldsymbol{t}_n|W, \boldsymbol{\mu}, \sigma) = \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{x}_n|W\boldsymbol{t}_n + \boldsymbol{\mu}, \sigma^2 I)\mathcal{N}(\boldsymbol{t}_n|\boldsymbol{0}, I).$$

Для поиска значений параметров модели (гиперплоскости) воспользуемся методом максимального правдоподобия:

$$p(X|W, \boldsymbol{\mu}, \sigma) = \prod_{n=1}^{N} p(\boldsymbol{x}_n|W, \boldsymbol{\mu}, \sigma) \to \max_{W,\boldsymbol{\mu},\sigma}. \tag{1}$$

Маргинальное распределение $p(\boldsymbol{x}_n|W, \boldsymbol{\mu}, \sigma)$ представляет собой свертку двух нормальных распределений и может быть вычислено аналитически:

$$p(\boldsymbol{x}_n|W, \boldsymbol{\mu}, \sigma) = \int p(\boldsymbol{x}_n|\boldsymbol{t}_n, W, \boldsymbol{\mu}, \sigma)p(\boldsymbol{t}_n)d\boldsymbol{t}_n =$$

$$= \int \mathcal{N}(\boldsymbol{x}_n|W\boldsymbol{t}_n + \boldsymbol{\mu}, \sigma^2 I)\mathcal{N}(\boldsymbol{t}_n|\boldsymbol{0}, I)d\boldsymbol{t}_n = \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}, \sigma^2 I + WW^T).$$

Таким образом, модель PCA представляет собой нормальное распределение, в котором матрица ковариации задается специальным образом:

$$C = WW^T + \sigma^2 I. \tag{2}$$

Заметим, что модель PCA инвариантна относительно выбора базиса в гиперплоскости. Пусть $R \in \mathbb{R}^d$ – произвольная ортогональная матрица, задающая поворот базиса гиперплоскости. Это соответствует использованию матрицы $\widetilde{W} = WR$. Тогда матрица ковариации равна

$$C = \widetilde{W}\widetilde{W}^T + \sigma^2 I = WRR^TW^T + \sigma^2 I = WW^T + \sigma^2 I.$$

Таким образом, матрица ковариации не зависит от R.

Вернемся теперь к задаче оптимизации (1). Дифференцируя логарифм совместного распределения $p(X, T|W, \boldsymbol{\mu}, \sigma)$ по параметрам $W, \boldsymbol{\mu}, \sigma$ и приравнивая производные к нулю, получим следующее решение задачи:

$$\begin{aligned} &\boldsymbol{\mu} = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n, \;\; S = \frac{1}{N}\sum_{n=1}^{N} (\boldsymbol{x}_n - \boldsymbol{\mu})(\boldsymbol{x}_n - \boldsymbol{\mu})^T, \\ &W = Q(\Lambda - \sigma^2 I)^{1/2} R, \\ &\sigma^2 = \frac{1}{D-d}\sum_{i=d+1}^{D} \lambda_i. \end{aligned} \tag{3}$$

Здесь $Q = (\boldsymbol{q}_1|\ldots|\boldsymbol{q}_d) \in \mathbb{R}^{D\times d}$, $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_d$ – нормированные собственные вектора выборочной матрицы ковариации S, отвечающие наибольшим собственным значениям $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$, $\|\boldsymbol{q}_i\| = 1$, $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$, R – произвольная ортогональная матрица размера $d \times d$.

Рассмотрим подробнее решение (3). Гиперплоскость, наилучшим образом объясняющая данные, определяется собственными векторами выборочной матрицы ковариации. Нормировочные коэффициенты собственных векторов $\sqrt{\lambda_i - \sigma^2}$ в решении для W служат для корректного восстановления дисперсии данных по всем направлениям. Действительно, величина дисперсии нормально распределенных данных с матрицей ковариации C вдоль единичного направления $\boldsymbol{v} : \boldsymbol{v}^T\boldsymbol{v} = 1$ составляет $\boldsymbol{v}^T C\boldsymbol{v}$. Если $\boldsymbol{v}$ лежит в подпространстве, ортогональном гиперплоскости, то $\boldsymbol{v}^T C\boldsymbol{v} = \sigma^2$. Теперь пусть $\boldsymbol{v}$ совпадает с одним из собственных векторов $\boldsymbol{q}_i$. Тогда $\boldsymbol{v}^T C\boldsymbol{v} = \lambda_i - \sigma^2 + \sigma^2 = \lambda_i$. Таким образом, модель PCA корректно восстанавливает дисперсии данных в пространстве гиперплоскости и аппроксимирует дисперсию средним значением в ортогональном пространстве.

Зная параметры $W, \boldsymbol{\mu}, \sigma$, задача поиска для объекта $\boldsymbol{x}$ представления $\boldsymbol{t}$ в пространстве $\mathbb{R}^d$ сводится к вычислению математического ожидания условного распределения

$$\begin{aligned} &p(\boldsymbol{t}|\boldsymbol{x}) = \frac{p(\boldsymbol{t}, \boldsymbol{x})}{\int p(\boldsymbol{t}, \boldsymbol{x}) d\boldsymbol{x}} = \mathcal{N}(\boldsymbol{t}|(\sigma^2 I + W^TW)^{-1}W^T(\boldsymbol{x} - \boldsymbol{\mu}), I + \sigma^{-2}W^TW), \\ &\mathbb{E}_{\boldsymbol{t}|\boldsymbol{x}}\boldsymbol{t} = (\sigma^2 I + W^TW)^{-1}W^T(\boldsymbol{x} - \boldsymbol{\mu}). \end{aligned}$$

Матрица W состоит из собственных векторов выборочной матрицы ковариации. Известно, что собственные вектора, отвечающие различным собственным значениям, являются ортогональными. Остальные собственные вектора также можно выбрать ортогональными путем преобразования базиса гиперплоскости. Таким образом, матрица ковариации для компонент $\boldsymbol{t}$ $I + \sigma^{-2}W^TW$ является диагональной, следовательно, новые признаки являются некоррелированными.

EM-алгоритм для PCA

По аналогии с моделью смеси нормальных распределений, модель PCA можно рассматривать как модель со скрытыми переменными и решать задачу максимизации правдоподобия (1) с помощью итерационного EM-алгоритма:
E-шаг:

$$\begin{aligned}
&p(T|X, W_{old}, \sigma^2_{old}, \boldsymbol{\mu}_{old}) = \prod_{n=1}^{N} p(\boldsymbol{t}_n|\boldsymbol{x}_n, W_{old}, \sigma^2_{old}, \boldsymbol{\mu}_{old}),\\
&p(\boldsymbol{t}_n|\boldsymbol{x}_n, W_{old}, \sigma^2_{old}, \boldsymbol{\mu}_{old}) = \mathcal{N}(\boldsymbol{t}_n|\boldsymbol{\mu}_n, \Sigma_n),\\
&\boldsymbol{\mu}_n = M_{old} W^T_{old}(\boldsymbol{x}_n - \boldsymbol{\mu}_{old}),\\
&\Sigma_n = \sigma^2_{old} M_{old},\\
&M_{old} = (W^T_{old} W_{old} + \sigma^2_{old} I)^{-1}.
\end{aligned} \tag{4}$$

M-шаг:

$$\begin{aligned}
&\boldsymbol{\mu}_{new} = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n,\\
&W_{new} = \left(\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{\mu}_{new})\mathbb{E}\boldsymbol{t}^T_n\right)\left(\sum_{n=1}^{N}\mathbb{E}\boldsymbol{t}_n\boldsymbol{t}^T_n\right)^{-1},\\
&\sigma^2_{new} = \frac{1}{ND}\sum_{n=1}^{N}\Big((\boldsymbol{x}_n - \boldsymbol{\mu}_{new})^T(\boldsymbol{x}_n - \boldsymbol{\mu}_{new}) - 2\mathbb{E}\boldsymbol{t}^T_n W^T_{new}(\boldsymbol{x}_n - \boldsymbol{\mu}_{new}) + \mathrm{tr} W^T_{new} W_{new}\mathbb{E}\boldsymbol{t}_n\boldsymbol{t}^T_n\Big).
\end{aligned} \tag{5}$$

При этом необходимые статистики вычисляются следующим образом:

$$\begin{aligned}
&\mathbb{E}\boldsymbol{t}_n = \boldsymbol{\mu}_n,\\
&\mathbb{E}\boldsymbol{t}_n\boldsymbol{t}^T_n = \Sigma_n + \boldsymbol{\mu}_n\boldsymbol{\mu}^T_n.
\end{aligned}$$

Заметим, что в процессе EM-итераций значение $\boldsymbol{\mu}$ не меняется и равно выборочному среднему. Поэтому при практическом применении EM-алгоритма для PCA выборка X сначала центрируется, а затем все вычисления проводятся без пересчета $\boldsymbol{\mu}$.

EM-алгоритм сходится к решению для W (3) с некоторой неединичной матрицей R. Таким образом, полученные признаки не обладают свойством некоррелированности. Кроме того, столбцы W не являются, вообще говоря, ортогональными. Поэтому для получения ортогонального базиса требуется дополнительно после завершения EM-итераций проводить процесс ортогонализации Грамма-Шмидта.

EM-алгоритм для PCA является вычислительно более эффективным по сравнению с аналитическим решением (3) для больших выборок и в ситуациях, когда $d \ll D$. Действительно, вычисление выборочной матрицы ковариации требует $O(ND^2)$, а поиск ее собственных значений и собственных векторов – $O(D^3)$ или $O(N^3)$. В EM-алгоритме самые сложные операции требуют $O(NDd)$ и $O(d^3)$, что может дать существенный выигрыш при больших N и $d \ll D$.

Учет пропусков в данных

Одним из преимуществ использования EM-алгоритма для максимизации правдоподобия в модели PCA является возможность прямого обобщения метода на случай наличия пропусков в

данных. Обозначим через K_n множество номеров известных значений признаков для объекта $\boldsymbol{x}_n$ и U_n — множество пропущенных значений признаков для объекта $\boldsymbol{x}_n$, $K_n \cup U_n = \{1, \ldots, D\}$. Соответственно $W_{K_n} = \{w_{ij}\}_{i \in K_n, j \in \{1,\ldots,d\}}$. Вероятностная модель PCA с пропусками в данных выглядит следующим образом:

$$
\begin{aligned}
&p(X_K, X_U, T | W, \sigma^2, \boldsymbol{\mu}) = \prod_{n=1}^{N} p(\boldsymbol{x}_{n,K_n}, \boldsymbol{x}_{n,U_n} | \boldsymbol{t}_n, W, \boldsymbol{\mu}, \sigma^2) p(\boldsymbol{t}_n), \\
&p(\boldsymbol{x}_{n,K_n}, \boldsymbol{x}_{n,U_n} | \boldsymbol{t}_n, W, \boldsymbol{\mu}, \sigma^2) = \mathcal{N}((\boldsymbol{x}_{n,K_n}, \boldsymbol{x}_{n,U_n}) | (W_{K_n}\boldsymbol{t}_n + \boldsymbol{\mu}_{K_n}, W_{U_n}\boldsymbol{t}_n + \boldsymbol{\mu}_{U_n}); \sigma^2 I), \\
&p(\boldsymbol{t}_n) = \mathcal{N}(\boldsymbol{t}_n | \boldsymbol{0}, I).
\end{aligned}
$$

Нетрудно показать, что EM-алгоритм для этой модели состоит в следующем:
E-шаг:

$$
\begin{aligned}
&p(X_U, T | X_K, W_{old}, \sigma^2_{old}) = \prod_{n=1}^{N} p((\boldsymbol{x}_{n,U_n}, \boldsymbol{t}_n) | \boldsymbol{x}_{n,K_n}, W_{old}, \sigma^2_{old}), \\
&p((\boldsymbol{x}_{n,U_n}, \boldsymbol{t}_n) | \boldsymbol{x}_{n,K_n}, W_{old}, \sigma^2_{old}) = \mathcal{N}\Big((\boldsymbol{x}_{n,U_n}, \boldsymbol{t}_n) | \boldsymbol{m}_n, S_n\Big), \\
&\boldsymbol{m}_n = (W_{U_n} M W_{K_n}^T \boldsymbol{x}_{n,K_n}, M W_{K_n}^T \boldsymbol{x}_{n,K_n}), \\
&S_n = \sigma^2_{old} \begin{pmatrix} I + W_{U_n} M W_{U_n}^T & -W_{U_n} M \\ -M W_{U_n}^T & M \end{pmatrix}, \\
&M = (W_{K_n}^T W_{K_n} + \sigma^2_{old} I)^{-1}.
\end{aligned}
$$

M-шаг:

$$
\begin{aligned}
&W_i^{new} = \left(\sum_{n: i \in K_n} x_{ni} \mathbb{E}\boldsymbol{t}_n^T + \sum_{n: i \in U_n} \mathbb{E} x_{ni} \boldsymbol{t}_n^T \right) \left(\sum_{n=1}^{N} \mathbb{E}\boldsymbol{t}_n \boldsymbol{t}_n^T \right)^{-1}, \\
&\sigma^2_{new} = \frac{1}{ND} \sum_{n=1}^{N} \Big(\boldsymbol{x}_{n,K_n}^T \boldsymbol{x}_{n,K_n} + \mathrm{tr}\mathbb{E}\boldsymbol{x}_{n,U_n} \boldsymbol{x}_{n,U_n}^T - 2\mathbb{E}\boldsymbol{t}_n^T W_{K_n}^T \boldsymbol{x}_{n,K_n} - 2\mathrm{tr} W_{U_n}^T \mathbb{E}\boldsymbol{x}_{n,U_n} \boldsymbol{t}_n^T + \\
&\qquad + \mathrm{tr} W_{K_n}^T W_{K_n} \mathbb{E}\boldsymbol{t}_n \boldsymbol{t}_n^T + \mathrm{tr} W_{U_n}^T W_{U_n} \mathbb{E}\boldsymbol{t}_n \boldsymbol{t}_n^T \Big).
\end{aligned}
$$

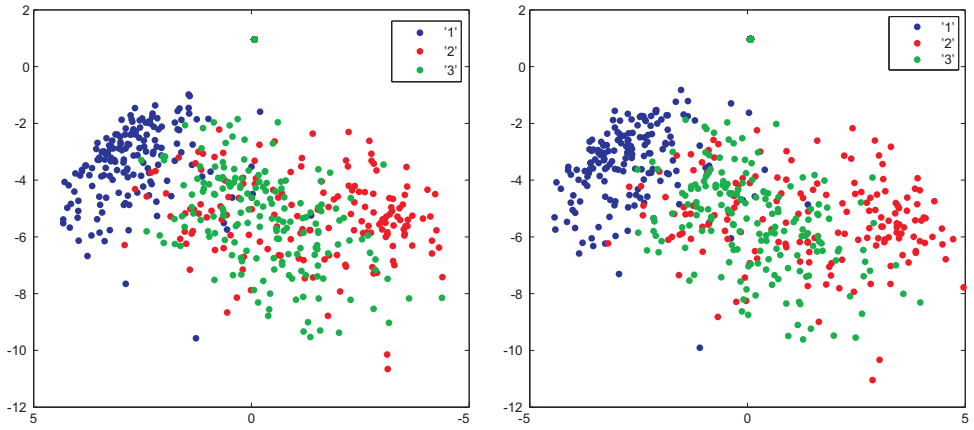
При этом выборка предварительно центрируется на величину

$$
\boldsymbol{\mu} : \mu_i = \frac{\sum_{n: i \in K_n}^{N} \boldsymbol{x}_{ni}}{\sum_{n: i \in K_n} 1}.
$$

Заметим, что формулы EM-алгоритма для модели PCA с пропусками переходят в соответствующие формулы EM-алгоритма для PCA в том случае, если пропусков в данных нет.

В качестве иллюстративного примера вернемся к задаче выбора признакового пространства для базы данных рукописных цифр MNIST. На рис. 4 приведена проекция исходной выборки на первые две главные компоненты (совпадает с рис. 2b), а также аналогичная проекция для выборки, в которой 30% случайно выбранных значений считаются пропущенными. Как видно, результаты практически совпадают между собой.

Рассмотренный метод учета пропусков в данных является адекватным для случая, когда места пропусков в данных определяются случайными факторами и, в частности, не зависят



(a) (b)

Рис. 4: Проекция выборки изображений цифр ’1’, ’2’, ’3’ на первые две главные компоненты для полных данных (a) и для выборки, в которой 30% случайно выбранных значений считаются пропущенными.

от истинных значений признаков в местах пропуска. Если, например, измерительный датчик дает сбой только для экстремальных значений измеряемой характеристики, то здесь необходима модификация вероятностной модели с пропусками, учитывающая модель образования пропущенных значений.

Байесовский метод главных компонент

Рассмотрим задачу автоматического выбора размерности редуцированного пространства d. Параметр d является типичным структурным параметром, выбор значения которого представляет собой компромисс между минимизацией размерности и минимизацией ошибки при проектировании выборки X в T.

Для решения задачи выбора d воспользуемся аналогией с методом релевантных векторов для задачи классификации с K классами и введем априорное распределение на матрицу W следующим образом:

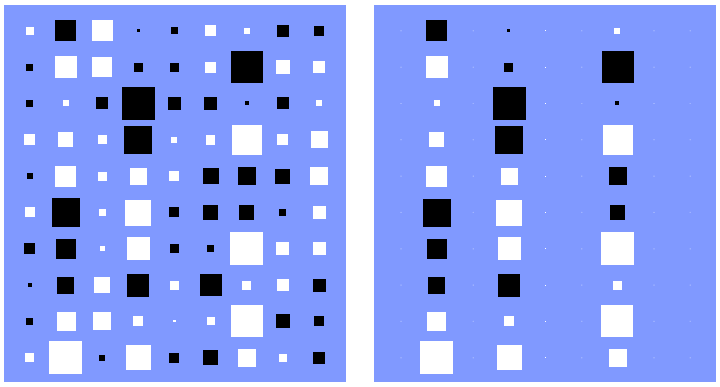
$$p(W|\boldsymbol{\alpha}) = \prod_{i=1}^{D} \sqrt{\frac{\alpha_i}{2\pi}}^{D} \exp\left(-\frac{\alpha_i}{2}\boldsymbol{w}_i^T\boldsymbol{w}\right).$$

Здесь $\boldsymbol{w}_i$ – i-ый столбец матрицы W (i-ый базисный вектор гиперплоскости), $\boldsymbol{\alpha}$ – параметры регуляризации для каждого базисного вектора. Устремление параметра α_i к плюс бесконечности соответствует бесконечному штрафу за любое отклонение $\boldsymbol{w}_i$ от нулевого вектора, что в свою очередь соответствует удалению базисного вектора из модели и снижению размерности d на единицу. Теперь полная вероятностная модель главных компонент выглядит как

$$p(X, T, W|\boldsymbol{\mu}, \sigma, \boldsymbol{\alpha}) = p(X|T, W, \boldsymbol{\mu}, \sigma)p(T)p(W|\boldsymbol{\alpha}).$$

Обучение этой модели с помощью метода максимального правдоподобия приводит к задаче

$$p(X|\boldsymbol{\mu}, \sigma, \boldsymbol{\alpha}) = \int p(X|W, \boldsymbol{\mu}, \sigma)p(W|\boldsymbol{\alpha})dW \to \max_{\boldsymbol{\mu},\sigma,\boldsymbol{\alpha}}.$$



(a) (b)

Рис. 5: Иллюстрация байесовского метода главных компонент. Модельные данные состоят из 300 объектов, сгенерированных из нормального распределения в пространстве размерности $D = 10$. При этом данные имеют стандартное отклонение 1 по трем направлениям в этом пространстве и стандартное отклоненение 0.5 по остальным семи направлениям. На рис. a показана матрица W, полученная с помощью стандартного метода главных компонент (белые и черные квадраты соответствуют положительным и отрицательным значениям, величина квадрата пропорциональна модулю значения). На рис. b показана матрица W, полученная с помощью байесовского метода главных компонент. Как видно, было выделено три направления, соответствующих загаданным направлениям с наибольшей дисперсией.

Эта задача может быть решена с помощью итерационного приближения Лапласа, где на каждой итерации сначала при текущих параметрах регуляризации $\boldsymbol{\alpha}$ находится максимум подынтегральной функции

$$(W_{MP}, \boldsymbol{\mu}_{MP}, \sigma_{MP}) = \arg\max_{W,\boldsymbol{\mu},\sigma} p(X|W, \boldsymbol{\mu}, \sigma) p(W|\boldsymbol{\alpha}), \tag{6}$$

а затем при текущих $(W_{MP}, \boldsymbol{\mu}_{MP}, \sigma_{MP})$ новые значения параметров регуляризации определяются как

$$\alpha_i^{new} = \frac{D}{\boldsymbol{w}_{MP,i}^T \boldsymbol{w}_{MP,i}}.$$

Задача оптимизации (6) может быть решена с помощью EM-алгоритма (4)-(5), где формула пересчета для W меняется на следующую:

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

Здесь $A = \mathrm{diag}(\alpha_1, \ldots, \alpha_D)$. Пример применения байесовского метода главных компонент показан на рис. 5.

Вероятностная смесь главных компонент

Одним из ограничений метода главных компонент является его линейность. В том случае, если выборка данных образует скрытую поверхность, которая является существенно нелинейной,

метод главных компонент может приводить к неадекватным результатам (большая ошибка при восстановлении данных или маленькая редукция размерности пространства). Простым обобщением метода главных компонент, которое позволяет преодолеть это ограничение, является рассмотрение вероятностной смеси главных компонент.

Рассмотрим следующую вероятностную модель:

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k p_k(\boldsymbol{x}|W_k, \sigma_k^2, \boldsymbol{\mu}_k) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, W_k W_k^T + \sigma_k^2 I),\ \sum_{k=1}^{K} \pi_k = 1,\ \pi_k \geq 0.$$

Эта модель представляет собой смесь нормальных распределений, в которой матрицы ковариации задаются специальным образом. Введем эквивалентную вероятностную модель путем добавления скрытых переменных $z_n \in \{1, \ldots, K\}$ для каждого объекта $\boldsymbol{x}_n$, отвечающих за номер компоненты смеси:

$$p(z_n = k) = \pi_k,$$
$$p(\boldsymbol{x}_n|z_n = k) = p_k(\boldsymbol{x}_n).$$

Можно показать, что EM-алгоритм максимизации правдоподобия в этой модели по параметрам $\boldsymbol{\pi}$, $M = \{\boldsymbol{\mu}_k\}_{k=1}^K$, $\mathcal{W} = \{W_k\}_{k=1}^K$ и $\boldsymbol{\sigma}$ выглядит следующим образом:
E-шаг:

$$\gamma_{nk} = p(z_n = k|\boldsymbol{x}_n, \boldsymbol{\pi}, M, \mathcal{W}, \boldsymbol{\sigma}) = \frac{\pi_k p_k(\boldsymbol{x}_n|W_k, \sigma_k^2, \boldsymbol{\mu}_k)}{\sum_{j=1}^K \pi_j p_j(\boldsymbol{x}_n|W_j, \sigma_j^2, \boldsymbol{\mu}_j)}.$$

M-шаг:

$$\pi_k^{new} = \frac{1}{N} \sum_{n=1}^{N} \gamma_{nk},$$
$$\boldsymbol{\mu}_k^{new} = \frac{\sum_{n=1}^N \gamma_{nk} \boldsymbol{x}_n}{\sum_{n=1}^N \gamma_{nk}},$$
$$S_k = \frac{\sum_{n=1}^N \gamma_{nk} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T}{\sum_{n=1}^N \gamma_{nk}}.$$

При этом параметры W_k и σ_k вычисляются по формулам (3) путем разложения по собственным векторам матрицы S_k. Можно также вывести альтернативные формулы пересчета этих параметров без привлечения промежуточной матрицы ковариации S_k с помощью EM-алгоритма. Эти формулы также можно легко обобщить на случай наличия пропусков в данных.

Заметим, что формулы для γ_{nk}, π_k и $\boldsymbol{\mu}_k$ полностью совпадают с соответствующими формулами EM-алгоритма для разделения смеси нормальных распределений.

Восстановление вероятностной смеси главных компонент соответствует построению K линейных подпространств, определяемых параметрами $W_k, \boldsymbol{\mu}_k, \sigma_k$. Таким образом, для заданного объекта $\boldsymbol{x}_n$ можно получить проекцию $\boldsymbol{t}_n$ на подпространство с номером k по формуле $(\sigma_k^2 I + W_k^T W_k)^{-1} W_k^T (\boldsymbol{x}_n - \boldsymbol{\mu}_k)$. При этом по аналогии с применением смеси нормальных распределений для решения задачи кластеризации, номер подпространства k выбирается как

$$k = \arg\max_k \gamma_{nk}. \tag{7}$$

С точки зрения задачи уменьшения размерности в данных, для каждого объекта $\boldsymbol{x}_n$ сохраняется номер подпространства k и проекция объекта на это подпространство $\boldsymbol{t}_n$.

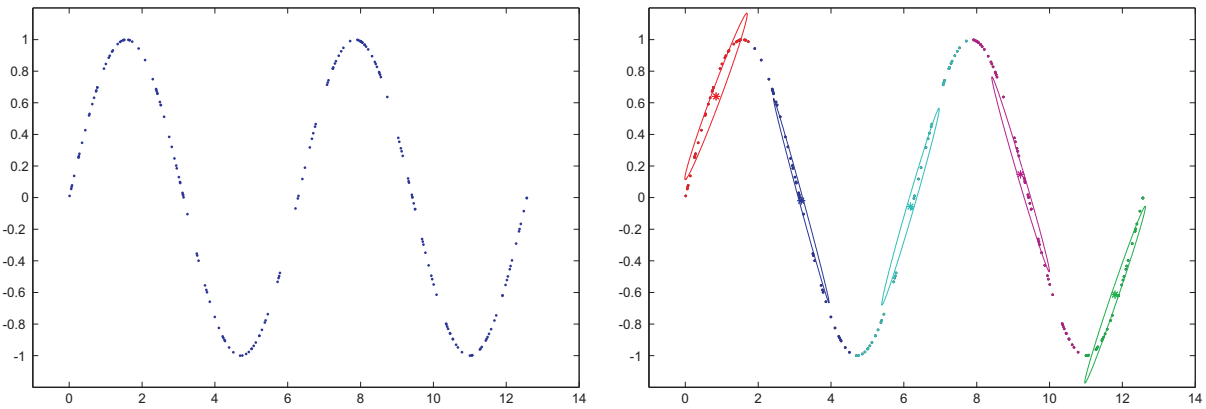


Рис. 6: Кластеризация двухмерной выборки (рис. a) на 5 кластеров с помощью смеси главных компонент (рис. b). Цветами обозначены объекты соответствующих кластеров. Кроме того, показаны центры и эллипсы рассеивания для каждой компоненты смеси.

Применение смеси главных компонент

Модель смеси главных компонент имеет широкую область применения. Помимо решения задачи уменьшения размерности и сжатия данных, эту модель можно использовать для решения задачи кластеризации и для восстановления плотности распределения выборки.

Рассмотрим применение модели смеси главных компонент для задачи кластеризации. Как уже было отмечено выше, эта модель является частным случаем общей модели смеси нормальных распределений, в которой матрица ковариации для каждой компоненты задается специальным образом: $C_k = W_k W_k^T + \sigma_k^2 I$. Как и в общем случае, для решения задачи кластеризации на K кластеров сначала восстанавливаются параметры модели смеси $\{W_k, \boldsymbol{\mu}_k, \sigma_k\}_{k=1}^K$ с помощью описанного выше EM-алгоритма, а затем номер кластера для объекта $\boldsymbol{x}_n$ определяется с помощью формулы (7). Пример применения смеси главных компонент для решения задачи кластеризации показан на рис. 6.

Матрица ковариации в модели смеси главных компонент требует задания $dD + 1$ параметров, а произвольная симметричная неотрицательно определенная матрица размера $D \times D$ задается $D(D+1)/2$ параметрами. Таким образом, модель смеси главных компонент имеет смысл применять для решения задачи кластеризации в том случае, когда восстановление смеси произвольных нормальных распределений не представляется возможным в силу ограниченности выборки. Кроме того, при применении смеси произвольных нормальных распределений кластеры представляют собой компактные шарообразные формы, а то время как в смеси главных компонент кластеры образуют объекты, лежащие в одном линейном подпространстве заданной размерности.

Другим возможным применением модели смеси главных компонент является блочное сжатие изображений. Пусть имеется некоторое изображение в шкале серого цвета (см. рис. 7a). Разобьем это изображение на набор непересекающихся блоков размера 8×8, и каждый блок вытянем в вектор длины 64. Таким образом, мы получим некоторую выборку размера <Число_блоков>×64. Например, для изображения размера 304×200 соответствующая выборка будет иметь размер 950×64. Применим к этой выборке методы уменьшения размерности в данных для решения задачи сжатия изображения. На рис. 7 приведен пример применения метода блочного сжатия изображения с помощью метода главных компонент с $d = 6$ (см. рис. 7b)

Рис. 7: Иллюстрация сжатия изображения (рис. a) в 10 раз с помощью метода главных компонент (рис. b) и смеси главных компонент (рис. c).

и смеси главных компонент с $d = 5$ и $K = 20$ (см. рис. 7c). В обоих случаях коэффициент сжатия составляет порядка 10 (в смеси главных компонент помимо проекции на подпространство сохраняется дополнительно номер этого подпространства). Как видно из рисунка, смесь главных компонент обеспечивает меньшую величину ошибки и, соответственно, более высокое качество восстановления изображения.

Стоит отдельно подчеркнуть, что рассмотренный метод является скорее иллюстративным примером к вероятностной модели смеси главных компонент, чем реальным методом сжатия изображений, т.к., например, он никак не учитывает специфику предметной области и особен-

ности реалистичных изображений.

Еще одним примером применения модели смеси главных компонент является восстановление плотности классов при решении задачи классификации. Пусть имеется задача классификации на K классов. Восстановим по обучающей выборке плотность каждого из классов $p(\boldsymbol{x}|k)$ с помощью вероятностной модели смеси главных компонент. После этого можно воспользоваться байесовским классификатором и классифицировать объекты по следующему правилу:

$$\hat{k}(\boldsymbol{x}) = \arg\max_k p(k|\boldsymbol{x}) = \arg\max_k p(\boldsymbol{x}|k)p(k).$$

Здесь $p(k)$ – априорная вероятность появления класса k. Заметим, что решение задачи классификации с помощью восстановления плотности каждого из классов требует большого объема обучающей выборки. Как уже было отмечено выше, модель смеси главных компонент задается значительно меньшим числом параметров, чем модель смеси произвольных нормальных распределений. В результате для применения модели смеси главных компонент требуется меньший объем обучающей выборки.