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 $\Phi \rightarrow$ 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) $x, z \in X$
   (c) Делаем это так, что $< \phi(x), \phi(z) >$ выражается через $< x, z >$
   (d) Используем метод, который использует скалярные произведения объектов
   (e) 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 являюися 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:

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

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

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

1.1 Семинар: Задачи условной оптимизации

Учебник: Boyd, Convex Optimization

$$\begin{cases} f_0(x) \to min_{x \in R^d} \\ f_i(x) \le 0, i = 1, \dots, m \\ h_i(x) = 0, i = 1, \dots, p \end{cases}$$

$$G(x) = f_0(x) + \sum_{i=1}^{m} I_-(f_i(x)) + \sum_{i=1}^{p} I_0(h_i(x)) \to min$$

Штрафы за нарушение ограничений:

$$I_-(z) = \begin{cases} 0, z \le 0 \\ +\infty, z > 0 \end{cases}$$

$$I_0 = \begin{cases} 0, z = 0 \\ +\infty, z \neq 0 \end{cases}$$

$G(x) \to \infty$ в точках где не выполняется условие
Проблема: Недифференцируема
Заменяем функции на их аппроксимации ($\hat{I}_- = ax$)
Лагранжиан:

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x)$$

$$\lambda_i \ge 0$$

x - прямые (primal) переменные
λ, ν - двойственные переменные
Двойственная функция

$$g(\lambda, \nu) = \inf_x L(x, \lambda, \nu)$$

- Двойственная функция всегда вогнутая
- Дает нижнюю оценку на минимум функции в прямой задаче
  x' - допустимая точка (все условия выполнены)

$$L(x', \lambda, \nu) = f_0(x') + \sum_{i=1}^{m} \lambda_i f_i(x') + \sum_{i=1}^{p} \nu_i h_i(x')$$

$$f_i(x) \le 0, h_i(x) = 0 \to L(x', \lambda, \nu) \le f_0(x')$$

$$\inf_x L(x, \lambda, \nu) \le \inf_{x'} L(x', \lambda, \nu) \le \inf_{x'} f_0(x')$$

$\uparrow$ - это и есть решение исходной задачи

$$g(\lambda, \nu) \leq f_0(x_\star)$$

$$g(\lambda, \nu) \rightarrow \max_{\lambda, \nu}, \lambda_i \geq 0$$

$\lambda^\star, \nu^\star$ - решение двойственной задачи

$g(\lambda^\star, \nu^\star) \leq f_0(x_*)$ - слабая двойственность

$g(\lambda^\star, \nu^\star) = f_0(x_*)$ - сильная двойственность

<u>Достаточное условие сильной двойственности (Условие Слейтера)</u>

- Задача выпуклая:
  $f_0, f_1, \ldots, f_m$ - выпуклые
  $h_1, \ldots, h_p$ - линейные
- $\exists x'$, что все ограничения выполнены строго

Пусть имеет место сильная двойственность:

$$g(\lambda^\star, \nu^\star) = f_0(x_*)$$

$$g(\lambda^\star, \nu^\star) = \inf_x (f_0(x) + \sum \lambda^\star f_i(x) + \sum \nu^\star h_i(x)) \leq f_0(x_\star) + \sum \lambda^\star f_i(x_\star) + \sum \nu^\star h_i(x_\star) \leq f_0(x_\star)$$

Все неравенства являются равенствами:

- Если решить безусловную задачу при подставлении $\lambda^\star, \nu^\star$, то получим решение прямой задачи
- $\lambda_i^\star f_i(x^\star) = 0$ - условие дополняющей нежесткости

Теорема Куна-Такера

Необходимые условия для

$$\begin{cases} \nabla_x L(x_\star, \lambda^\star, \nu^\star) = 0 \\ f_i(x) \leq 0 \\ h_i(x) = 0 \\ \lambda_i \geq 0 \\ \lambda_i f_i(x_\star) = 0 \\ \text{Сильная двойственность} \end{cases} \qquad \leftrightarrow x_\star, \lambda^\star, \nu^\star \text{решения}$$

2 Аппроксимации ядер, ЕМ алгоритм

Скалярные произведения тяжело хранить из-за размера матрицы.

Есть ли возможность построить $\tilde{\phi}(x) \to < \tilde{\phi}(x_i), \tilde{\phi}(x_j) > \approx K(x_i, x_j)$

2.1 Метод случайных признаков Фурье

$$K(x, z) = K(x - z)$$

K - непрерывная функция

Теорема Бохнера

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

Используем:

$$K(x-z) = \int_{R^d} p(w) e^{iw^T(x-z)} dw \xrightarrow{\text{Формула Эйлера } ^1} \int_{R^d} p(w) cos(w^T(x-z)) + i \int_{R^d} p(w) cos(w^T(x-z))$$

$$\xrightarrow{K(x-z) \text{ - веществ.}} \text{Комплексная часть} = 0 \to K(x-z) = \int_{R^d} p(w) cos(w^T(x-z)) dw$$

$$\xrightarrow{\text{Монте-Карло } ^2} K(x - z) \approx \{w_j \sim p(w)\} : \frac{1}{n} \sum_{i=1}^{n} cos w_j^T (x - z)$$

$$= \frac{1}{n} \sum_{i=1}^{n} cos w_j^T x cos w_j^T z + sin w_j^T x sin w_j^T z$$

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

$$K(x - z) = < \tilde{\phi}(x), \tilde{\phi}(z) >$$

Для гауссова ядра:

$$p(w) = \mathcal{N}(0, 1)$$

[1] $e^{ix} = cos x + i sin x$

[2] $\int_a^b f(x) dx = \frac{b-a}{n} \sum_{i=1}^{N} f(u_i)$

2.2 EM алгоритм

Смесь распределений:

$$\begin{cases} p(x) = \sum_{k=0}^{K} \pi_k p_k(x) \\ \sum \pi_k = 1 \end{cases}$$

Вероятностный эксперимент:
Выбираем K из $[\pi_1, \dots, \pi_K]$, выбираем x из $pi_k(x)$
Z - скрытые переменные

$$Z = \{0, 1\}^K, \sum Z_k = 1$$

$$p(Z_k = 1) = \pi_k$$

$$p(z) = \prod_{k=1}^{K} \pi_k^{Z_k}$$

$$p(x \mid Z_k = 1) = p_k(x)$$

$$p(x \mid z) = \prod_{k=1}^{K} (p_k(x)^{Z_k})$$

$$p(x, z) = p(x \mid z)p(z) = \prod_{k=1}^{K} (\pi_k p_k(x))^{Z_k}$$

$$p(x) = \sum_{k=1}^{K} p(x, z = k) = \sum_{k=1}^{K} \pi_k p_k(x)$$

Вероятностная кластеризация:
$p_k(x)$ - распределение k-го кластера

$$x \to (p_1(x), \dots, p_k(x))$$

Хотим описать X смесью распределений

$$p(x) = \sum_{k=1}^{K} \pi_k \phi(x \mid \theta_k), \phi(x \mid \theta_k) \sim \mathcal{N}(\mu, \Sigma), \theta = (\mu, \Sigma)$$

Неполное правдоподобие:

$$ln(P(X \mid \Theta)) = \sum_{i=1}^{l} log \sum_{k=1}^{K} \pi_k \phi(x_i \mid \theta_k) \to max_\theta$$

Логарифм многооптимальная функция - просто оптимизировать ее сложно

Используем функцию полного правдоподобия

$$log(P, X \mid \Theta) = \sum_{i=1}^{l} log \sum_{k=1}^{K} (\pi_k \phi(x_i \mid \theta_k))^{Z_k}$$

$$\sum_{i=1}^{l} \sum_{k=1}^{K} Z_{ik}(log\pi_k + log\phi(x_i \mid \theta_k)) \to \max_{\Theta}$$

Известно аналитическое решение для нормального распределения.

Не знаем $Z_i k$

$$\Theta = (\pi_1, \dots, \pi_k, \theta_1, \dots, \theta_k)$$

Используем метод ALS для поиска Z, Θ

1. Оптимизация по скрытым переменным

   Апостериорное распределение: $p(Z \mid X, \Theta) = \frac{P(X,Z|\Theta)}{p(X|\Theta)}$

$$Z^{\star} = \arg\max_Z p(Z \mid X, \Theta)$$

2. Оптимизировать по Θ

$$logp(X, Z^{\star} \mid \Theta) \to \max_{\Theta}$$

3. Повторять до сходимости

Можно лучше. Не гарантирует сходимости

EM-алгоритм - метод обучения моделей со скрытыми переменными

EM-алгоритм

1. E-шаг - вычисляем $p(Z \mid X, \Theta)$ и запоминаем

2. M-шаг

$$E_{Z \sim p(Z|X,\Theta)} logp(X, Z \mid \Theta) = \sum_{Z} p(Z \mid X, \Theta) logp(X, Z \mid \Theta) \to \max_{\Theta}$$

Вывод EM-алгоритма

$$logp(X \mid \Theta) = Z(q, \Theta) + KL(q \mid\mid p)$$

$$L(q, \Theta) = \sum_Z q(Z) log \frac{p(X, Z \mid \Theta)}{q(Z)}$$

$$KL(q \mid\mid p) = -\sum_Z q(Z) log \frac{p(Z \mid X, \Theta)}{q(Z)}$$

$$\forall q(Z)$$

$L(q, \Theta)$ - нижняя оценка

Берем $q(Z) = p(Z \mid X, \Theta)$ - получаем E-шаг

$L(q, \Theta) = \sum_{Z \sim q(Z)} p(Z) log(...)$ - M-шаг

EM-алгоритм дает гарантии на рост правдоподобия