Ядровые обобщения методов

В данной лекции рассматривается один из подходов к изменению признакового пространства — ядра, которые позволяют повышать размерность пространства без вычислительных трудностей.

1 Основные понятия

Ядром мы будем называть функцию $k(x, z)$, представимую в виде скалярного произведения в некотором пространстве: $k(x, z) = \langle \phi(x), \phi(z) \rangle$, где $\phi : \mathcal{X} \to H$ — отображение из исходного признакового пространства в некоторое *спрямляющее пространство*.

1.1 Построение ядер

Самый простой способ задать ядро — в явном виде построить отображение $\phi(x)$ в спрямляющее признаковое пространство. Тогда ядро определяется как скалярное произведение в этом пространстве: $k(x, z) = \langle \phi(x), \phi(z) \rangle$. При таком способе, однако, возникают проблемы с ростом вычислительной сложности, о которых уже было сказано выше.

Допустим, в качестве новых признаков мы хотим взять всевозможные произведения исходных признаков. Определим соответствующее отображение

$$\phi(x) = (x_i x_j)_{i,j=1}^{d} \in \mathbb{R}^{d^2}$$

и найдём ядро:

$$\begin{aligned}
k(x,z) = \langle \phi(x), \phi(z)\rangle = \langle (x_i x_j)_{i,j=1}^d, (z_i z_j)_{i,j=1}^d \rangle = \\
= \sum_{i,j=1}^d x_i x_j z_i z_j = \\
= \sum_{i=1}^d x_i z_i \sum_{j=1}^d x_j z_j = \\
= \langle x, z\rangle^2.
\end{aligned}$$

Таким образом, ядро выражается через скалярное произведение в исходном пространстве, и для его вычисления необходимо порядка d операций (в то время как прямое вычисление ядра потребовало бы $O(d^2)$ операций).

1.2 Неявное задание ядра

Пример с мономами показал, что можно определить ядро так, что оно не будет в явном виде использовать отображение объектов в новое признаковое пространство. Но как убедиться, что функция $k(x,z)$ определяет скалярное произведение в некотором пространстве? Ответ на этот вопрос даёт

Теорема 1 (Мерсер)**.** Функция $k(x,z)$ является ядром тогда и только тогда, когда:

1. Она симметрична: $k(x,z) = k(z,x)$.
2. Она неотрицательно определена, то есть для любой конечной выборки $\{x_i\}_{i=1}^{\ell}$ матрица $K = \big(k(x_i,x_j)\big)_{i,j=1}^{\ell}$ неотрицательно определена.

Проверять условия теоремы Мерсера, однако, может быть достаточно трудно. Поэтому для построения ядер, как правило, пользуются несколькими базовыми ядрами и операциями над ними, сохраняющими симметричность и неотрицательную определённость.

Теорема 2 ([2])**.** Пусть $k_1(x,z)$ и $k_2(x,z)$ — ядра, заданные на множестве X, $f(x)$ — вещественная функция на X, $\phi : X \to \mathbb{R}^N$ — векторная функция на X, k_3 — ядро, заданное на $\mathbb{R}^N$. Тогда следующие функции являются ядрами:

1. $k(x,z) = k_1(x,z) + k_2(x,z)$,
2. $k(x,z) = \alpha k_1(x,z)$, $\alpha > 0$,
3. $k(x,z) = k_1(x,z) k_2(x,z)$,
4. $k(x,z) = f(x) f(z)$,
5. $k(x,z) = k_3(\phi(x), \phi(z))$.

Доказательство.
Докажем третий пункт.

Пусть ядро k_1 соответствует отображению $\phi_1 : \mathbb{X} \to \mathbb{R}^{d_1}$, а ядро k_2 — отображению $\phi_2 : \mathbb{X} \to \mathbb{R}^{d_2}$. Определим новое отображение, которое соответствует всевозможным произведениям признаков из первого и второго спрямляющих пространств:

$$\phi_3(x) = \Big((\phi_1(x))_i \, (\phi_2(x))_j \Big)_{i,j=1}^{d_1,d_2}.$$

Соответствующее этому спрямляющему пространству ядро примет вид

$$\begin{aligned} k_3(x,z) &= \langle \phi_3(x), \phi_3(z) \rangle = \\ &= \sum_{i=1}^{d_1} \sum_{j=1}^{d_2} (\phi_3(x))_{ij} \, (\phi_3(z))_{ij} = \\ &= \sum_{i=1}^{d_1} (\phi_1(x))_i \, (\phi_1(z))_i \sum_{j=1}^{d_2} (\phi_2(x))_j \, (\phi_2(z))_j = \\ &= k_1(x,z) k_2(x,z). \end{aligned}$$

Таким образом, произведение двух ядер соответствует скалярному произведению в некотором спрямляющем пространстве, а значит является ядром. □

Теорема 3 ([3])**.** Пусть $k_1(x,z), k_2(x,z), \dots$ — последовательность ядер, причем предел

$$k(x,z) = \lim_{n \to \infty} k_n(x,z)$$

существует для всех x и z. Тогда $k(x,z)$ — ядро.

1.3 Примеры построения ядер

Рассмотрим некоторые примеры построения ядер.

1.3.1 Полиномиальные ядра

Пусть $p(x)$ — многочлен с положительными коэффициентами. Покажем, что $k(x, z) = p(\langle x, z \rangle)$ — ядро.

Пусть многочлен имеет вид

$$p(x) = \sum_{i=0}^{m} a_i x^i.$$

Будем доказывать требуемое утверждение по шагам.

1. $\langle x, z \rangle$ —ядро по определению ($\phi(x) = x$);
2. $\langle x, z \rangle^i$ — ядро как произведение ядер;
3. $a_i \langle x, z \rangle^i$ — ядро как произведение положительной константы на ядро;
4. константный член a_0 — ядро по пункту 4 теоремы 2, где $f(x) = \sqrt{a_0}$;
5. $\sum_{i=0}^{m} a_i \langle x, z \rangle^i$ — ядро как линейная комбинация ядер.

Также известно, что $p(k(x, z))$ — ядро для любого ядра $k(x, z)$.

Рассмотрим частный случай полиномиального ядра:

$$k_m(x, z) = \left(\langle x, z \rangle + R\right)^m, \ R > 0.$$

Распишем степень, воспользовавшись формулой бинома Ньютона:

$$k_m(x, z) = \sum_{i=0}^{m} C_m^i R^{m-i} \langle x, z \rangle^i.$$

Поскольку коэффициенты при скалярных произведениях $C_m^i R^{m-i}$ положительны, то данное ядро действительно является ядром. Если расписать скалярные произведения, то можно убедиться, что оно соответствует переводу набора признаков во всевозможные мономы над признаками степени не больше m, причем моном степени i имеет вес $\sqrt{C_m^i R^{m-i}}$.

Заметим, что параметр R контролирует относительный вес при мономах больших степеней. Например, отношение веса при мономе степени $m-1$ к весу при мономе первой степени равно

$$\sqrt{\frac{C_m^{m-1}R}{C_m^1 R^{m-1}}} = \sqrt{\frac{1}{R^{m-2}}},$$

то есть по мере увеличения R вес при мономах старших степеней будет становиться очень небольшим по сравнению с весом при остальных мономах. Можно сказать, что параметр R контролирует сложность модели.

1.3.2 Графовые ядра

Рассмотрим пару размеченных графов $G_1 = (V_1, E_1)$ и $G_2 = (V_2, E_2)$. *Прямым произведением размеченных графов* G_1 и G_2 называется граф $G_\times = (V_\times, E_\times)$, где $V_\times = \{(v_1, v_2) \in V_1 \times V_2 | \text{label}(v_1) = \text{label}(v_2)\}$ и $E_\times \{((u_1, u_2), (v_1, v_2)) \in V_\times^2 | (u_1, v_1) \in E_1, (u_2, v_2) \in E_2, \text{label}(u_1, v_1) = \text{label}(u_2, v_2)\}$.

Пусть $A_\times$ — матрица смежности прямого произведения G_1 и G_2. *Ядром прямого произведения графов* G_1 и G_2 называется

$$k_\times(G_1, G_2) = \sum_{i,j=1}^{|V_\times|} \left[\sum_{n=0}^{\infty} \lambda_n A_\times^n \right]_{ij},$$

где $\lambda_i \in \mathbb{R}_+, i \in \mathbb{N}$.

2 Спрямляющее пространство

Иногда может оказаться полезным знать не только вид ядра $k(x, z)$, но и вид преобразования $\phi(x)$, и наоборот. Рассмотрим данный переход на нескольких примерах.

2.1 Ядра для подмножеств

Рассмотрим ядро на пространстве всех подмножеств конечного множества D:

$$k(A_1, A_2) = 2^{|A_1 \cap A_2|}.$$

Докажем, что оно соответствует отображению в $2^{|D|}$-мерное пространство

$$(\phi(A))_U = \begin{cases} 1, U \subseteq A, \\ 0, \text{иначе}, \end{cases}$$

где U пробегает по всем подмножествам множества D.

Покажем, что при использовании указанного отображения $\phi(A)$ скалярное произведение в спрямляющем пространстве действительно имеет указанный вид:

$$\langle \phi(A_1), \phi(A_2) \rangle = \sum_{U \subseteq D} (\phi(A_1))_U (\phi(A_2))_U .$$

Заметим, что $(\phi(A_1))_U (\phi(A_2))_U = 1$ только в том случае, если $(\phi(A_1))_U = 1$ и $(\phi(A_2))_U = 1$, т.е. если $U \subseteq A_1$ и $U \subseteq A_2$. Таким образом,

$$\langle \phi(A_1), \phi(A_2) \rangle = |\{U \subseteq D | U \subseteq A_1, U \subseteq A_2\}| .$$

Подсчитаем количество таких множеств. Рассмотрим некоторое $U \subseteq A_1 \cap A_2$. Заметим, что все прочие подмножества D не будут удовлетворять хотя бы одному из условий, в то время как для таким образом выбранного U выполняются оба, поэтому необходимое число — число различных подмножеств $A_1 \cap A_2$. Оно, в свою очередь, равно $2^{|A_1 \cap A_2|}$.

2.2 Произведение мономов

Рассмотрим ядро

$$k(x, z) = \prod_{j=1}^{d} (1 + x_j z_j).$$

Какому спрямляющему пространству оно соответствует?

Раскроем скобки в выражении для $k(x, z)$. Заметим, что итоговое выражение будет включать мономы всех чётных степеней от 0 до $2d$ включительно. При этом мономы степени $2k$, $k \in \{0, \dots, d\}$, формируются следующим образом: из d скобок, входящих в произведение, случайным образом выбираются k, после чего входящие в них слагаемые вида $x_j z_j$ умножаются на единицы, входящие в состав остальных $d - k$ скобок.

Таким образом, в итоговое выражение входят все мономы степени $2k$ над всеми наборами из k различных исходных признаков, и только они. Запишем это формально:

$$k(x, z) = (1 + x_1z_1)(1 + x_2z_2)\dots(1 + x_dz_d) = \sum_{k=0}^{d} \sum_{\substack{D\subseteq\{1,\dots,d\}\\|D|=k}} \prod_{j\in D} x_jz_j.$$

Для простоты понимания приведем вид итогового выражения для $d = 2, 3$ (несложно убедиться в его справедливости путём раскрытия скобок):

$$\begin{aligned}
k((x_1, x_2), (z_1, z_2)) &= 1 + x_1z_1 + x_2z_2 + x_1x_2z_1z_2,\\
k((x_1, x_2, x_3), (z_1, z_2, z_3)) &= 1 + x_1z_1 + x_2z_2 + x_3z_3 + x_1x_2z_1z_2+\\
&\quad x_1x_3z_1z_3 + x_2x_3z_2z_3 + x_1x_2x_3z_1z_2z_3.
\end{aligned}$$

Таким образом, объект x в спрямляющем пространстве представим в следующем виде:

$$\phi(x) = (1, x_1, \dots, x_d, x_1x_2, \dots, x_1x_d, \dots, x_{d-1}x_d, \dots, x_1x_2\dots x_d) = \left(\prod_{j\in D} x_j\right)_{D\subseteq\{1,\dots,d\}},$$

то есть в виде вектора мономов всех степеней над наборами различных признаков в исходном пространстве.

2.3 Гауссовские ядра

Гауссовское ядро определяется как

$$k(x, z) = \exp\left(-\frac{\|x - z\|^2}{2\sigma^2}\right).$$

Покажем, что оно действительно является ядром.

Покажем сначала, что функция $\exp(\langle x, z\rangle)$ является ядром. Представим ее в виде предела последовательности:

$$\exp(\langle x, z\rangle) = \sum_{k=0}^{\infty} \frac{\langle x, z\rangle^k}{k!} = \lim_{n\to\infty} \sum_{k=0}^{n} \frac{\langle x, z\rangle^k}{k!}.$$

Каждый член предельной последовательности является многочленом с положительными коэффициентами, и поэтому является ядром. Предел существует во всех точках (x, z), поскольку ряд Тейлора для функции e^x сходится на всей числовой прямой. Значит, по теореме 3 данная функция является ядром.

Аналогично можно доказать, что функция $\exp(\langle x, z\rangle/\sigma^2)$ также является ядром. Гауссовское ядро легко получить из данного путём замены преобразования $\phi(x)$ на $\phi(x)/\|\phi(x)\|$.

Заметим, что можно построить гауссово ядро, используя любое другое ядро $k(x, z)$. В этом случае оно примет вид

$$\exp\left(-\frac{\|\phi(x) - \phi(z)\|^2}{2\sigma^2}\right) = \exp\left(-\frac{k(x, x) - 2k(x, z) + k(z, z)}{2\sigma^2}\right).$$

Здесь мы расписали расстояние между векторами $\|\phi(x)-\phi(z)\|^2$ в спрямляющем пространстве через функцию ядра.

Спрямляющее пространство. Какому спрямляющему пространству соответствует гауссовское ядро? Оно является пределом последовательности полиномиальных ядер при стремлении степени ядра к бесконечности, что наталкивает на мысль, что и спрямляющее пространство будет бесконечномерным. Чтобы показать это формально, нам понадобится следующее утверждение.

Утверждение 1 ([4]). Пусть $x_1, \dots, x_\ell$ — различные точки пространства $\mathbb{R}^d$. Тогда матрица

$$G = \left[\exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)\right]_{i,j=1}^{\ell}$$

является невырожденной при $\sigma > 0$.

Вспомним также факт из линейной алгебры: матрица Грама системы точек $x_1, \dots, x_\ell$ невырождена тогда и только тогда, когда эти точки линейно независимы. Поскольку матрица из утверждения 1 является матрицей Грама для точек $x_1, \dots, x_\ell$ в спрямляющем пространстве гауссова ядра, то заключаем, что в данном пространстве существует сколь угодно много линейно независимых точек. Значит, данное пространство является бесконечномерным.

Можно показать это и менее формально. Распишем функцию ядра:

$$
\begin{aligned}
k(x,z) &= \exp\left(-\frac{\|x-z\|^2}{2\sigma^2}\right) = \exp\left(-\frac{\|x\|^2}{2\sigma^2}\right)\exp\left(-\frac{\|z\|^2}{2\sigma^2}\right)\exp\left(-\frac{\langle x,z\rangle}{\sigma^2}\right) = \\
&= \{\text{раскладываем экспоненту в ряд}\} = \\
&= \exp\left(-\frac{\|x\|^2}{2\sigma^2}\right)\exp\left(-\frac{\|z\|^2}{2\sigma^2}\right)\sum_{k=0}^{\infty}\frac{\langle x,z\rangle^k}{k!\sigma^{2k}}.
\end{aligned}
$$

Легко видеть, что для получения k-го слагаемого в спрямляющем пространстве должны быть все мономы степени k над исходными признаками. Поскольку всего в сумме бесконечное число слагаемых, то и размерность спрямляющего пространства должна быть бесконечной.

Роль параметра σ. Заметим, что параметр σ в разложении гауссова ядра в ряд Тейлора входит в коэффициент перед слагаемым $\langle x,z\rangle^k$ как $1/\sigma^{2k}$. Его роль аналогична параметру R в полиномиальных ядрах. Маленькие значения σ соответствуют большим значениям R: чем меньше σ, тем больше вес при мономах большой степени, тем больше риск переобучения.

3 Выразительная способность ядерных методов

Пусть $\mathcal{X}$ — множество объектов. Зафиксируем положительно определённое ядро k и рассмотрим отображение $\Phi : \mathcal{X} \to \mathbb{R}^{\mathcal{X}}, x \mapsto k(\cdot, x)$. Построим с помощью Φ пространство со скалярным произведением. Для этого для произвольных функций $f(\cdot) = \sum\limits_{i=1}^{n}\alpha_i k(\cdot, x_i)$ и $g(\cdot) = \sum\limits_{j=1}^{n'}\beta_j k(\cdot, x'_j)$, представимых в виде линейных комбинаций элементов образа Φ, определим скалярное произведение между ними по правилу

$$\langle f, g\rangle = \sum_{i=1}^{n}\sum_{j=1}^{n'}\alpha_i\beta_j k(x_i, x'_j).$$

k называется *воспроизводящим ядром*, а гильбертово пространство $\mathcal{H}$, получаемое из построенного пространства со скалярным произведением путём его пополнения по норме — *гильбертовым пространством с воспроизводящим ядром* (reproducing kernel Hilbert space, RKHS).

Теорема 4 (Representer theorem)**.** Пусть $\{(x_i, y_i)\}_{i=1}^n \subseteq \mathcal{X} \times \mathbb{R}$ — обучающая выборка, $c : (\mathcal{X} \times \mathbb{R}^2)^n \to \mathbb{R} \cup \{\infty\}$ — произвольная функция потерь, $\Omega : [0, \infty) \to \mathbb{R}$ — возрастающая функция. Тогда любой минимизатор $f \in \mathcal{H}$ функционала эмпирического риска с регуляризацией

$$c((x_1, y_1, f(x_1)), \dots, (x_n, y_n, f(x_n))) + \Omega(||f||^2_{\mathcal{H}})$$

допускает представление в форме

$$f(x) = \sum_{i=1}^n \alpha_i k(x, x_i).$$

4 Свойства регуляризации ядерных методов

В данном разделе ограничимся рассмотрением положительно определённых ядер, инвариантных относительно сдвига. Их можно представить в виде

$$k(x, z) = h(x - z).$$

В таком случае h называется *положительно определённой функцией.*

Теорема 5 (Бохнер)**.** Непрерывная на $\mathbb{R}^d$ функция h является положительно определённой тогда и только тогда, когда на $\mathbb{R}^d$ существует конечная неотрицательная борелевская мера μ такая, что

$$h(x) = \int_{\mathbb{R}^d} e^{-i\langle x, \omega \rangle} d\mu(\omega).$$

Теорема Бохнера позволяет анализировать ядра вида $k(x, z) = h(x - z)$ в частотной области [5].

Пусть $\mathcal{X} = \mathbb{R}^d$, Υ — положительно определённый оператор из $\mathcal{H}$ в функциональное пространство со скалярным произведением, заданным стандартным образом ($\langle f, g \rangle = \int f(x)\overline{g(x)}dx$):

$$\langle f, g \rangle_k = \langle \Upsilon f, \Upsilon g \rangle = \langle \Upsilon^2 f, g \rangle$$

Если в $k(x, z) = h(x - z)$ функция $h \in L_1(\mathbb{R}^d)$ является непрерывной и строго положительно определённой, то теорема Бохнера принимает вид:

$$k(x, z) = \int e^{-i\langle x - z, \omega \rangle} v(\omega) d\omega.$$

Преобразование Фурье $k(x,\cdot)$ имеет вид

$$\mathcal{F}[k(x,\cdot)](\omega) = (2\pi)^{-d/2} \int \int (v(\omega')e^{-i\langle x,\omega'\rangle})e^{i\langle x',\omega'\rangle} d\omega' e^{-i\langle x',\omega\rangle} dx' = (2\pi)^{d/2} v(\omega) e^{-i\langle x,\omega\rangle}.$$

Следовательно, $k(x,z)$ можно представить в виде

$$k(x,z) = (2\pi)^{-d} \int \frac{\mathcal{F}[k(x,\cdot)](\omega)\overline{\mathcal{F}[k(z,\cdot)](\omega)}}{v(\omega)} d\omega.$$

Если Υ отображает f в $(2\pi)^{-d/2} v^{-1/2} \mathcal{F}[f]$, то

$$k(x,z) = \int (\Upsilon k(x,\cdot))(\omega) \overline{(\Upsilon k(z,\cdot))(\omega)} d\omega.$$

Таким образом, можно анализировать свойства регуляризации ядра k в терминах преобразования Фурье $v(\omega)$. Малые значения $v(\omega)$ усиливают соответствующие частоты в $\Upsilon f = (2\pi)^{-d/2} v^{-1/2} \mathcal{F}[f]$. Поэтому штрафование $\langle f, f\rangle_k$ равносильно сильному ослаблению соответствующих частот. Малые значения $v(\omega)$ для больших $||\omega||$ поощряются, так как высокочастотные компоненты $\mathcal{F}[f]$ соответствуют быстрым изменениям f.

Таким образом, при штрафовании $||f||^2_{\mathcal{H}} = \langle f, f\rangle_k$ происходит фильтрация высокочастотных компонент $\mathcal{F}[f]$, соответствующих быстрым изменениям f, что приводит к снижению риска переобучения на шум в данных.

Список литературы

[1] *Bishop, C.M.* Pattern Recognition and Machine Learning. // Springer, 2006.

[2] *Shawe-Taylor, J., Cristianini, N.* Kernel Methods for Pattern Analysis. // Cambridge University Press, 2004.

[3] *Sholkopf, B.A., Smola, A.J.* Learning with kernels. // MIT Press, 2002.

[4] *Micchelli, C.A.* Algebraic aspects of interpolation. // Proceedings of Symposia in Applied Mathematics, 36:81–102, 1986.

[5] *Hofmann, T., Schölkopf, B., Smola, A.J.* Kernel methods in machine learning. // The Annals of Statistics, 36:1171–1220, 2008.