1 Метод Группового Учета Аргументов (МГУА)

Целью данного алгоритма является построение линейной регрессионной модели с помощью ограничнного числа столбцов МД-матрицы.
Нам дана МД-матрица X, размера $N \times M$ и целевой вектор-столбец Y, длины N. Сначала приведем эти данные к "хорошему"виду, чтобы каждый столбец имел нулевое среднее и единичную норму.
Введем обозначения:

- $\hat{Y}$ - вектор, приближающий Y, вектор прогноза. Он может быть получен через решающую функцию F.
- $Cor(a,b) = \frac{\sum_{i=1}^{N} a_i b_i}{\sqrt{(\sum a_i^2)(\sum b_i^2)}}$ - функция корреляции двух приближений a и b.
- $R^2(Y,\hat{Y}) = 1 - \frac{\sum_{i=1}^{N} \varepsilon_i^2}{\sum (Y_i - Y_{cp})^2}$ - функция качества оценки $\hat{Y}$.
  - $\varepsilon_i = Y_i - \hat{Y}_i$
  - Y_{cp} - среднее значение Y.
- Q - размер буфера. Это параметр, фиксированное число.
- C_T - порог корреляции. Это параметр, фиксированное число.

Создаем буфер размера Q, в который будем добавлять лучшие приближения $\hat{Y}_i$, относительно $R^2(Y,\hat{Y})$, корреляция между которыми меньше C_T.
Пусть $I > 0$ - количество итераций в нашем алгоритме. На первом шаге заполняем буфер, строя приближения по двум столбцам каким-либо алгоритмом, напрмер линейной регрессией. На втором шаге заполняем буфер приближениями, построенными тем же алгоритмом, но между столбцом из X и вектором из буфера из предыдущего шага. Таким образом максимальное количество столбцов, использованное на итерации i равно i+1.
Далее повторяем последний шаг, пока не будет выполнено I итераций.
В качестве искомой зависимости берем ту, которой соответствует вектор оценки из буфера с наилучшим значением функции качества.
Плюсы:

- выбор не более заданного количества дескрипторов и построение на их основе результирующей модели;
- возможность использования в качестве базового алгоритма регрессии любого алгоритма для построения приближающей функции;
- таким образом с помощью МГУА можно выбирать наиболее значимые признаки.

Минусы:

- количество итераций алгоритма должен задавать эксперт.
  Вариант решения: использование кросс-валидации (метод скользящего контроля), то есть нахождение количества итераций, при котором среднее значение функции качества наилучшее. Необходимо только задать ограничение сверху на количество итераций, либо оно ограничится количеством дескрипторов, что может привести к переобучению.
- возможная неоптимальность решающей функции.