Разработка эффективных методов построения ансамблевых регрессионных линейных моделей, основанных на минимизации ошибки

Александр Сергеевич Суржанский

Московский Государственный Университет

22 ноября 2024 г.

Постановка задачи

Предположим, что у нас имеется набор из l предикторов, прогнозирующих значения некоторой переменной Y. Прогноз, вычисляемый i-м предиктором для некоторого объекта ω, далее будет обозначаться через $z_i(\omega)$. Пусть $c = (c_1, \ldots, c_l)$ – вектор действительных неотрицательных коэффициентов, удовлетворяющий условию $\sum_{i=1}^{l} c_i = 1$. Будем рассматривать выпуклые корректирующие процедуры (ВКП), вычисляющие коллективный прогноз $Z(\omega, c)$ в виде

$$Z(\omega, c) = \sum_{i=1}^{l} c_i z_i(\omega)$$

Квадратичная ошибка

$\sum_{i=1}^{l} c_i[Y(\omega) - z_i(\omega)]^2 =$
$[Y(\omega) - Z(\omega, c)]^2 + \sum_{i=1}^{l} c_i[z_i(\omega) - Z(\omega, c)]^2$
Таким образом, квадратичная ошибка ВКП при прогнозировании Y для объекта ω может быть выражена в виде разности:
$[Y(w) - Z(\omega, c)]^2 =$
$\sum_{i=1}^{l} c_i[Y(w) - z_i(\omega)]^2 - \sum_{i=1}^{l} c_i[z_i(\omega) - Z(\omega, c)]^2$
Задача поиска оптимальной ВКП может быть представлена как задача минимизации математического ожидания ошибки на пространстве Ω всевозможных прогнозируемых объектов:

$$\Delta_{\text{ccp}} = E_{\Omega}\left(\sum_{i=1}^{l} c_i[Y(\omega) - z_i(\omega)]^2 - \sum_{i=1}^{l} c_i[z_i(\omega) - Z(\omega, \mathbf{c})]^2\right)$$

Задача квадратичного программирования

Обозначим $\delta_i = E_\Omega[Y(\omega - z_i(\omega))]^2, \rho_{ij} = E_\Omega = [z_i(\omega) - z_j(\omega)]^2$
Функционал обобщенной ошибки может быть представлен в виде

$$\Delta_{\text{ccp}} = \sum_{i=1}^{l} c_i \delta_i + E_\Omega Z^2(\omega, \mathbf{c}) - \sum_{i=1}^{l} c_i E_\Omega z_i^2(\omega) =$$

$$= \sum_{i=1}^{l} c_i \delta_i + \sum_{i'=1}^{l} \sum_{i''=1}^{l} c_{i'} c_{i''} E_\Omega [z_{i'}(\omega) z_{i''}(\omega)] - \sum_{i=1}^{l} c_i E_\Omega z_i^2(\omega).$$

Задачу минимизации обобщенной ошибки можно свести к задаче квадратичного программирования

$$\sum_{i=1}^{l} c_i \delta_i - \frac{1}{2} \sum_{i'=1}^{l} \sum_{i''=1}^{l} c_{i'} c_{i''} \rho_{i' i''} \longrightarrow \min$$

с ограничениями $\sum_{i=1}^{l} c_i = 1, c_i \geq 0, i = 1, 2, \ldots, l$

Нескоратимые предикторы

Под условием несократимости набора предикторов понимается условие невозможности удаления из набора какого-либо элемента без уменьшения точности соответствующей оптимальной ВКП.

Пусть:

$$\overline{D}_l = \left\{ \mathbf{c} \middle| \sum_{i=1}^{l} c_i = 1, c_i \geq 0, i = \overline{1, l} \right\},$$

$$D_l = \left\{ \mathbf{c} \middle| \sum_{i=1}^{l} c_i = 1, c_i > 0, i = \overline{1, l} \right\}.$$

Набор предикторов будет называться несократимым, если существует точка $c \in D_l$ такая, что

$$\Delta_{\text{ccp}}(\mathbf{c}) < \Delta_{\text{ccp}}(\mathbf{c}') \quad \forall \mathbf{c}' \in \overline{D}_l \backslash D_l.$$

Несократимость в случае двух предикторов

Для того чтобы обобщенная квадратичная ошибка оптимальной ВКП по двум предикторам была меньше, чем обобщенная квадратичная ошибка каждого из предикторов, необходимо и достаточно, чтобы выполнялось неравенство

$$|\delta_2 - \delta_1| < \rho_{12}$$

В этом случае обобщенная квадратичная ошибка достигает своего минимума, если

$$c_1 = \frac{(\rho_{12} + \delta_2 - \delta_1)}{2\rho_{12}},$$

$$c_2 = \frac{(\rho_{12} + \delta_1 - \delta_2)}{2\rho_{12}}.$$

Несократимость в общем случае

Поэтому одновременное выполнение следующих условий является необходимым и достаточным условием несократимости набора предикторов:

$$\sum_{i'=1}^{l}\left[\delta_{i'}\bar{\rho}_{i'i}+\frac{\frac{1}{2}-\sum_{j'=1}^{l}\sum_{j=1}^{l}\delta_{j}\bar{\rho}_{j'i}}{\sum_{j'=1}^{l}\sum_{j=1}^{l}\bar{\rho}_{j'j}}\bar{\rho}_{i'i}\right]>0.$$

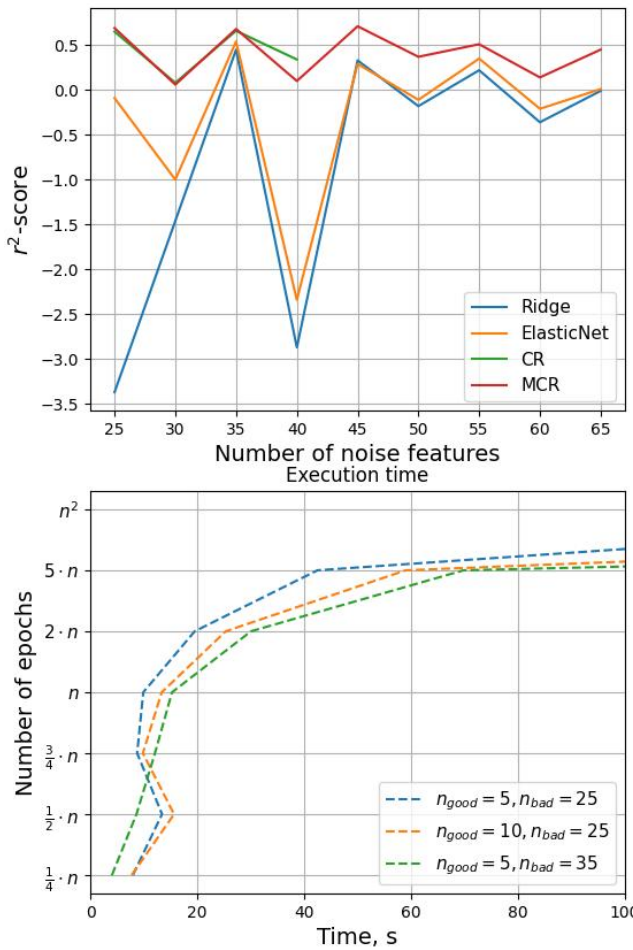
$$-\frac{1}{2}\sum_{i'=1}^{l}\sum_{i''=1}^{l}\rho_{i'i''}z_{i'}z_{i''}>0.$$

Важным свойством ВКП является снижение дисперсии коллективных прогнозов $Z(\omega)$ по отношению к средней дисперсии по ансамблю, так как выполнено следующее утверждение:

$$\sqrt{\mathrm{Var}(Z)}\leq\sum_{i=1}^{l}c_i\sqrt{\mathrm{Var}(z_i)}.$$

Эксперименты

r^2-score and execution time from the number of epochs



Модель	Датасет 1	Датасет 2
ВКП	0.9367	0.9521
CMC	0.9527	0.9613
Ridge	0.9603	0.9611
Lasso	0.8427	0.9492
ElasticNet	0.8849	0.9527

Сравнение моделей по $r^2 - score$

Заключение

Поиск коэффициентов оптимального ВКП сводится к задаче квадратичного программирования, которая решается в терминах избыточности набора предикторов. Установлены необходимые и достаточные условия избыточности для наборов из двух предикторов, а также из произвольного числа предикторов. Была рассмотрена линейная регрессия, основанная на оптимизации ВКП, которая естественно включает в себя отбор значимых переменных. Планируется также проведения исследования по использованию других базовых предикторов, одним из которых может быть дивергентный лес.