

Лекция 16

Ядра в машинном обучении

Е. А. Соколов
ФКН ВШЭ

26 января 2018 г.

Ядра позволяют превращать линейные методы машинного обучения в нелинейные за счёт подмены признакового пространства. При этом, поскольку подмена производится через скалярное произведение, сложность методов не повышается. Мы уже знаем, как конструируются ядра, а также изучили несколько их распространённых примеров — например, полиномиальные и гауссовы ядра. Теперь мы перейдём к теоретическим вопросам и выясним, как выглядит оптимальный алгоритм в спрямляющем пространстве в общем виде. Далее мы обсудим вычислительные трудности, связанные с ядровыми методами, и разберём методы их устранения с помощью рандомизации. Наконец, мы разберём ещё одно применение ядер — а именно, в методе главных компонент.

1 Гильбертовы пространства и теорема о представлении

Рассмотрим гильбертово пространство (т.е. линейное векторное пространство со скалярным произведением, которое является полным) функций над объектами $H \subset \{f : \mathbb{X} \rightarrow \mathbb{R}\}$. Нас будут интересовать *гильбертовы пространства с воспроизводящими ядрами* (reproducing kernel Hilbert space, RKHS) — грубо говоря, это такие пространства функций, в которых результат применения функции $f \in H$ к объекту $x \in \mathbb{X}$ представим как скалярное произведение f на некоторый элемент пространства $\varphi(x) \in H$:

$$f(x) = \langle f, \varphi(x) \rangle$$

Заметим, что $\varphi(x)$ — это тоже функция из H ; значит, она представима в виде

$$\varphi(x)(z) = \langle \varphi(x), \varphi(z) \rangle$$

Данная функция $K : \mathbb{X} \times \mathbb{X} \rightarrow \mathbb{R}$ называется *воспроизводящим ядром* и является симметричной и положительно определённой. Из функционального анализа известно, что любому симметричному положительно определённому ядру соответствует некоторое гильбертово пространство с воспроизводящим ядром — а значит, при использовании ядер мы переводим объекты в некоторые функциональные гильбертовы пространства.

При использовании ядер мы строим линейную модель в спрямляющем пространстве — то есть модель вида $a(x) = \langle w, \varphi(x) \rangle$. Поскольку спрямляющее пространство может быть очень сложным (и даже бесконечномерным), есть риск, что мы не сможем найти w и, как следствие, $a(x)$ в явном виде. Тем не менее, до сих пор в линейной регрессии и в SVM оказывалось, что оптимальная модель имеет вид

$$a(x) = \sum_{i=1}^{\ell} \alpha_i K(x, x_i)$$

Оказывается, это правило является достаточно общим, и в большинстве случаев решение задачи машинного обучения будет иметь такой вид. Данный результат называется *теоремой о представлении* (representer theorem). Сформулируем и докажем её.

Теорема 1.1. Пусть $K(x, z)$ — симметричное положительно определённое ядро, соответствующее гильбертову пространству H . Пусть заданы функция потерь $L((x_1, y_1, a(x_1)), \dots, (x_\ell, y_\ell, a(x_\ell)))$ и регуляризатор $g(\|a\|)$, где $g : [0, +\infty) \rightarrow \mathbb{R}$ — монотонно возрастающая функция. Тогда решение задачи

$$a_* = \arg \min_{a \in H} \{L((x_1, y_1, a(x_1)), \dots, (x_\ell, y_\ell, a(x_\ell))) + g(\|a\|)\} \quad (1.1)$$

имеет вид

$$a_*(x) = \sum_{i=1}^{\ell} \alpha_i K(x, x_i).$$

Доказательство.

Рассмотрим базис, состоящий из элементов $\varphi(x_1), \dots, \varphi(x_\ell)$. Любой элемент гильбертова пространства $a \in H$ можно представить в виде суммы двух компонент: одна будет принадлежать линейной оболочке элементов $\varphi(x_1), \dots, \varphi(x_\ell)$, другая — ортогональному дополнению:

$$a = \sum_{i=1}^{\ell} \alpha_i \varphi(x_i) + v,$$

где $\langle v, \varphi(x_i) \rangle = 0$ для всех $i = 1, \dots, \ell$.

Как мы выяснили выше, применение функции a к объекту x равносильно вычислению скалярного произведения a на $\varphi(x)$. Тогда для объектов обучающей выборки будет выполнено

$$a(x_j) = \left\langle \sum_{i=1}^{\ell} \alpha_i \varphi(x_i) + v, \varphi(x_j) \right\rangle = \sum_{i=1}^{\ell} \alpha_i \langle \varphi(x_j), \varphi(x_i) \rangle.$$

Мы получили, что ответ модели на объектах обучающей выборки не зависит от v — значит, и значение функции потерь не зависит от v .

Теперь выясним, как элемент v влияет на регуляризатор. Воспользуемся его монотонностью и запишем неравенство:

$$\begin{aligned} g(\|a\|) &= g\left(\left\|\sum_{i=1}^{\ell} \alpha_i \varphi(x_i) + v\right\|\right) \\ &= g\left(\sqrt{\left\|\sum_{i=1}^{\ell} \alpha_i \varphi(x_i)\right\|^2 + \|v\|^2}\right) \\ &\geq g\left(\left\|\sum_{i=1}^{\ell} \alpha_i \varphi(x_i)\right\|\right) \end{aligned}$$

Мы получили, что зануление v приводит к уменьшению значения регуляризатора и никак не влияет на значение функции потерь. Значит, в решении задачи (1.1) компонента v всегда будет равна нулю. Отсюда получаем, что это решение имеет вид

$$a_*(x) = \sum_{i=1}^{\ell} \alpha_i K(x, x_i).$$

■

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) [2], 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:

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

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

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

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

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

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

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

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 .

3 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 собственному 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 $\alpha_{ji} = \lambda_j^{-1/2} (v_j)_i$.

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.

Список литературы

- [1] *Drineas, Petros and Mahoney, Michael W.* On the Nyström Method for Approximating a Gram Matrix for Improved Kernel-Based Learning. // Journal of Machine Learning Research, 2005.

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