Лекция 16
Ядровые варианты PCA и SVM

Е. А. Соколов
ФКН ВШЭ

23 апреля 2017 г.

1 Ядровой метод главных компонент

Вспомним, что в методе главных компонент вычисляются собственные векторы $u_1, \dots, u_d$ ковариационной матрицы $X^T X$, соответствующие наибольшим собственным значениям. После этого новое признаковое описание объекта x вычисляется с помощью его проецирования на данные компоненты:

$$(\langle u_j, x\rangle)_{j=1}^d.$$

Попробуем теперь воспользоваться методом главных компонент в ядровом пространстве, где объекты описываются векторами $\varphi(x)$. Поскольку зачастую отображение $\varphi(x)$ нельзя выписать в явном виде, сформулируем метод главных компонент в терминах матрицы Грама $K = \Phi\Phi^T$ и ядра $K(x, z)$. Отметим, что напрямую пользоваться ковариационной матрицей $\Phi^T\Phi$ нельзя, поскольку она имеет размер $d \times d$, а число признаков d в спрямляющем пространстве может быть слишком большим; более того, спрямляющее пространство может быть бесконечномерным, и в этом случае ковариационную матрицу получить вообще не получится.

Пусть v_j — собственный вектор матрицы Грама K, соответствующий собственном значению λ_j. Рассмотрим цепочку уравнений:

$$\Phi^T\Phi(\Phi^T v_j) = \Phi^T(\Phi\Phi^T v_j) = \lambda_j \Phi^T v_j,$$

из которой следует, что $\Phi^T v_j$ является собственным вектором ковариационной матрицы $\Phi^T\Phi$, соответствующим собственному значению λ_j. Найдём норму данного вектора:

$$\|\Phi^T v_j\|^2 = v_j^T(\Phi\Phi^T v_j) = \lambda_j v_j^T v_j = \lambda_j,$$

где мы воспользовались нормированностью собственных векторов v_j. Значит, векторы $u_j = \lambda_j^{-1/2}\Phi^T v_j$ будут являться ортонормированной системой собственных векторов ковариационной матрицы.

Преобразуем выражение для u_j:

$$u_j = \lambda_j^{-1/2}\sum_{i=1}^{\ell}(v_j)_i\varphi(x_i) = \sum_{i=1}^{\ell}\alpha_{ji}\varphi(x_i),$$

где $\alpha_{ji} = \lambda_j^{-1/2} v_j$.

Мы выразили главные компоненты через признаковые описания объектов обучающей выборки в ядровом пространстве. Теперь найдём проекции объекта $\varphi(x)$ на эти компоненты:

$$\begin{aligned}\langle u_j, \varphi(x)\rangle &= \left\langle \sum_{i=1}^{\ell} \alpha_{ji}\varphi(x_i), \varphi(x) \right\rangle \\ &= \sum_{i=1}^{\ell} \alpha_{ji}\langle \varphi(x_i), \varphi(x)\rangle \\ &= \sum_{i=1}^{\ell} \alpha_{ji} K(x_i, x).\end{aligned}$$

Итак, мы выразили проекции на главные компоненты через ядро и через собственные векторы матрицы Грама — этого достаточно, чтобы вычислять проекции, не используя напрямую признаковые описания объектов из спрямляющего пространства.

2 Двойственный метод опорных векторов

§2.1 Условия Куна-Таккера

Рассмотрим задачу оптимизации

$$\begin{cases} f_0(x) \to \min\limits_{x\in\mathbb{R}^d} \\ f_i(x) \leqslant 0, \quad i = 1, \ldots, m, \\ h_i(x) = 0, \quad i = 1, \ldots, p. \end{cases} \tag{2.1}$$

Соответствующий ей лагранжиан имеет вид

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x),$$

где $\lambda_i \geqslant 0$.

Двойственной функцией для задачи (2.1) называется функция, получающаяся при взятии минимума лагранжиана по x:

$$g(\lambda, \nu) = \inf_x L(x, \lambda, \nu).$$

Можно показать, что данная функция всегда является вогнутой.

Зачем нужна двойственная функция? Оказывается, она дает нижнюю оценку на минимум в исходной оптимизационной задаче. Обозначим решение задачи (2.1) через x_*. Пусть x' — *допустимая* точка, т.е. $f_i(x') \leqslant 0$, $h_i(x') = 0$. Пусть также $\lambda_i > 0$. Тогда

$$L(x', \lambda, \nu) = f_0(x') + \sum_{i=1}^{m} \lambda_i f_i(x') + \sum_{i=1}^{p} \nu_i h_i(x') \leqslant f_0(x').$$

Если взять в левой части минимум по всем допустимым x, то неравенство останется верным; оно останется верным и в случае, если мы возьмем минимум по всем возможным x:

$$\inf_x L(x,\lambda,\nu) \leqslant \inf_{x\ -\ \text{допуст.}} L(x,\lambda,\nu) \leqslant L(x',\lambda,\nu).$$

Итак, получаем

$$\inf_x L(x,\lambda,\nu) \leqslant f_0(x').$$

Поскольку решение задачи x_* также является допустимой точкой, получаем, что при $\lambda \geqslant 0$ двойственная функция дает нижнюю оценку на минимум:

$$g(\lambda,\nu) \leqslant f_0(x_*).$$

Итак, двойственная функция для любой пары (λ,ν) с $\lambda > 0$ дает нижнюю оценку на минимум в оптимизационной задаче. Попробуем теперь найти наилучшую нижнюю оценку:

$$\begin{cases} g(\lambda,\nu) \to \max\limits_{\lambda,\nu} \\ \lambda_i \geqslant 0, \quad i = 1,\ldots,m. \end{cases} \tag{2.2}$$

Данная задача называется *двойственной* к задаче (2.1). Если известно решение данной задачи, то от условной задачи (2.1) можно перейти к безусловной.

Можно записать следующие условия, которые выполнены для решений прямой и двойственной задач x_* и (λ^*,ν^*):

$$\begin{cases} \nabla f_0(x_*) + \sum\limits_{i=1}^{m} \lambda_i^* \nabla f_i(x_*) + \sum\limits_{i=1}^{p} \nu_i^* \nabla h_i(x_*) = 0 \\ f_i(x_*) \leqslant 0, \quad i = 1,\ldots m \\ h_i(x_*) = 0, \quad i = 1,\ldots p \\ \lambda_i^* \geqslant 0, \quad i = 1,\ldots m \\ \lambda_i^* f_i(x_*) = 0, \quad i = 1,\ldots m \end{cases} \tag{KKT}$$

Данные выражения являются необходимыми и достаточными условиями для решения задачи (2.1). Отметим, что это верно лишь при выполнении ряда не очень сложных требований, обсуждение которых выходит за рамки нашей лекции.

§2.2 Вывод двойственной задачи SVM

Вспомним, что метод опорных векторов сводится к решению задачи оптимизации

$$\begin{cases} \dfrac{1}{2}\|w\|^2 + C\sum\limits_{i=1}^{\ell} \xi_i \to \min\limits_{w,b,\xi} \\ y_i\left(\langle w, x_i\rangle + b\right) \geqslant 1 - \xi_i, \quad i = 1,\ldots,\ell, \\ \xi_i \geqslant 0, \quad i = 1,\ldots,\ell. \end{cases} \tag{2.3}$$

Построим двойственную к ней. Запишем лагранжиан:

$$L(w,b,\xi,\lambda,\mu) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{\ell}\xi_i - \sum_{i=1}^{\ell}\lambda_i\left[y_i\left(\langle w,x_i\rangle + b\right) - 1 + \xi_i\right] - \sum_{i=1}^{\ell}\mu_i\xi_i.$$

Выпишем условия Куна-Таккера:

$$\nabla_w L = w - \sum_{i=1}^{\ell}\lambda_i y_i x_i = 0 \quad \Longrightarrow \quad w = \sum_{i=1}^{\ell}\lambda_i y_i x_i \tag{2.4}$$

$$\nabla_b L = -\sum_{i=1}^{\ell}\lambda_i y_i = 0 \quad \Longrightarrow \quad \sum_{i=1}^{\ell}\lambda_i y_i = 0 \tag{2.5}$$

$$\nabla_{\xi_i} L = C - \lambda_i - \mu_i \quad \Longrightarrow \quad \lambda_i + \mu_i = C \tag{2.6}$$

$$\lambda_i\left[y_i\left(\langle w,x_i\rangle + b\right) - 1 + \xi_i\right] = 0 \quad \Longrightarrow \quad (\lambda_i = 0) \text{ или } \left(y_i\left(\langle w,x_i\rangle + b\right) = 1 - \xi_i\right) \tag{2.7}$$

$$\mu_i\xi_i = 0 \quad \Longrightarrow \quad (\mu_i = 0) \text{ или } (\xi_i = 0) \tag{2.8}$$

$$\xi_i \geqslant 0, \lambda_i \geqslant 0, \mu_i \geqslant 0. \tag{2.9}$$

Проанализируем полученные условия. Из (2.4) следует, что вектор весов, полученный в результате настройки SVM, можно записать как линейную комбинацию объектов, причем веса в этой линейной комбинации можно найти как решение двойственной задачи. В зависимости от значений ξ_i и λ_i объекты x_i разбиваются на три категории:

1. $\xi_i = 0$, $\lambda_i = 0$.
   Такие объекты не влияют решение w (входят в него с нулевым весом λ_i), правильно классифицируются ($\xi_i = 0$) и лежат вне разделяющей полосы. Объекты этой категории называются *периферийными*.

2. $\xi_i = 0$, $0 < \lambda_i < C$.
   Из условия (2.7) следует, что $y_i\left(\langle w,x_i\rangle + b\right) = 1$, то есть объект лежит строго на границе разделяющей полосы. Поскольку $\lambda_i > 0$, объект влияет на решение w. Объекты этой категории называются *опорными граничными*.

3. $\xi_i > 0$, $\lambda_i = C$.
   Такие объекты могут лежать внутри разделяющей полосы ($0 < \xi_i < 2$) или выходить за ее пределы ($\xi_i \geqslant 2$). При этом если $0 < \xi_i < 1$, то объект классифицируется правильно, в противном случае — неправильно. Объекты этой категории называются *опорными нарушителями*.

Отметим, что варианта $\xi_i > 0$, $\lambda_i < C$ быть не может, поскольку при $\xi_i > 0$ из условия дополняющей нежесткости (2.8) следует, что $\mu_i = 0$, и отсюда из уравнения (2.6) получаем, что $\lambda_i = C$.

Итак, итоговый классификатор зависит только от объектов, лежащих на границе разделяющей полосы, и от объектов-нарушителей (с $\xi_i > 0$).

Построим двойственную функцию. Для этого подставим выражение (2.4) в лагранжиан, и воспользуемся уравнениями (2.5) и (2.6) (данные три уравнения выполнены для точки минимума лагранжиана при любых фиксированных λ и μ):

$$L = \frac{1}{2}\left\|\sum_{i=1}^{\ell}\lambda_i y_i x_i\right\|^2 - \sum_{i,j=1}^{\ell}\lambda_i\lambda_j y_i y_j\langle x_i, x_j\rangle - b\underbrace{\sum_{i=1}^{\ell}\lambda_i y_i}_{0} + \sum_{i=1}^{\ell}\lambda_i + \sum_{i=1}^{\ell}\xi_i\underbrace{(C-\lambda_i-\mu_i)}_{0}$$

$$= \sum_{i=1}^{\ell}\lambda_i - \frac{1}{2}\sum_{i,j=1}^{\ell}\lambda_i\lambda_j y_i y_j\langle x_i, x_j\rangle.$$

Мы должны потребовать выполнения условий (2.5) и (2.6) (если они не выполнены, то двойственная функция обращается в минус бесконечность), а также неотрицательность двойственных переменных $\lambda_i \geqslant 0$, $\mu_i \geqslant 0$. Ограничение на μ_i и условие (2.6), можно объединить, получив $\lambda_i \leqslant C$. Приходим к следующей двойственной задаче:

$$\begin{cases} \displaystyle\sum_{i=1}^{\ell}\lambda_i - \frac{1}{2}\sum_{i,j=1}^{\ell}\lambda_i\lambda_j y_i y_j\langle x_i, x_j\rangle \to \max\limits_{\lambda} \\ 0 \leqslant \lambda_i \leqslant C, \quad i = 1, \ldots, \ell, \\ \displaystyle\sum_{i=1}^{\ell}\lambda_i y_i = 0. \end{cases} \tag{2.10}$$

Она также является вогнутой, квадратичной и имеет единственный максимум.

Двойственная задача SVM зависит только от скалярных произведений объектов — отдельные признаковые описания никак не входят в неё. Значит, можно легко сделать ядровой переход:

$$\begin{cases} \displaystyle\sum_{i=1}^{\ell}\lambda_i - \frac{1}{2}\sum_{i,j=1}^{\ell}\lambda_i\lambda_j y_i y_j K(x_i, x_j) \to \max\limits_{\lambda} \\ 0 \leqslant \lambda_i \leqslant C, \quad i = 1, \ldots, \ell, \\ \displaystyle\sum_{i=1}^{\ell}\lambda_i y_i = 0. \end{cases} \tag{2.11}$$

Вернемся к тому, какое представление классификатора дает двойственная задача. Из уравнения (2.4) следует, что вектор весов w можно представить как линейную комбинацию объектов из обучающей выборки. Подставляя это представление w в классификатор, получаем

$$a(x) = \operatorname{sign}\left(\sum_{i=1}^{\ell}\lambda_i y_i\langle x_i, x\rangle + b\right). \tag{2.12}$$

Таким образом, классификатор измеряет сходство нового объекта с объектами из обучения, вычисляя скалярное произведение между ними. Это выражение также

зависит только от скалярных произведений, поэтому в нём тоже можно перейти к ядру.

В представлении (2.12) фигурирует переменная b, которая не находится непосредственно в двойственной задаче. Однако ее легко восстановить по любому граничному опорному объекту x_i, для которого выполнено $\xi_i = 0, 0 < \lambda_i < C$. Для него выполнено $y_i\left(\langle w, x_i\rangle + b\right) = 1$, откуда получаем

$$b = y_i - \langle w, x_i\rangle.$$

Как правило, для численной устойчивости берут медиану данной величины по всем граничным опорным объектам:

$$b = \operatorname{med}\{y_i - \langle w, x_i\rangle \,|\, \xi_i = 0, 0 < \lambda_i < C\}.$$