Машинное обучение, ФКН ВШЭ
Семинар №12

1 Построение ядер

Напомним, что ядром мы называем функцию $K(x, z)$, представимую в виде скалярного произведения в некотором пространстве: $K(x, z) = \langle \varphi(x), \varphi(z) \rangle$, где $\varphi : \mathbb{X} \to H$ — отображение из исходного признакового пространства в некоторое *спрямляющее пространство* H.

Вспомним, какие функции в принципе могут быть ядрами — по теореме Мерсера 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 $(x_1, \ldots, x_\ell)$ матрица $K = \big(K(x_i, x_j)\big)_{i,j=1}^{\ell}$ неотрицательно определена.

Задача 1.1. *Покажите, что если $K(x, z)$ — ядро, то оно симметрично и неотрицательно определено.*

Решение. Функция $K(x, z)$ — ядро, то есть она определяет скалярное произведение в некотором пространстве: $K(x, z) = \langle \varphi(x), \varphi(z) \rangle$. Симметричность этой функции вытекает из симметричности скалярного произведения.

Покажем неотрицательную определенность. Пусть $(x_1, \ldots, x_\ell)$ — выборка, а $K = \big(K(x_i, x_j)\big)_{i,j=1}^{\ell}$ — матрица ядра, соответствующая ей. Тогда для произвольного вектора v:

$$\begin{aligned}
\langle Kv, v \rangle &= \sum_{i,j=1}^{\ell} v_i v_j K(x_i, x_j) = \\
&= \sum_{i,j=1}^{\ell} v_i v_j \langle \varphi(x_i), \varphi(x_j) \rangle = \\
&= \sum_{i,j=1}^{\ell} \langle v_i \varphi(x_i), v_j \varphi(x_j) \rangle = \\
&= \left\langle \sum_{i=1}^{\ell} v_i \varphi(x_i), \sum_{j=1}^{\ell} v_j \varphi(x_j) \right\rangle = \\
&= \left\| \sum_{i=1}^{\ell} v_i \varphi(x_i) \right\|^2 \geqslant 0.
\end{aligned}$$

Мы доказали неотрицательную определенность матрицы K, а значит и ядра $K(x,z)$.

■

Вместо того, чтобы проверять эти свойства, можно сразу составлять ядра по фиксированным правилам. Вспомним две следующие теоремы.

Теорема 1.1. *Пусть $K_1(x,z)$ и $K_2(x,z)$ — ядра, заданные на множестве $\mathbb{X}$, $f(x)$ — вещественная функция на $\mathbb{X}$, $\varphi : \mathbb{X} \to \mathbb{R}^N$ — векторная функция на $\mathbb{X}$, 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. $K(x,z) = K_3(\varphi(x), \varphi(z))$.

Теорема 1.2. *Пусть $K_1(x,z), K_2(x,z), \ldots$ — 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.

Задача 1.2. *Покажите, что произведение ядер является ядром (третий пункт теоремы 1.1).*

Решение. Пусть ядро K_1 соответствует отображению $\varphi_1 : \mathbb{X} \to \mathbb{R}^{d_1}$, а ядро K_2 — отображению $\varphi_2 : \mathbb{X} \to \mathbb{R}^{d_2}$. Определим новое отображение, которое соответствует всевозможным произведениям признаков из первого и второго спрямляющих пространств:

$$\varphi_3(x) = \left((\varphi_1(x))_i (\varphi_2(x))_j \right)_{i,j=1}^{d_1,d_2}.$$

Соответствующее этому спрямляющему пространству ядро примет вид

$$\begin{aligned} K_3(x,z) &= \langle \varphi_3(x), \varphi_3(z) \rangle = \\ &= \sum_{i=1}^{d_1} \sum_{j=1}^{d_2} (\varphi_3(x))_{ij} (\varphi_3(z))_{ij} = \\ &= \sum_{i=1}^{d_1} (\varphi_1(x))_i (\varphi_1(z))_i \sum_{j=1}^{d_2} (\varphi_2(x))_j (\varphi_2(z))_j = \\ &= K_1(x,z)K_2(x,z). \end{aligned}$$

Мы показали, что произведение двух ядер соответствует скалярному произведению в некотором спрямляющем пространстве, а значит является ядром.

■

Задача 1.3. *Пусть $p(x)$ — многочлен с положительными коэффициентами. Покажите, что $K(x,z) = p(\langle x,z\rangle)$ — ядро.*

Решение. 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 ($\varphi(x) = x$);
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 1.1, где $f(x) = \sqrt{a_0}$;
5. $\sum_{i=0}^{m} a_i\langle x,z\rangle^i$ — ядро как линейная комбинация ядер.

■

2 Спрямляющие пространства

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

Задача 2.1. *Рассмотрим ядро на пространстве всех подмножеств конечного множества D:*

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

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

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

где U пробегает по всем подмножествам множества D.

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

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

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

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

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 $2^{|A_1 \cap A_2|}$.

■

Задача 2.2. *Рассмотрим ядро*

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

Какому спрямляющему пространству оно соответствует?

Решение. 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 $2k$, $k \in \{0, \ldots, d\}$, 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:

$$K(x, z) = (1 + x_1 z_1)(1 + x_2 z_2) \ldots (1 + x_d z_d) = \sum_{k=0}^{d} \sum_{\substack{D \subseteq \{1,\ldots,d\} \\ |D|=k}} \prod_{j \in D} x_j z_j.$$

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

$$\begin{aligned} K((x_1, x_2), (z_1, z_2)) &= 1 + x_1 z_1 + x_2 z_2 + x_1 x_2 z_1 z_2, \\ K((x_1, x_2, x_3), (z_1, z_2, z_3)) &= 1 + x_1 z_1 + x_2 z_2 + x_3 z_3 + x_1 x_2 z_1 z_2 + \\ &\quad x_1 x_3 z_1 z_3 + x_2 x_3 z_2 z_3 + x_1 x_2 x_3 z_1 z_2 z_3. \end{aligned}$$

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

$$\varphi(x) = (1, x_1, \ldots, x_d, x_1 x_2, \ldots, x_1 x_d, \ldots, x_{d-1} x_d, \ldots, x_1 x_2 \ldots x_d) = \left(\prod_{j \in D} x_j \right)_{D \subseteq \{1,\ldots,d\}},$$

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

■

Задача 2.3. *Пусть* $\{(x_i, y_i)\}_{i=1}^{\ell}$, $y_i \in \{-1, +1\}$ — *произвольная выборка, в которой все объекты различны, а* $\varphi(x)$ — *отображение в спрямляющее пространство, соответствующее гауссову ядру. Покажите, что в данном спрямляющем пространстве существует линейный классификатор, безошибочно разделяющий выборку* $\varphi(x_1), \ldots, \varphi(x_\ell)$.

Решение. Покажем, что вектор весов w в спрямляющем пространстве может быть найден как линейная комбинация объектов выборки $\varphi(x_1), \ldots, \varphi(x_\ell)$, т.е. $w = \sum_{i=1}^{\ell} \alpha_i \varphi(x_i)$. Запишем условие верной классификации каждого из объектов выборки в спрямляющем пространстве:

$$\langle w, \varphi(x_i) \rangle = y_i,\ i = \overline{1, \ell}.$$

Заметим, что записанное нами условие является более строгим, чем необходимо, однако в дальнейшем мы покажем существование w, удовлетворяющего этим более строгим ограничениям. Преобразуем:

$$\left\langle \sum_{j=1}^{\ell} \alpha_j \varphi(x_j), \varphi(x_i) \right\rangle = y_i,\ i = \overline{1, \ell},$$

$$\sum_{j=1}^{\ell} \alpha_j \langle \varphi(x_j), \varphi(x_i) \rangle = y_i,\ i = \overline{1, \ell},$$

$$\sum_{j=1}^{\ell} \alpha_j K(x_i, x_j) = y_i,\ i = \overline{1, \ell}.$$

Таким образом, мы получили систему из ℓ линейных уравнений на $\alpha_1, \ldots, \alpha_\ell$, при этом матрицей системы является матрица Грама, являющаяся невырожденной (согласно утв. 1.3 лекции 13), а потому система имеет решение, и соответствующий вектор w существует.

■

3 Ядра в метрических методах

Теперь, когда у нас есть общее представление о природе ядер, попробуем использовать их для усовершенствования уже известных нам методов — например, метрических. Как вы знаете, для использования данного класса алгоритмов необходимо задать функцию расстояния на пространстве объектов — однако при использовании ядер у нас не всегда есть возможность выразить $\varphi(x)$ в явном виде. Тем не менее, оказывается, ядро содержит в себе много информации о спрямляющем пространстве, и позволяет производить в нем различные операции, не зная самого отображения $\varphi(x)$.

Задача 3.1. *Как вычислить норму вектора* $\varphi(x)$, *зная лишь ядро* $K(x, z)$?

Решение.

$$\|\varphi(x)\| = \sqrt{\|\varphi(x)\|^2} = \sqrt{\langle \varphi(x), \varphi(x) \rangle} = \sqrt{K(x, x)}.$$

■

Задача 3.2. *Как вычислить расстояние между векторами $\varphi(x)$ и $\varphi(z)$, зная лишь ядро $K(x,z)$?*

Решение.

$$\begin{aligned}\rho^2(\varphi(x),\varphi(z)) &= \|\varphi(x)-\varphi(z)\|^2 = \langle\varphi(x)-\varphi(z),\varphi(x)-\varphi(z)\rangle = \\ &= \langle\varphi(x),\varphi(x)\rangle - 2\langle\varphi(x),\varphi(z)\rangle + \langle\varphi(z),\varphi(z)\rangle = \\ &= K(x,x) - 2K(x,z) + K(z,z).\end{aligned}$$

■

Таким образом, ядра можно использовать и в метрических методах (например, kNN) — достаточно подставить в них в качестве функции расстояния величину $\sqrt{K(x,x)-2K(x,z)+K(z,z)}$.

4 Ядровой персептрон Розенблатта

Ранее мы изучали линейные модели классификации и узнали, что они отличаются друг от друга тем, какая верхняя оценка пороговой функции потерь используется в задаче оптимизации, т.е. от выбора функции $\tilde{L}(M)$:

$$Q(w,X) = \sum_{i=1}^{\ell}[y_i\langle w,x_i\rangle < 0] = \sum_{i=1}^{\ell}[M_i<0] \le \sum_{i=1}^{\ell}\tilde{L}(M_i) \to \min_w.$$

Модель классификации, использующая $\tilde{L}(M) = \max(-M,0)$, называется *персептроном Розенблатта*. Для данной модели доказана *теорема Новикова* о сходимости, которая звучит следующим образом:

Теорема 4.1. *Пусть $\mathbb{X}$ — пространство объектов, $\mathbb{Y}=\{-1,+1\}$ — пространство ответов, а выборка $X=\{(x_i,y_i)\}_{i=1}^{\ell}$ линейно разделима. Тогда персептрон Розенблатта, обученный при помощи метода стохастического градиентного спуска, позволяет за конечное число итераций найти вектор весов w^*, безошибочно разделяющий обучающую выборку, для любого начального приближения $w^{(0)}$ и для любого темпа обучения $\eta>0$, независимо от порядка предъявления объектов обучающей выборки.*

Задача 4.1. *Пусть дана обучающая выборка $X=\{(x_i,y_i)\}_{i=1}^{\ell}$, $y_i\in\{-1,+1\}$, в которой все объекты различны. Будем обучать персептрон Розенблатта при помощи стохастического градиентного спуска, при этом пусть начальное приближение вектора весов представимо в виде линейной комбинации объектов обучающей выборки: $w^{(0)}=\sum_{i=1}^{\ell}\alpha_i^{(0)}x_i$.*

1. *Покажите, что вектор весов $w^{(t)}$ после t-ой итерации может быть представлен как линейная комбинация объектов обучающей выборки.*

2. *Переформулируйте алгоритм обучения и построения прогноза для персептрона таким образом, чтобы он не использовал признаковые описания объектов в явном виде, а лишь скалярные произведения между объектами обучающей выборки.*

3. *Приведите пример спрямляющего пространства (или, что то же самое, ядра), при использовании которого ядровой алгоритм персептрона, обученный при помощи стохастического градиентного спуска, позволяет за конечное число итераций найти вектор весов, безошибочно разделяющий обучающую выборку, для любого темпа обучения $\eta > 0$, независимо от порядка предъявления объектов обучающей выборки.*

Решение.

1. Отметим, что

$$\frac{\partial \tilde{L}(M_i)}{\partial w} = \begin{cases} -y_i x_i, M_i < 0, \\ \nexists, M_i = 0, \\ 0, M_i > 0. \end{cases}$$

Для простоты во втором случае положим градиент равным $-y_i x_i$. Таким образом, имеем $\frac{\partial \tilde{L}(M_i)}{\partial w} = -y_i x_i [y_i \langle w, x_i \rangle \leq 0]$ (способ обновления весов, использующий данное предположение, называется *правилом Хэбба*, и именно для него, на самом деле, доказывается теорема Новикова).

Пусть на t-ой итерации стохастического градиентного спуска для обновления весов был выбран i_t-ый объект выборки, тогда на этой итерации выполняются следующие шаги:

(а) вычисляется значение $M_i = y_i \langle w^{(t-1)}, x_i \rangle$ для текущего классификатора;

(б) если $M_i > 0$, т.е. объект классифицируется верно, то обновление весов не производится; в противном случае вектор весов обновляется следующим образом:

$$w^{(t)} = w^{(t-1)} + \eta x_i y_i,$$

где η — темп обучения.

Таким образом, на каждой итерации к линейной комбинации объектов обучающей выборки прибавляется один из объектов с некоторым весом, а потому вектор весов на любой итерации можно представить как линейную комбинацию объектов обучающей выборки.

2. В предыдущем пункте мы показали, что алгоритм обучения персептрона состоит из двух чередующихся шагов: вычисления отступа выбранного объекта и обновления весов. Запишем отступ M_i после $(t-1)$-ой итерации, используя знания из предыдущего пункта:

$$M_i = y_i \langle w, x_i \rangle = y_i \left\langle \sum_{j=1}^{\ell} \alpha_j^{(t-1)} x_j, x_i \right\rangle = y_i \sum_{j=1}^{\ell} \alpha_j^{(t-1)} \langle x_i, x_j \rangle.$$

Таким образом, мы можем определить, допускает ли алгоритм ошибку на объекте x_i, используя лишь скалярные произведения. В случае, если ошибка имеет место, нам необходимо лишь обновить значение α_i:

$$\alpha_i^{(t)} = \alpha_i^{(t-1)} + \eta y_i.$$

Таким образом, мы переписали алгоритм обучения персептрона без использования признаковых описаний объектов в явном виде. Проделаем то же самое для прогноза обученной модели для произвольного объекта $x \in \mathbb{X}$:

$$a(x) = \operatorname{sign} \langle w^*, x \rangle = \operatorname{sign} \left\langle \sum_{i=1}^{\ell} \alpha_i^* x_i, x \right\rangle = \operatorname{sign} \left(\sum_{i=1}^{\ell} \alpha_i^* \langle x_i, x \rangle \right).$$

Заметим, что в описанных выше вычислениях вместо скалярных произведений $\langle x_i, x_j \rangle, \langle x_i, x \rangle$ в исходном признаковом пространстве можно использовать значения произвольного ядра $K(x_i, x_j), K(x_i, x)$ соответственно, тем самым будет осуществлен переход к задаче обучения персептрона в соответствующем спрямляющем пространстве.

3. Рассмотрим спрямляющее пространство, соответствующее гауссовскому ядру. В задаче 2.3 было показано, что в данном спрямляющем пространстве существует вектор весов w^*, позволяющий безошибочно разделять выборку X, т.е. она является линейно разделимой. Тогда согласно теореме Новикова за конечное число итераций стохастического градиентного спуска будет получен вектор весов, позволяющий безошибочно разделить обучающую выборку.

■