Ансамблирование линейных моделей с помощью выпуклых комбинаций через максимизацию корреляции с откликом

Борисов Иван Максимович

МГУ

Научный руководитель: Сенько Олег Валентинович

2024

Цель исследования

Разработка нового метода линейной регрессии, основанного на ансамблировании выпуклых комбинаций "элементарных"регрессий, с акцентом на максимизацию корреляции предсказаний с целевой переменной. Ожидается, что предложенная модель будет демонстрировать качество, сопоставимое с Эластичной сетью на данных малого объема.

Литература

- А. А. Докукин, О. В. Сенько, “Регрессионная модель, основанная на выпуклых комбинациях, максимально коррелирующих с откликом”, Ж. вычисл. матем. и матем. физ., 55:3 (2015), 530–544; Comput. Math. Math. Phys., 55:3 (2015), 526–539
- Senko O., Dokukin A. Optimal forecasting based on convex correcting procedures // New Trends in Classifica* tion and Data Mining. ITHEA, Sofia, 2010. P. 62–72.
- Ensembles of Regularized Linear Models//Anthony Christidis, Laks V.S. Lakshmanan, Ezequiel Smucler, Ruben Zamar (2001)

Постановка задачи

Решается задача

$X = \{x^1, x^2, \ldots, x^n\}, x^i \in \mathbb{R}^d$ - объекты,
$Y = \{y^1, \ldots, y^n\}, y^i \in \mathbb{R}$ - отклики.
Решаем задачу линейной регрессии $a : X \to Y$, то есть $a(x) = \langle w, x \rangle + b$, где $w \in \mathbb{R}^d, b \in \mathbb{R}$ — обучаемые параметры линейной модели.

Проблема

Мультиколлинеарность — высокая корреляция между переменными. Особенно данная проблема существенна в случае $d >> n$.

Борьба с мультиколлинеарностью.

Начальная задача

$$L(\theta) \to min_\theta$$

где $\theta = (w, b) \in \mathbb{R}^{d+1}$ — вектор обучаемых параметров.

Новая задача

$$L(\theta) + C(\theta) \to min_\theta$$

где $C : \Theta \to \mathbb{R}$.

Борьба с мультиколлинеарностью.

Положим

Пусть $L(\theta) = MSE(\theta)$ и $C_i = w_i \rho(y, x_i)$,
где $\rho(y, x_i)$ — коэффициент корреляции Пирсона.

Тогда:

$$\begin{cases} \sum_{i=1}^{n} (y^i - b - \langle w, x^i \rangle)^2 \to min_\theta \\ C_1 = w_1 \rho(y, x_1) \geq 0 \\ \dots \\ C_k = w_k \rho(y, x_k) \geq 0 \end{cases} \tag{1}$$

Переход от задачи оптимизации к поиску наилучшей выпуклой комбинации.

Решение (1) эквивалентно следующему алгоритму:

1. Методом наименьших квадратов строятся d "элементарных"регрессоров:
$$R_i = b_i + w_i x_i, \quad \overline{R} = (R_1, \ldots, R_d).$$
2. Находится выпуклая комбинация с максимальной корреляцией с откликом:
$$\sum_{i=1}^{d} c_i = 1, c_i \geq 0 \Rightarrow \rho(P(\overline{c}^*, \overline{R}), y) \geq \rho(P(\overline{c}, \overline{R}), y)$$
$$\forall \overline{c} = (c_1, \ldots, c_d)$$
где $P(\overline{c}, \overline{R}) = \sum_{i=1}^{d} c_i^* R^i$ и $\overline{c}^*$ — оптимальная комбинация.
3. Строится линейная регрессия для прогнозирования y:
$$a(x) = \beta + \alpha P(\overline{c}^*, \overline{R}).$$

Оптимизационная задача

Корреляция Пирсона для выпуклой комбинации

$$\rho(Y, P(\overline{c}, \overline{R})) = \frac{cov(\overline{P}, Y)}{\sqrt{\mathbb{D}\overline{P}}\sqrt{\mathbb{D}Y}} = \frac{\mathbb{E}[(\overline{P} - \mathbb{E}\overline{P})(Y - \mathbb{E}Y)]}{\sqrt{\mathbb{D}\overline{P}}\sqrt{\mathbb{D}Y}}$$

$$= \frac{\sum_{i=1}^{d} c_i \mathbb{D}R_i}{\sqrt{\mathbb{D}Y}\sqrt{\sum_{i=1}^{l} c_i \mathbb{D}R_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} c_i c_j \varrho(R_i, R_j)}} =$$

$$= \frac{\theta}{\sqrt{\mathbb{D}Y}\sqrt{\theta - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} c_i c_j \varrho(R_i, R_j)}} \to \max_{\theta}$$

Решение для двух элементарных предикторов

$$\rho(Y, P(r,\theta)) = \frac{\theta}{\sqrt{\mathbb{D}Y}\sqrt{\theta + \varrho(r_1, r_2)\frac{(\theta-\mathbb{D}r_1)(\theta-\mathbb{D}r_2)}{(\mathbb{D}r_1-\mathbb{D}r_2)^2}}} \to max_\theta$$

Взяв производную по θ и приравняв ее к 0, получим:

$$\theta^* = \frac{-2\varrho(r_1, r_2)\mathbb{D}r_1\mathbb{D}r_2}{(\mathbb{D}r_1 - \mathbb{D}r_2)^2 - \varrho(r_1, r_2)(\mathbb{D}r_1 + \mathbb{D}r_2)}$$

Утверждение 1:
Ансамбль $\bar{r} = (r_1, r_2)$ является несократимым $\Longleftrightarrow$

$$\begin{cases} \theta^* = \frac{-2\varrho(r_1,r_2)\mathbb{D}r_1\mathbb{D}r_2}{(\mathbb{D}r_1-\mathbb{D}r_2)^2-\varrho(r_1,r_2)(\mathbb{D}r_1+\mathbb{D}r_2)} \\ \theta^* \in (\mathbb{D}r_1, \mathbb{D}r_2) \\ \exists i \in \{1,2\} : \rho(Y, P(r,\theta^*)) \geq \rho(Y, r_i) \end{cases} \tag{2}$$

Многомерный случай

$$P = \left|\left|\rho(r_i, r_j)\right|\right|_{d\times d}, V = \left|\left|\mathbb{D}r_i\right|\right|_{1\times d}, I = \left|\left|1\right|\right|_{1\times d}, O = \left|\left|0\right|\right|_{1\times d}$$

$$A_k = \sum_{i=1}^{l} P_{ki}^{-1}\mathbb{D}r_i, B_k = \sum_{i=1}^{l} P_{ki}^{-1}$$

$$C_k = \frac{\alpha B_k - \beta A_k}{\alpha\gamma - \beta^2}, D_k = \frac{\gamma A_k - \beta B_k}{\alpha\gamma - \beta^2}$$

$$Q_0 = \sum_{i=1}^{d}\sum_{j=1}^{d} C_i C_j \varrho_{ij}, Q_1 = \sum_{i=1}^{d}\sum_{j=1}^{d}(C_i D_j + C_j D_i)\varrho_{ij}$$

$$Q_2 = \sum_{i=1}^{d}\sum_{j=1}^{d} D_i D_j \varrho_{ij}$$

Тогда:

$$\boxed{c_k^* = C_k + D_k\theta} \qquad (3)$$

Многомерный случай

Утверждение 2: Если ансамбль $\overline{r}$ является несократимым относительно коэффициента корреляции, и $\exists \mathrm{P}^{-1}$, $(\theta_{min}, \theta_{max})$ — интервал значений, на котором $\forall k = 1, \ldots, d \Rightarrow c_k^* > 0$, тогда выполнены неравенства:

$$\begin{cases} \theta_{min} < \theta^* < \theta_{max} \\ \kappa(\theta^*) > \kappa(\theta_{min}) \\ \kappa(\theta^*) > \kappa(\theta_{max}), \end{cases}$$

$$\text{где } c_k^* = C_k + D_k\theta, \theta^* = \frac{Q_0}{(1 - 0.5Q_1)}$$

Также максимум корреляции $\rho(Y, P(\overline{r}, c)) = \frac{\kappa(\theta)}{\mathbb{D}Y}$ на $\overline{D}_d$ достигается при θ^* в точке c^*.
Верно и обратное утверждение.

Основной алгоритм.

Двумерный случай

- Обучаем на i-ом признаке d МНК-регрессий.
- На валидационной выборке оцениваем:

$$\mathbb{D}R_i = \frac{1}{n}\sum_{k=1}^{n}(R_i(x_i^k) - \mathbb{E}R_i)^2, \mathbb{E}R_i = \frac{1}{n}\sum_{k=1}^{n} R_i(x_i^k)$$

$$\varrho(R_i, R_j) = \frac{1}{n}\sum_{k=1}^{n}(R_i(x_i^k) - R_j(x_j^k))$$

- Вычисляем θ^* для всех пар «элементарных» предикторов по формуле (2).
- Проверяем $\theta^*_{i,j} \in (\mathbb{D}R_i, \mathbb{D}R_j)$.
- Оставляем только те θ^*_{ij}, для которых $\forall k \in \{i,j\} : \rho(Y, P(\overline{R}, \theta^*_{ij})) \geq \tau\rho(Y, R_k), \tau \geq 1$

Основной алгоритм.

Многомерный случай

- Создаем словарь, где ключами являются индексы "элементарных"регрессоров в ансамбле, а значениями — их веса c_k.
- Для каждого ненулевого θ^*_{ij} из двумерного случая находим коэффициенты по формуле (3) и записываем их в словарь.
- Проходим по переменным, не входящим в текущий ансамбль, и добавляем соответствующий элементарный предиктор.

Основной алгоритм.

Многомерный случай (продолжение)

- Проверяем выполнение условий утверждения 2 для нового ансамбля.
- Если условия выполнены, обновляем словарь, удаляя старый ансамбль и добавляя новый, и продолжаем перебор переменных.
- Если условия нарушены, завершаем перебор для текущего ансамбля и возвращаемся к предыдущему.

Альтернативный алгоритм

- Бутстрапируем выборку.
- Запускаем алгорит построения Оптимальных Выпуклых Комбинаций.
- Вместо полного перебора на каждой итерации фиксируем максимально коррелирующую с целевой переменной комбинацию.
- Для добавления вариативности в комбинации используем метод случайных подпространств.
- Повторяем заданное число $n\_bootstrap$ раз.

Получение итогового предсказания

Обозначим

l - число выпуклых комбинаций; $\text{ВПК}_i(x)$ - предсказание i-ой комбинации на x; Y - целевые переменные тренировочной выборки; $\overline{\text{ВПК}}(X)$ - матрица, столбцы которой - предсказания каждой выпуклой комбинации на тренировочной выборке; ρ_i - коэффициент корреляции Пирсона i-ой комбинации с целевой переменной.

Тогда

- $\text{ВПК}_{\text{ср}}(x) = \alpha_1(\frac{1}{l}\sum_{i=1}^{l}\text{ВПК}_i(x)) + \beta_1$
- $\text{ВПК}_{\text{кор}}(x) = \alpha_2(\sum_{i=1}^{l}\frac{1}{1-\rho_i^2}\text{ВПК}_i(x)) + \beta_2$
- $\text{ВПК}_{\text{лин}}(x) = \text{Ridge}[\overline{\text{ВПК}}(X), Y](x)$

где α_i, β_i — коэффициенты, подобранные по MSE на тренировочной выборке.

Эксперименты

Данные

2 датасета размеров 176×94 и 92×100. Разиты в отношении 8:2 на обучение и тест с $random\_seed = 42$.
Домен: химические элементы. К примеру, CaAuBi, CdAgSb, CdAuSb, CdCuSb, CePdBi, ..., ZrNiSn, ZrPdSn, ZrPtSn, ZrRhSb, ZrRuSb.
Предлагается по набору признаков химических элементов предсказать некоторый параметр данного химического элемента

Эксперименты

Модели для сравнения

Если $L = \sum_{i=1}^{n}(y_i - a(x_i))^2 + R(w)$, то в зависимости от функции $R(w)$ определим:

- Ridge: $R(w) = ||w||_2^2$
- Lasso: $R(w) = ||w||_1$
- ElasticNet: $R(w) = 0.5 \cdot ||w||_1 + 0.25 \cdot ||w||_2^2$

Также для сравнения обучим ARD-регрессию (RVR) и Байесовскую Ridge регрессию.

Эксперименты

Гиперпараметры

1. $\tau = 1.25$ - сила, с которой растет корреляция при добавлении нового предиктора.
2. $p = 0.5$ - вероятность вхождения i-ого признака в методе случайных подпространств.
3. $n_{\text{bootstrap}} = 10$ - число бутстрапирований выборки (максимальное число предикторов в ансамбле.)

Результаты

Модель	r^2	Корреляция Пирсона
$\text{ВПК}_{\text{ср}}$	0.566/0.89	0.794/0.946
$\text{ВПК}_{\text{кор}}$	0.598/0.894	0.81/0.949
$\text{ВПК}_{\text{лин}}$	0.953/0.918	0.977/0.962
Ridge	**0.9603**	**0.9809**
Lasso	0.843	0.922
ElasticNet	0.885	0.943
ARD	0.911	0.958
Байесовская	0.944	0.973

Таблица: Данные 1

Результаты

Модель	r^2	Корреляция Пирсона
$\text{ВПК}_{\text{ср}}$	0.9/0.924	0.949/0.962
$\text{ВПК}_{\text{кор}}$	0.882/0.921	0.939/0.961
$\text{ВПК}_{\text{лин}}$	0.961/0.935	0.981/0.97
Ridge	0.961	0.981
Lasso	0.949	0.975
ElasticNet	0.953	0.9767
ARD	**0.963**	**0.982**
Байесовская	0.962	0.982

Таблица: Данные 2

Результаты

По результатам третий метод усреднения классического алгоритма демонстрирует наилучшее качество, в то время как первый и второй методы значительно уступают. Сравнение ВПК с другими моделями будет основываться на лучших результатах. Новый алгоритм превосходит Лассо и Эластичную сеть, а также показывает сопоставимые результаты с Ridge, ARD и Баейсовской регрессиями. В целом, новый метод имеет право на существование и может показывать результаты не хуже устоявшихся решений.

Итоги

В результате работы были обоснованы теоретические основы модели, основанной на ансамблировании линейных моделей с выпуклыми комбинациями для максимизации корреляции с целевой переменной. Алгоритм был реализован и протестирован на реальных данных, показав лучшие результаты по сравнению с некоторыми существующими решениями. Исследованы более эффективные методы агрегации ансамбля, включая метод случайных подпространств и жадного отбора, который снижает вычислительные сложности перебора d! комбинаций. В дальнейшем алгоритм можно развить с помощью идеи дивергентного леса.