Последовательное восстановление обобщенных линейных моделей зависимостей по возрастающей обучающей совокупности

Морозов Алексей Олегович[1,2] ao.morozov@phystech.edu
Моттль Вадим Вячеславович[2] vmottl@yandex.ru
Сулимова Валентина Вячеславовна[3] vsulimova@yandex.ru

[1]Московский физико-технический институт
[2]ФИЦ ИУ РАН
[3]Тульский государственный университет

«Математические методы распознавания образов»
Москва
28 ноября 2019

Обобщенная задача восстановления зависимости в линейном пространстве наблюдений

Дано:
$\mathbf{x} \in \mathbb{R}^n$ — объекты реального мира
$y = f(\mathbf{x}) : \ \mathbb{R}^n \to \mathbb{Y}$ — неизвестная зависимость
Найти:
$\hat{y} = \hat{f}(\mathbf{x}), \ \mathbb{R}^n \to \mathbb{Y}$
$\hat{y} \approx y$
Частные случаи:

- $\mathbb{Y} = \mathbb{R}$ оценивание регрессионной зависимости: $\mathbb{R}^n \to \mathbb{R}$
- $\mathbb{Y} = \{-1, 1\}$ бинарная классификация: $\mathbb{R}^n \to \{-1, 1\}$

Обобщенная линейная модель (Generalized Linear Model)

$z(\mathbf{x}|\mathbf{a}, b) = \mathbf{a}^{\mathsf{T}}\mathbf{x} + b : \mathbb{R}^n \to \mathbb{R}$ — обобщенная линейная модель зависимости

$q(y, z) : \ \mathbb{Y} \times \mathbb{R} \to \mathbb{R}^+$ — функция связи (штрафа, потерь, loss), выпуклая по $z \in \mathbb{R}$

$\hat{y}(\mathbf{x}|\mathbf{a}, b) = \arg\min\limits_{y \in \mathbb{Y}} q(y, z(\mathbf{x}|\mathbf{a}, b))$ — решающее правило

Частные случаи:

- Регрессия: $q(y, z) = (y - z)^2$
- Классификация (бинарная): выпуклая функция, удовлетворяющая следующим свойствам $\lim\limits_{z \to -\infty} q(y = +1, z) = \infty$, $\lim\limits_{z \to \infty} q(y = +1, z) = 0$, $\lim\limits_{z \to -\infty} q(y = -1, z) = 0$, $\lim\limits_{z \to \infty} q(y = -1, z) = \infty$.
- Логистическая регрессия: $q(y, z) = \ln\left(1 + \dfrac{1}{e^{yz}}\right)$

Общепринятый принцип обучения по прецедентам: Минимум регуляризованного эмпирического риска

Множество прецедентов: $\{(\mathbf{x}_j, y_j), j = 1 \ldots, N\}$
Нужно найти параметры линейной модели: $\mathbf{a} \in \mathbb{R}^n$, $b \in \mathbb{R}$.
Критерий: минимум потерь в пределах обучающей совокупности.
Эмпирический риск на обучающей совокупности вместо среднего риска по всем объектам реального мира.

$$EmpR(\mathbf{a}, b) = \sum_{j=1}^{N} q\left(y_j, \mathbf{a}^\mathsf{T}\mathbf{x}_j + b\right) \to \min$$

Однако, если $n > N$, то существует континуум точек минимума $(\mathbf{a}, b) \in \mathbb{R}^{n+1}$

Минимум регуляризованного эмпирического риска

Ридж регуляризация, $\gamma > 0$

$$J(\mathbf{a}, b) = \gamma \mathbf{a}^{\mathsf{T}} \mathbf{a} + \sum_{j=1}^{N} q\left(y_j, \mathbf{a}^{\mathsf{T}} \mathbf{x}_j + b\right)$$

Селективная регуляризация $\mu > 0$

$$J(\mathbf{a}, b|\mu) = \gamma \sum_{i=1}^{n} \begin{pmatrix} 2\mu|a_i|, & |a_i| \leqslant \mu \\ \mu^2 + a_i^2, & |a_i| > \mu \end{pmatrix} + \sum_{j=1}^{N} q\left(y_j, \mathbf{a}^{\mathsf{T}} \mathbf{x}_j + b\right)$$

При увеличении параметра селективности μ, коэффициенты при «лишних» признаках, мало способствующих уменьшению эмпирического риска, стремятся к нулю.

В результате имеем малое подмножество активных признаков: $\mathbb{I}(\mu) = \{i : a_i \neq 0\} \subseteq \{1, \ldots, n\}$

Двойственная формулировка и численное решение задачи минимизации эмпирического риска

$$\begin{cases}\gamma \sum\limits_{i=1}^{n} \begin{pmatrix} 2\mu|a_i|, & |a_i| \leqslant \mu \\ \mu^2 + a_i^2, & |a_i| > \mu \end{pmatrix} + \sum\limits_{j=1}^{N} q\left(y_j, z_j\right) \to \min(a_1, \dots, a_n, z_1, \dots, z_n|\mu) \\ z_j = \mathbf{a}^\mathsf{T}\mathbf{x}_j + b,\ j = 1, \dots, N \end{cases}$$

Теорема

Решение разделенной задачи полностью определяется решением выпуклой двойственной задачи

$$\begin{cases}\frac{1}{2}\sum\limits_{i=1}^{n} \max\left[0, \left(\sum\limits_{j=1}^{N}\lambda_j x_{ji}\right)^2 - \mu^2\right] - \sum\limits_{j=1}^{N}\inf\limits_{z\in\mathbb{R}}\left[\frac{1}{2\gamma}q(y_j, z) + \lambda_j z\right] \to \\ \to \min(\lambda_1, \dots, \lambda_N) \\ \sum\limits_{j=1}^{N}\lambda_j = 0 \end{cases}$$

Способ последовательного приближенного решения двойственной задачи

$$\hat{a}_i = \begin{cases} 0, & \left|\sum_{j=1}^{N} \hat{\lambda}_j x_{ji}\right| \leqslant \mu \\ \left|\sum_{j=1}^{N} \hat{\lambda}_j x_{ji}\right|, & \left|\sum_{j=1}^{N} \hat{\lambda}_j x_{ji}\right| > \mu \end{cases}$$

Ограничимся случаем ридж-регуляризации с фиксированным числом признаков

Двойственная задача для N объектов:

$$W_N(\lambda_1, \dots, \lambda_N) = \frac{1}{2} \sum_{i=1}^{n} \left(\sum_{j=1}^{N} \lambda_j x_{ji} \right)^2 - \sum_{j=1}^{N} \inf_{z \in \mathbb{R}} \left[\frac{1}{2\gamma} q(y_j, z) + \lambda_j z \right]$$

Пусть она решена $(\hat{\lambda}_{N,1}, \dots, \hat{\lambda}_{N,N})$ и пришел новый объект $(\mathbf{x}_{N+1}, y_{N+1})$

Задача выпуклой оптимизации только с двумя переменными (ξ, λ)

$$W_{N+1}(\lambda_1, \ldots, \lambda_N, \lambda_{N+1}) = \frac{1}{2} \sum_{i=1}^{n} \left(\sum_{j=1}^{N} \lambda_j x_{ji} + \lambda_{N+1} x_{N+1,i} \right)^2 -$$

$$- \sum_{j=1}^{N} \inf_{z \in \mathbb{R}} \left[\frac{1}{2\gamma} q(y_j, z) + \lambda_j z \right] - \inf_{z \in \mathbb{R}} \left[\frac{1}{2\gamma} q(y_{N+1}, z) + \lambda_{N+1} z \right]$$

Идея: искать заново только одну новую переменную λ_{N+1} и один общий коэффициент ξ при прежних переменных $(\xi \hat{\lambda}_{N,1}, \ldots, \xi \hat{\lambda}_{N,N})$

$$W_{N+1}(\xi, \lambda) = \frac{1}{2} \sum_{i=1}^{n} \left(\xi \sum_{j=1}^{N} \hat{\lambda}_{N,j} x_{ji} + \lambda x_{N+1,i} \right)^2 -$$

$$- \sum_{j=1}^{N} \inf_{z \in \mathbb{R}} \left[\frac{1}{2\gamma} q(y_j, z) + \xi \hat{\lambda}_{N,j} z \right] - \inf_{z \in \mathbb{R}} \left[\frac{1}{2\gamma} q(y_{N+1}, z) + \lambda z \right]$$

Двойственная задача для регрессии

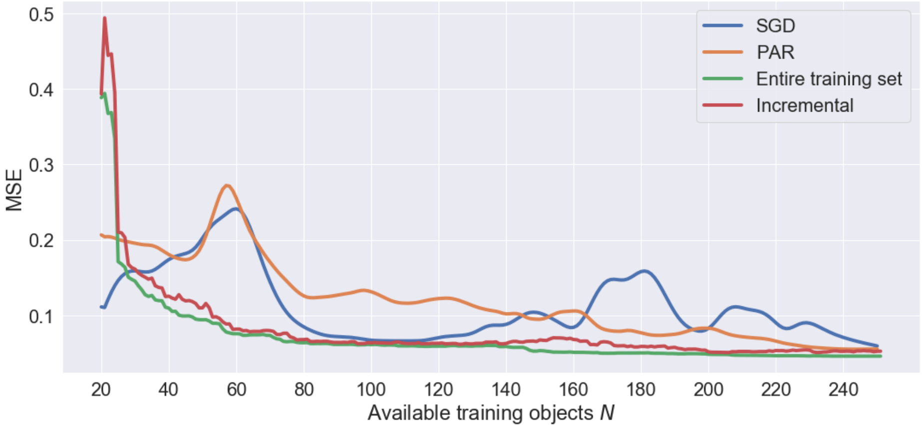
$$W_{N+1}(\xi,\lambda)=\frac{1}{2}\sum_{i=1}^{n}\left(\xi\sum_{j=1}^{N}\hat{\lambda}_{N,j}x_{ji}+\lambda x_{N+1,i}\right)^2+$$
$$+\sum_{j=1}^{N}\left(\frac{1}{2}\gamma(\xi\hat{\lambda}_{N,j})^2-y_j\xi\hat{\lambda}_{N,j}\right)+\left(\frac{1}{2}\gamma\lambda^2-y_{N+1}\lambda\right)$$

Её решение сводится к решению системы двух линейных уравнений:

$$\left(\begin{array}{c|c}\sum\limits_{i=1}^{n}\left(\sum\limits_{j=1}^{N}\hat{\lambda}_{N,j}x_{ji}\right)^2+\gamma\sum\limits_{j=1}^{N}\hat{\lambda}_{N,j}^2 & \sum\limits_{i=1}^{n}\left(\sum\limits_{j=1}^{N}\hat{\lambda}_{N,j}x_{ji}\right)x_{N+1,i} \\ \hline \sum\limits_{i=1}^{n}\left(\sum\limits_{j=1}^{N}\hat{\lambda}_{N,j}x_{ji}\right)x_{N+1,i} & \sum\limits_{i=1}^{n}x_{N+1,i}^2+\gamma\end{array}\right)\begin{pmatrix}\xi\\ \lambda\end{pmatrix}=\left(\begin{array}{c}\sum\limits_{j=1}^{N}y_j\hat{\lambda}_{N,j}\\ \hline y_{N+1}\lambda\end{array}\right)$$

Эксперимент

Данные: временные ряды ежемесячных доходностей активов на бирже за 20 лет ($N = 250$). Мы вычислили временной ряд доходности $\mathbf{y} = \mathbf{X}^T\mathbf{a} + \psi \in \mathbb{R}^N$ с 5%-ой дисперсией шума ψ. Взяли первые 20 наблюдений и инкрементально обучались



Использованная литература

- Nelder J. A., Wedderburn R. W. M. Generalized linear models //Journal of the Royal Statistical Society: Series A (General). – 1972. – Т. 135. – №. 3. – С. 370-384. www.rss.onlinelibrary.wiley.com/doi/abs/10.2307/2344614
- Crammer K. et al. Online passive-aggressive algorithms //Journal of Machine Learning Research. – 2006. – Т. 7. – №. Mar. – С. 551-585. www.jmlr.org/papers/volume7/crammer06a/crammer06a.pdf
- Tatarchuk A. et al. A support kernel machine for supervised selective combining of diverse pattern-recognition modalities //International Workshop on Multiple Classifier Systems. – Springer, Berlin, Heidelberg, 2010. – С. 165-174. www.link.springer.com/chapter/10.1007/978-3-642-12127-2_17

Спасибо за внимание!