MO2

March 2021

1 Ядровые методы

Данные: $x = (x_1, ...x_m)$
Базисные функции: $\phi(x_1, ...)$
Модель принимает вид: $a(x) = \sum_{j=1}^{m} w_j \phi_j(x)$
Для хорошего качества нужно много базисных функций $\rightarrow$ Ядровые методы позволяют не перебирать большое количество базисных функций

- Быстрое обучение

Ядровые методы

1. Двойственное представление для линейной регрессии

   $Q(w) = \frac{1}{2}\sum_{i=1}^{l}(\sum_{j=1}^{m}(w_j\phi_j(x_i) - y_i)^2 + \frac{\lambda}{2}||w||_2^2 = \frac{1}{2}||\Phi w - y||_2^2 + \frac{\lambda}{2}||w||_2^2$

   $\Phi = \begin{pmatrix} \phi_1(x_1) & ... & \phi_m(x_1) \\ ... & ... & ... \\ \phi_1(x_l) & ... & \phi_m(x_l) \end{pmatrix}$
   $\nabla_w Q = \Phi^T(\Phi w - y) + \lambda w \rightarrow w = -\frac{1}{\lambda}\Phi^T(\Phi w - y) \rightarrow w = \Phi^T a$

   w является линейной комбинацией строк Φ $\rightarrow$ Решение можно искать из $w = \Phi^T a$
   $Q(a) = \frac{1}{2}||\Phi\Phi^T a - y|| + \frac{\lambda}{2}a^T\Phi\Phi^T a \rightarrow min_a$

   $\Phi\Phi^T$ - матрица Грама (попарных скалярных произведений объектов)

   Можно записать Q(w) так, что он зависит только от скалярных произведений объектов

2. SVM

$$\begin{cases} \sum_{i=1}^{l} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{l} \lambda_i \lambda_j y_i y_j < x_i, x_j > \to max_\lambda \\ 0 \geq \lambda_i \leq C \\ \sum_{i=1}^{l} \lambda_i y_i = 0 \end{cases}$$

   Такая формулировка задачи зависит от скалярных произведений объектов

3. Алгоритм

   (a) Добавляем новые признаки

   (b) Делаем это так, что $< \phi(x), \phi(z) >$ выражается через $< x, z >$

   (c) Используем метод, который использует скалярные произведения объектов

   (d) В этом методе $< x, z > \to < \phi(x), \phi(z) >$ (*Kernel trick*)

4. **Ядро** - функция $K(x, z) =< \phi(x), \phi(z) >$, где $\phi : X \to H$

   (a) H - спрямляющее пространство

   (b) ϕ - спрямляющее отображение

5. **Теорема Мерсера**

   (a) $K(x, z)$ - ядро $\leftrightarrow \begin{cases} K(x, z) = K(z, x) \\ K \text{ неотрицательно определенная} \end{cases}$

   (b) НО $\to \forall l, \forall (x_1, ..., x_l) \in R^d \to (K(x_i, x_j))_{i,j=1}^{l}$ НО

   (c) *На практике теорема Мерсера слишком сложна для применения*

6. **Теорема 1**

   (a) Если

      i. $K_1(x, z), K_2(x, z)$ - ядра, $x, z \in X$

      ii. $f^{(x)}$ - вещественная функция на X

      iii. $\phi : X \to R^n$

      iv. K_3 - ядро заданное на R^n

   (b) То *следующие функции являюся ядрами:*

      i. $K(x, z) = K_1(x, z) + K_2(x, z)$

      ii. $K(x, z) = \alpha K_1(x, z)$

      iii. $K(x, z) = K_1 K_2$

iv. $K(x,z) = f^{(x)} f^{(z)}$
v. $K(x,z) = K(\phi(x), \phi(z))$

7. **Теорема 2**
   (a) Если:
      i. $K_1(x,z), K_2(x,z), ...$ - последовательность ядер
      ii. $\exists K(x,z) = \lim_{n\to\infty} K_n(x,z), \forall x, z$
   (b) То:
      i. K - ядро

8. **Полиномиальные ядра**
   (a) p(v) - многочлен с неотриц. коэфф
   (b) $K(x,z) = w_0 + w_1 < x, z > + w_2 < x, z >^2 + ...$
   (c) Является ядром по теореме 1
   (d) $K(x,z) = (< x, z > + R)^m = \sum_{i=0}^{m} C_m^i R^{m-i} < x, z >^i$
      i. Если расписать все $< x, z >^i$, то получим все мономы степени i от исходных признаков
      ii. Зачем R? $\to$ коэффициент при мономе $= \sqrt{C_m^i R^{m-i}}$
      iii. Сравним веса при мономах 1 и (m - 1) $\sqrt{\frac{C_m^{m-1} R}{C_m^1 R^{m-1}}} = \sqrt{\frac{1}{R^{m-2}}}$
      iv. R больше - мономы высоких степеней имеют низкий вклад
      v. Конечномерное спрямляющее пространство, но можно сделать линейно разделимое пространство

9. **Гауссовы ядра**
   (a) Позволяет перевести в бесконечномерное спрямляющее пространство
   (b) $\boxed{K(x,z) = exp\left(-\frac{||x-z||^2}{2\sigma^2}\right)}$
      i. $exp(< x, z >) = \sum_{k=0}^{\infty} \frac{<x,z>^k}{k}, \forall x, z = \lim_{n\to\infty} \sum_{k=0}^{\infty} \frac{<x,z>^k}{k}$
         A. Разложение через ряд Тейлора
         B. Ядро, как последовательность ядер
      ii. $\frac{exp(<x,z>)}{2\sigma^2}$ - ядро, аналогично
      iii. $exp\left(-\frac{||x-z||^2}{2\sigma^2}\right) = exp\left(-\frac{<x-z,x-z>}{2\sigma^2}\right) = exp\left(-\frac{<x,x>-<z,z>,+<x,z>}{2\sigma^2}\right) = \frac{exp(<x,z>/\sigma^2}{exp(||x||^2/\sigma^2)exp(||z||^2/\sigma^2}\big)$

iv. $exp(< x, z > /\sigma^2) = K(x, z) =< \phi(x), \phi(z) >$

v. $\tilde{\phi(x)} = \frac{\phi(x)}{||\phi(x)||} = \frac{\phi(x)}{\sqrt{K(x,x)}}$

vi. $< \tilde{\phi(x)}, \tilde{\phi(z)} >= \frac{<\phi(x),\phi(z)>}{\sqrt{K(x,x)K(z,z)}}$

(c) Какое спрямляющее пространство? - бесконечная сумма всех мономов

(d) *Утверждение:* $x_1, ..., x_l$ - различные векторы из $\mathbb{R}^d$

Тогда:

$G = (exp\left(-\frac{||x-z||^2}{2\sigma^2}\right))_{i,j=1}^{l}$ - невырожденная при $\sigma^2 > 0$

(e) $x_1, ..., x_l \in \mathbb{R}^d$ - их матрица Грамма невырождена $\to \phi(x_1, ..., x_l)$ ЛНЗ $\to$ бесконечное количество ЛНЗ векторов $\to$ бесконечномерное пространство

10. **Ядровой SVM**

(a) $\begin{cases} \frac{1}{2}||w||^2 + C\sum_{i=1}^{l}\xi_i \to min_{w,b,\xi} \\ y_i(< w, x_i > +b) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases}$

$$L(w, b, \xi, \lambda, \mu) = \frac{1}{2}||w||^2 + C\sum_{i=1}^{d}\xi_i - \sum_{i=1}^{l}\lambda_i(y_i(< w, x_i > +b) - 1 + \xi_i) - \sum_{i=1}^{l}\mu_i\xi_i$$

В точке оптимума $\nabla_w L = 0$

$$\nabla_w L = w - \sum_{i=1}^{l}\lambda y_i x_i = 0 \to w = \sum_{i=1}^{l}\lambda_i y_i x_i$$

$$\nabla_b L = \sum_{i=1}^{l}\lambda_i y_i = 0$$

$$\nabla_{\xi_i} L = C - \lambda_i - \mu_i = 0 \to \lambda_i + \mu_i = C$$

Условие дополняющей нежесткости:

$$\lambda_i(y_i(< w, x_i > +b) - 1 + \xi_i) = 0 \to \lambda_i = 0 \text{ или } (y_i(< w, x_i > +b) - 1 + \xi_i)$$

$$\mu_i\xi_i = 0 \to \mu_i = 0 \text{ или } \xi_i = 0$$

Свойства лагранжиана:

$$\lambda \geq 0, \mu \geq 0$$

(b) Типы объектов

   i. $\lambda_i = 0 \to \mu_i = C \to \xi_i = 0 \to x_i$ лежит с правильной стороны от разделяющей гиперплоскости и на достаточном расстоянии от нее. $w = \sum_{i=1}^{l} \lambda y_i x_i \to$ объект не влияет на веса. Называется **периферийный.**

   ii. $0 < \lambda_i < 1 \to \mu \neq 0 \to \xi_0 = 0$. x_i не залезает на разделяющую полосу, но $y_i(< w, x_i > +b) = 1 \to x_i$ лежит прямо на границе. Дает вклад в w. x_i - **опорный граничный.**

   iii. $\lambda_i = C \to \xi_i > 0$. x_i дает вклад в w. $\xi_i > 0 \to x_i$ нарушает границу - **Опорные нарушители**.

(c) Подставляем $w = \sum_{i=1}^{l} \lambda y_i x_i$ в лагранжиан, учтем ограничения $\sum_{i=1}^{l} \lambda_i y_i = 0$ и $C - \lambda_i - \mu_i = 0$

   Двойственная задача SVM

$$\begin{cases} L = \sum_{i=1}^{l} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{l} \lambda_i \lambda_j y_i y_j < x_i, x_j > \to max_\lambda \\ \sum_{i=1}^{l} \lambda_i y_i = 0 \\ 0 \leq \lambda_i \leq C \end{cases}$$

(d) Если λ - решение, то $w = \sum_{i=1}^{l} \lambda_i y_i x_i$ - решение исходной задачи

(e) Задача зависит от объектов только через скалярное произведение $\to$ можно заменить его на ядро

(f) Находим b Берем $x_i : 0 < \lambda_i < C \to \xi_i = 0 \to y_i(< w, x_i > +b) = 1 \to b = y_i - < w, x_i >$

(g) Минусы ядрового SVM

   i. Сложно контролировать переобучение

   ii. Необходимо хранить в памяти матрицу Грамма

   iii. Нельзя менять функцию потерь

11. **Применение ядерной модели**

   (a) $a(x) = sign(< w, x > +b) = sign(< \sum_{i=1}^{l} \lambda y_i x_i, x > +b) = sign(\sum_{i=1}^{l} \lambda_i y_i < x_i, x > +b)$