

# Семинары по методу главных компонент

Евгений Соколов  
sokolov.evg@gmail.com

24 ноября 2013 г.

## 1 Метод главных компонент

В машинном обучении часто возникает задача уменьшения размерности признакового пространства. Одним из подходов к ее решению является поиск новых признаков, каждый из которых является линейной комбинацией исходных признаков. В случае использования квадратичной функции ошибки при поиске такого приближения получается *метод главных компонент* (principal component analysis, PCA), о котором и пойдет речь.

### §1.1 Векторное дифференцирование

Выведем некоторые формулы векторного дифференцирования, которые понадобятся нам в дальнейшем.

*Следом* квадратной матрицы  $A \in \mathbb{R}^{n \times n}$ ,  $A = (a_{ij})_{i,j=1}^n$  называется сумма ее диагональных элементов:

$$\text{tr } A = \sum_{i=1}^n a_{ii}.$$

*Нормой Фробениуса* матрицы  $A \in \mathbb{R}^{m \times n}$  называется величина

$$\|A\| = \sqrt{\sum_{i=1}^m \sum_{j=1}^n a_{ij}^2}.$$

Нам пригодится следующее соотношение.

**Задача 1.1.** *Покажите, что*

$$\|A\|^2 = \text{tr } A^T A.$$

**Решение.**

$$\text{tr } A^T A = \sum_{i=1}^n (A^T A)_{ii} = \sum_{i=1}^n \sum_{j=1}^m a_{ij}^T a_{ji} = \sum_{i=1}^n \sum_{j=1}^m a_{ji}^2 = \|A\|^2.$$

■

**Задача 1.2.** Покажите, что матрицы  $A \in \mathbb{R}^{m \times n}$  и  $A \in \mathbb{R}^{n \times m}$ , можно переставлять под знаком следа:

$$\operatorname{tr} AB = \operatorname{tr} BA.$$

**Решение.** Легко доказывается путем расписывания левой и правой частей равенства. ■

Отсюда вытекает *циклическое свойство* следа:

$$\operatorname{tr} ABC = \operatorname{tr} CAB = \operatorname{tr} BCA,$$

при условии, что размерности матриц допускают такие перестановки.

Пусть  $f : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}$  — вещественнозначная функция, заданная на пространстве матриц. Производная этой функции по матрице определяется как матрица производных по отдельным элементам

$$\nabla_X f(X) = \begin{pmatrix} \frac{\partial f}{\partial x_{11}} & \cdots & \frac{\partial f}{\partial x_{1n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial x_{m1}} & \cdots & \frac{\partial f}{\partial x_{mn}} \end{pmatrix}$$

**Задача 1.3.** Покажите, что для матриц  $X \in \mathbb{R}^{m \times n}$  и  $A \in \mathbb{R}^{n \times m}$  выполнено

$$\nabla_X \operatorname{tr} XA = A^T.$$

**Решение.** Найдем производную по  $x_{ij}$ :

$$\frac{\partial}{\partial x_{ij}} \operatorname{tr} XA = \frac{\partial}{\partial x_{ij}} \sum_{i=1}^m \sum_{j=1}^n x_{ij} a_{ji} = a_{ji}.$$

Получаем, что

$$\nabla_X \operatorname{tr} XA = A^T. \quad \blacksquare$$

**Задача 1.4.** Покажите, что

$$\nabla_X \operatorname{tr} AXB = A^T B^T.$$

**Решение.** Пользуясь циклическим свойством и предыдущей задачей, получаем

$$\nabla_X \operatorname{tr} AXB = \nabla_X \operatorname{tr} XBA = (BA)^T = A^T B^T. \quad \blacksquare$$

Также нам понадобится следующая формула, которую мы оставим без доказательства:

$$\nabla_X \operatorname{tr} BX^T X B^T = 2XB B^T.$$

## §1.2 Метод главных компонент как матричное разложение

Пусть  $X \in \mathbb{R}^{\ell \times D}$  — матрица «объекты-признаки», где  $\ell$  — число объектов, а  $D$  — число признаков. Поставим задачу уменьшить размерность пространства до  $d$ . Новую матрицу «объекты-признаки» обозначим через  $Z \in \mathbb{R}^{\ell \times d}$ . Потребуем, чтобы новые признаки линейно зависели от исходных:

$$x_{ij} = \sum_{s=1}^d z_{is} u_{js},$$

или, в векторном виде,  $x_i = z_i U^T$  (здесь мы ввели матрицу перехода  $U \in \mathbb{R}^{D \times d}$ ). Данные уравнения не могут быть выполнены точно при  $d < \text{rk } X$ , поэтому потребуем, чтобы левая и правая части равенств были как можно ближе друг к другу с точки зрения квадратичного отклонения:

$$F = \sum_{i=1}^{\ell} \|x_i - z_i U^T\|^2 = \|X - ZU^T\|^2 \rightarrow \min_{Z, U}. \quad (1.1)$$

Таким образом, мы пришли к задаче представления матрицы  $X$  в виде произведения двух матриц меньшей размерности. Эта задача называется *задачей матричного разложения*. В данном случае мы ищем приближение, оптимальное в смысле нормы Фробениуса, однако могут использоваться и другие нормы или метрики.

Везде далее мы будем предполагать, что матрицы  $Z$  и  $U$  невырождены, потому что иначе размерность нового пространства  $d$  может быть уменьшена без потери качества.

Приступим к решению этой задачи. Перепишем функционал:

$$\begin{aligned} F &= \|X - ZU^T\|^2 = \\ &= \text{tr}(X^T - UZ^T)(X - ZU^T) = \\ &= \text{tr}(X^T X) - \text{tr}(X^T ZU^T) - \text{tr}(UZ^T X) + \text{tr}(UZ^T ZU^T) = \\ &= \{\text{tr } A^T = \text{tr } A\} = \\ &= \text{tr}(X^T X) - 2 \text{tr}(X^T ZU^T) - \text{tr}(UZ^T ZU^T). \end{aligned}$$

Воспользовавшись в последнем выражении тем, что  $\text{tr } A^T = \text{tr } A$  и  $\text{tr } AB = \text{tr } BA$ , можно получить эквивалентное представление:

$$\text{tr}(X^T X) - 2 \text{tr}(X^T ZU^T) - \text{tr}(UZ^T ZU^T) = \text{tr}(X^T X) - 2 \text{tr}(UZ^T X) - \text{tr}(ZU^T UZ^T).$$

Теперь, пользуясь выведенными выше формулами векторного дифференцирования и двумя последними представлениями функционала  $F$ , найдем его производные по  $Z$  и  $U$  и приравняем их к нулю:

$$\begin{aligned} -2XU + 2ZU^T U &= (ZU^T - X)U = 0; \\ -2X^T Z + 2UZ^T Z &= Z^T(ZU^T - X) = 0. \end{aligned}$$

Пользуясь невырожденностью матриц  $Z$  и  $U$ , получаем

$$\begin{cases} Z = XU(U^T U)^{-1}; \\ U = X^T Z(Z^T Z)^{-1}. \end{cases} \quad (1.2)$$

Заметим, что данные решения не позволяют найти решение в явном виде. Это логично, поскольку задача (1.1) не имеет единственного решения: если пара  $(Z, U)$  является решением, то решением является и пара  $(ZR, UR^{-T})$ <sup>1</sup> для любой невырожденной матрицы  $R$ . Чтобы преодолеть эту проблему, наложим на решение дополнительное ограничение: матрицы  $Z^T Z$  и  $U^T U$  должны быть диагональными.

Пусть  $(\tilde{Z}, \tilde{U})$  — произвольное решение задачи (1.1). Матрица  $\tilde{U}^T \tilde{U}$  невырожденная (как произведение невырожденных матриц), поэтому существует такая невырожденная матрица  $S$ , что

$$S^{-1} \tilde{U}^T \tilde{U} S^{-T} = I.$$

Матрица  $S^T \tilde{Z}^T \tilde{Z} S$  невырожденная и симметричная, поэтому существует такая ортогональная матрица  $T$ ,  $T^T T = T T^T = I$ , что

$$T^T (S^T \tilde{Z}^T \tilde{Z} S) T = \Lambda,$$

где  $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_d)$  — диагональная матрица.

Возьмем  $R = ST$  и рассмотрим решение  $(Z, U) = (\tilde{Z}R, \tilde{U}R^{-T})$ . Тогда:

$$\begin{aligned} Z^T Z &= R^T \tilde{Z}^T \tilde{Z} R = T^T S^T \tilde{Z}^T \tilde{Z} S T = \Lambda; \\ U^T U &= R^{-1} \tilde{U}^T \tilde{U} R^{-T} = T^{-1} \underbrace{S^{-1} \tilde{U}^T \tilde{U} S^{-T}}_I T^{-T} = T^{-1} T^{-T} = (T T^T)^{-1} = I. \end{aligned}$$

Таким образом, мы нашли такое решение, что матрицы  $U^T U$  и  $Z^T Z$  являются диагональными. Учтем это в уравнениях (1.2):

$$\begin{cases} Z = XU; & (1.3) \\ U\Lambda = X^T Z. & (1.4) \end{cases}$$

Подставляя (1.3) в (1.4), получаем уравнение

$$U\Lambda = X^T XU.$$

Это означает, что столбцы матрицы  $U$  являются собственными векторами матрицы  $X^T X$ , и соответствующими им собственными значениями являются числа  $\lambda_1, \dots, \lambda_d$ , стоящие на диагонали матрицы  $\Lambda$ .

Подставим теперь уравнение (1.4) в (1.3):

$$Z\Lambda = X X^T Z.$$

Это означает, что столбцы матрицы  $Z$  являются собственными векторами матрицы  $X X^T$ , и им соответствуют собственные значения  $\lambda_1, \dots, \lambda_d$ .

<sup>1</sup> Здесь под  $S^{-T}$  мы понимаем  $(S^T)^{-1} = (S^{-1})^T$ .

Подставим матрицы  $U$  и  $Z$  в функционал (1.1):

$$\begin{aligned}
F &= \|X - ZU^T\|^2 = \text{tr}(X^T - Z^T U)(X - ZU^T) = \\
&= \text{tr} X^T(X - ZU^T) - \text{tr}(Z^T U X - Z^T U ZU^T) = \{(1.3) \Rightarrow X = ZU^T\} = \\
&= \text{tr} X^T(X - ZU^T) - \underbrace{\text{tr}(Z^T U ZU^T - Z^T U ZU^T)}_{=0} = \\
&= \text{tr} X^T(X - ZU^T) = \\
&= \text{tr} X^T X - \text{tr} X^T ZU^T = \{(1.4) \Rightarrow X^T Z = U\Lambda\} = \\
&= \text{tr} X^T X - \text{tr} U\Lambda U^T = \\
&= \text{tr} X^T X - \text{tr} \underbrace{\Lambda U^T U}_{=I} = \\
&= \{\text{след матрицы равен сумме ее собственных значений}\} = \\
&= \sum_{i=1}^D \lambda_i - \sum_{i=1}^d \lambda_i = \\
&= \sum_{i=d+1}^D \lambda_i.
\end{aligned}$$

В предпоследнем равенстве мы воспользовались тем, что  $\{\lambda_i\}$  — собственные значения матрицы  $X^T X$ . Из последнего равенства заключаем, что минимум функционала (1.1) достигается, если матрица  $\Lambda$  состоит из  $d$  наибольших собственных значений матрицы  $X^T X$ .

Получаем следующий алгоритм для нахождения новых  $d$  признаков, линейно зависящих от исходных:

1. Найти собственное разложение матрицы  $X^T X$ :

$$X^T X = Q\Lambda Q^T;$$

2. Построить матрицу  $U$ , столбцы которой — собственные векторы, соответствующие  $d$  наибольшим собственным значениям из  $\Lambda$ ;
3. Перейти к новой матрице признаков, пользуясь уравнением (1.3):

$$Z = XU.$$

### 1.2.1 Метод главных компонент как поиск проекционной плоскости

Рассмотрим иной подход к поиску новых признаков: найдем такую  $d$ -мерную плоскость в признаковом пространстве, что ошибка проецирования обучающих объектов на нее будет минимальной.

Будем искать направляющие векторы плоскости  $u_1, \dots, u_d$ . Если они представляют собой ортонормированную систему, то проекция  $z$  вектора  $x$  на определяемую ими плоскость находится по формуле  $z = U^T x$ . Ошибка проецирования на плоскость определяется как норма разности между исходным вектором  $x$  и его проекцией  $z$ .

По теореме Пифагора <sup>2</sup> эта ошибка равна  $\|x\|^2 - \|z\|^2$  <sup>3</sup> Ошибка проецирования всей выборке записывается как

$$\begin{aligned} \sum_{i=1}^{\ell} \|x_i\|^2 - \sum_{i=1}^{\ell} \|U^T x_i\|^2 &= \sum_{i=1}^{\ell} \|x_i\|^2 - \sum_{i=1}^{\ell} \sum_{j=1}^d \langle u_j, x_i \rangle^2 = \\ &= \sum_{i=1}^{\ell} \|x_i\|^2 - \sum_{i=1}^{\ell} \sum_{j=1}^d u_j^T x_i x_i^T u_j = \sum_{i=1}^{\ell} \|x_i\|^2 - \sum_{j=1}^d u_j^T \left( \sum_{i=1}^{\ell} x_i x_i^T \right) u_j. \end{aligned}$$

Матрица  $S = \sum_{i=1}^{\ell} x_i x_i^T$  называется *выборочной матрицей ковариации*. Заметим, что от  $\{u_j\}$  зависит лишь второе слагаемое. Получаем, что поиск плоскости с минимальной ошибкой проецирования сводится к следующей задаче оптимизации:

$$\begin{cases} \sum_{j=1}^d u_j^T S u_j \rightarrow \max \\ \|u_j\|^2 = 1, \quad j = 1, \dots, d. \end{cases} \quad (1.5)$$

Мы не включили в эту задачу условие, что система векторов  $u_1, \dots, u_d$  должна быть ортогональной, оставив лишь условия на нормировку. Позже мы увидим, что полученное нами решение все равно будет представлять собой ортонормированную систему.

Выпишем лагранжиан задачи (1.5):

$$L = \sum_{j=1}^d u_j^T S u_j + \sum_{j=1}^d \lambda_j (1 - \|u_j\|^2).$$

Дифференцируя его и приравнивая к нулю, получаем

$$\nabla_{u_j} L = 2S u_j - 2\lambda_j u_j = 0 \quad \Rightarrow \quad S u_j = \lambda_j u_j.$$

Отсюда следует, что векторы  $u_j$  являются собственными векторами матрицы  $S = X^T X$ , а двойственные переменные  $\lambda_j$  — соответствующими им собственными значениями. Собственные векторы всегда определены с точностью до скалярного множителя, поэтому их всегда можно выбрать такими, что  $\|u_j\|^2 = 1$ . Подставим их в функционал задачи (1.5):

$$\sum_{j=1}^d u_j^T S u_j = \sum_{j=1}^d \lambda_j \underbrace{u_j^T u_j}_{=1} = \sum_{j=1}^d \lambda_j.$$

Таким образом, функционал достигнет своего максимума, если взять в качестве  $\{u_j\}$  собственные векторы матрицы  $S$ , соответствующие ее наибольшим  $d$  собственным значениям. Из линейной алгебры известно, что различные собственные векторы симметричной матрицы ортогональны друг другу, поэтому система  $\{u_j\}$  будет ортонормированной. Новые признаковые описания объектов находятся путем проецирования объектов на полученную плоскость:  $Z = XU$ .

<sup>2</sup> Если векторы  $v_1$  и  $v_2$  ортогональны, то  $\|v_1 + v_2\|^2 = \|v_1\|^2 + \|v_2\|^2$ .

<sup>3</sup> Строго говоря, неправильно говорить о разности между векторами  $x$  и  $z$ , поскольку они имеют разные размерности ( $D$  и  $d$  соответственно). Правильнее было бы дополнить векторы  $u_1, \dots, u_d$  до ортонормированного базиса и перевести вектор  $x$  в этот базис, получив вектор  $x_u$ . Его норма не изменилась бы, поскольку ортогональное преобразование сохраняет длину. Вектор  $z$  при этом следовало бы рассматривать как  $D$ -мерный, у которого координаты с  $d + 1$  по  $D$  равны нулю.