Лекция 14
Ядра в машинном обучении

Е. А. Соколов
ФКН ВШЭ

6 февраля 2020 г.

1 Ядровой SVM

Вспомним, что метод опорных векторов сводится к решению задачи оптимизации

$$\begin{cases} \dfrac{1}{2}\|w\|^2 + C\sum\limits_{i=1}^{\ell}\xi_i \to \min\limits_{w,b,\xi} \\ y_i\left(\langle w, x_i\rangle + b\right) \geqslant 1 - \xi_i, \quad i = 1, \ldots, \ell, \\ \xi_i \geqslant 0, \quad i = 1, \ldots, \ell. \end{cases} \tag{1.1}$$

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:

$$\nabla_w L = w - \sum_{i=1}^{\ell}\lambda_i y_i x_i = 0 \quad \Longrightarrow \quad w = \sum_{i=1}^{\ell}\lambda_i y_i x_i \tag{1.2}$$

$$\nabla_b L = -\sum_{i=1}^{\ell}\lambda_i y_i = 0 \quad \Longrightarrow \quad \sum_{i=1}^{\ell}\lambda_i y_i = 0 \tag{1.3}$$

$$\nabla_{\xi_i} L = C - \lambda_i - \mu_i \quad \Longrightarrow \quad \lambda_i + \mu_i = C \tag{1.4}$$

$$\lambda_i\left[y_i\left(\langle w, x_i\rangle + b\right) - 1 + \xi_i\right] = 0 \quad \Longrightarrow \quad (\lambda_i = 0) \text{ или } \left(y_i\left(\langle w, x_i\rangle + b\right) = 1 - \xi_i\right) \tag{1.5}$$

$$\mu_i\xi_i = 0 \quad \Longrightarrow \quad (\mu_i = 0) \text{ или } (\xi_i = 0) \tag{1.6}$$

$$\xi_i \geqslant 0, \lambda_i \geqslant 0, \mu_i \geqslant 0. \tag{1.7}$$

Проанализируем полученные условия. Из (1.2) 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 (1.5) следует, что $y_i(\langle w, x_i\rangle + b) = 1$, 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 (1.6) следует, что $\mu_i = 0$, и отсюда из уравнения (1.4) 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 (1.2) в лагранжиан, и воспользуемся уравнениями (1.3) и (1.4) (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 (1.3) и (1.4) (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 (1.4), можно объединить, получив $\lambda_i \leqslant C$. Приходим к следующей двойственной задаче:

$$\begin{cases} \displaystyle\sum_{i=1}^{\ell}\lambda_i - \frac{1}{2}\sum_{i,j=1}^{\ell}\lambda_i\lambda_j y_i y_j\langle x_i, x_j\rangle \to \max\limits_{\lambda} \\ 0 \leqslant \lambda_i \leqslant C, \quad i = 1, \ldots, \ell, \\ \displaystyle\sum_{i=1}^{\ell}\lambda_i y_i = 0. \end{cases} \tag{1.8}$$

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:

$$\begin{cases} \sum\limits_{i=1}^{\ell} \lambda_i - \dfrac{1}{2} \sum\limits_{i,j=1}^{\ell} \lambda_i \lambda_j y_i y_j K(x_i, x_j) \to \max\limits_{\lambda} \\ 0 \leqslant \lambda_i \leqslant C, \quad i = 1, \ldots, \ell, \\ \sum\limits_{i=1}^{\ell} \lambda_i y_i = 0. \end{cases} \tag{1.9}$$

Вернемся к тому, какое представление классификатора дает двойственная задача. Из уравнения (1.2) следует, что вектор весов w можно представить как линейную комбинацию объектов из обучающей выборки. Подставляя это представление w в классификатор, получаем

$$a(x) = \operatorname{sign}\left(\sum_{i=1}^{\ell} \lambda_i y_i \langle x_i, x \rangle + b \right). \tag{1.10}$$

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 (1.10) 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:

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

Связь с kNN. Если использовать гауссовское ядро (или, как его еще называют, RBF-ядро) в методе опорных векторов, то получится следующее решающее правило:

$$a(x) = \operatorname{sign} \sum_{i=1}^{\ell} y_i \lambda_i \exp\left(-\frac{\|x - x_i\|^2}{2\sigma^2} \right).$$

Вспомним теперь, что решающее правило в методе k ближайших соседей выглядит как

$$a(x) = \arg\max_{y \in Y} \Gamma_y(x, X^\ell); \quad \Gamma_y(x, X^\ell) = \sum_{i=1}^{\ell} [y_x^{(i)} = y] w(i, x),$$

где $w(i, x)$ — оценка важности i-го соседа для классификации объекта x, а $y_x^{(i)}$ — метка i-го ближайшего соседа. Для случая двух классов $\{+1, -1\}$ решающее правило

можно записать как знак разности оценок за эти классы:

$$a(x) = \operatorname{sign}\left(\Gamma_{+1}(x, X^\ell) - \Gamma_{-1}(x, X^\ell)\right) =$$
$$= \operatorname{sign}\left(\sum_{i=1}^{\ell}[y_x^{(i)} = +1]w(i,x) - \sum_{i=1}^{\ell}[y_x^{(i)} = -1]w(i,x)\right) =$$
$$= \operatorname{sign}\sum_{i=1}^{\ell}([y_x^{(i)} = +1] - [y_x^{(i)} = -1])w(i,x) =$$
$$= \operatorname{sign}\sum_{i=1}^{\ell} y_x^{(i)} w(i,x).$$

Заметим, что решающие правила метода опорных векторов с RBF-ядром и метода k ближайших соседей совпадут, если положить

$$w(i,x) = \lambda_{(i)} \exp\left(-\frac{\|x - x_{(i)}\|^2}{2\sigma^2}\right).$$

То есть SVM-RBF — это метод ℓ ближайших соседей, использующий гауссово ядро в качестве функции расстояния, и настраивающий веса объектов путем максимизации отступов.

2 Аппроксимация спрямляющего пространства

Все ядровые методы используют матрицу Грама $G = XX^T$ вместо матрицы «объекты-признаки» X. Это позволяет сохранять сложность методов при сколь угодно большой размерности спрямляющего пространства, но работа с матрицей Грама для больших выборок может стать затруднительной. Так, уже при выборках размером в сотни тысяч объектов хранение этой матрицы потребует большого количества памяти, а обращение станет трудоёмкой задачей, поскольку требует $O(\ell^3)$ операций.

Решением данной проблемы может быть построение в явном виде такого преобразования $\tilde{\varphi}(x)$, которое переводит объекты в пространство не очень большой размерности, и в котором можно напрямую обучать любые модели. Мы разберём метод случайных признаков Фурье (иногда также называется Random Kitchen Sinks) [1], который обладает свойством аппроксимации скалярного произведения:

$$\langle\tilde{\varphi}(x), \tilde{\varphi}(z)\rangle \approx K(x,z).$$

Из комплексного анализа известно, что любое непрерывное ядро вида $K(x,z) = K(x-z)$ является преобразованием Фурье некоторого вероятностного распределения (теорема Бохнера):

$$K(x-z) = \int_{\mathbb{R}^d} p(w) e^{iw^T(x-z)} dw.$$

Преобразуем интеграл:

$$\int_{\mathbb{R}^d} p(w) e^{iw^T(x-z)} dw = \int_{\mathbb{R}^d} p(w)\cos(w^T(x-z))dw + i\int_{\mathbb{R}^d} p(w)\sin(w^T(x-z))dw =$$
$$= \int_{\mathbb{R}^d} p(w)\cos(w^T(x-z))dw.$$

Поскольку значение ядра $K(x-z)$ всегда вещественное, то и в правой части мнимая часть равна нулю — а значит, остаётся лишь интеграл от косинуса $\cos w^T(x-z)$. Мы можем приблизить данный интеграл методом Монте-Карло:

$$\int_{\mathbb{R}^d} p(w)\cos(w^T(x-z))dw \approx \frac{1}{n}\sum_{j=1}^{n}\cos(w_j^T(x-z)),$$

где векторы $w_1,\ldots,w_n$ генерируются из распределения $p(w)$. Используя эти векторы, мы можем сформировать аппроксимацию преобразования $\varphi(x)$:

$$\tilde{\varphi}(x) = \frac{1}{\sqrt{n}}(\cos(w_1^T x),\ldots,\cos(w_n^T x),\sin(w_1^T x),\ldots,\sin(w_n^T x)).$$

Действительно, в этом случае скалярное произведение новых признаков будет иметь вид

$$\begin{aligned}\tilde{K}(x,z) = \langle\tilde{\varphi}(x),\tilde{\varphi}(z)\rangle &= \frac{1}{n}\sum_{j=1}^{n}\big(\cos(w_j^T x)\cos(w_j^T z)+\sin(w_j^T x)\sin(w_j^T z)\big)\\ &= \frac{1}{n}\sum_{j=1}^{n}\cos(w_j^T(x-z)).\end{aligned}$$

Данная оценка является несмещённой для $K(x,z)$ в силу свойств метода Монте-Карло. Более того, с помощью неравенств концентрации меры можно показать, что дисперсия данной оценки достаточно низкая. Например, для гауссова ядра будет иметь место неравенство

$$\mathbb{P}\left[\sup_{x,z}|\tilde{K}(x,z)-K(x,z)|\geqslant\varepsilon\right]\leqslant 2^8(2d\sigma^2/\varepsilon)^2\exp(-d\varepsilon^2/4(d+2)).$$

Разумеется, найти распределение $p(w)$ можно не для всех ядер $K(x-z)$. Как правило, данный метод используется для гауссовых ядер $\exp(\|x-z\|^2/2\sigma^2)$ — для них распределение $p(w)$ будет нормальным с нулевым матожиданием и дисперсией σ^2.

Список литературы

[1] *Rahimi, Ali and Recht, Benjamin* Random Features for Large-scale Kernel Machines. // Proceedings of the 20th International Conference on Neural Information Processing Systems, 2007.