Кластерная регрессия

Дмитрий Федоряка

1 Введение

Здесь описывается алгоритм, предложенный автором в ходе написсания статьи о композициях моделей в задаче векторной авторегресии, но не показавший на практике улушения точности в сравнении с базовым алгоритмом.

Предложеный алгоритм является алгоритмом композиции моделей. Композиция моделей — это алгоритм f, построенный на основе нескольких базовых алгоритмов $f_1, \ldots, f_K$:

$$f(\mathbf{x}) = \sum_{i=1}^{K} w_i(\mathbf{x}) f_i(\mathbf{x})$$

Здесь $f_1, \ldots, f_K$ - базовые модели (в данной работе - линейные модели с различными матрицами весов). $w_1, \ldots, w_K$ - веса, с которыми модели учитываются. Для них должно выполняться $\sum_{i=1}^{K} w_i = 1$. Эти веса могут зависеть от объекта $\mathbf{x}$. В таком случае им можно придать вероятностный смысл: $w_i(\mathbf{x})$ - вероятность того, что объект $\mathbf{x}$ описыается моделью f_i.

При построении алгоритма использовалась идея алгоритма кластеризации K-means: в пространстве объектов выделяются некоторые группы объектов и центры, затем за несколько итерация объекты распределяются между центрами так, чтобы каждый объект был ближе к «своему» центру, чем остальным. Для этого сначала строится произвольное разбиение объектов, а затем на каждой итерации центры вычисляяются как центры масс в каждой группе и перераспределяются объекты.

Здесь вместо центров рассматриваются модели. Вместо вычисления центра масс производится обучение модели. «Ближайшую» модель для объектов находим как ту, которая даёт самую меньшую ошибку на этом объекте.

2 Описание алгоритма

2.1 Обучение

Опишем итерационный алгоритм, который строит модели f_i и находит веса w_i.

Произвольно разобьём все объекты на K непересекающихся подмножеств (классов): $\overline{1,m} = \bigsqcup_{j=1}^{K} C_j$. Предположим, что объекты i-го класса описыаются i-й моделью. Теперь запустим итерационный процесс, который должен осуществить это предположение.

На первом шагу итерации будем обучать модели на объектах соответсвующих им классов.

На втором шагу будем вычислять для i-го объекта и j-й модели, вес $w(i,j)$, показывающий насколько хорошо данная модель описывает объект (в качестве критерия качества используя погрешность предсказания по отношению к известному истинному значению), и на основании полученных значений перераспределять объекты между классами.

После перераспределения объектов может возникнуть одна из двух исключительных ситуаций. В некоторый класс может не попасть ни один объект. Тогда, скорее всего, начальное количество классов K слишком большое, и надо удалить из рассмотрения эти

пустые классы, уменьшив K. Может оказаться, что перераспределения не произошло. Тогда нужно остановить алгоритм, т.к. на следующих итерациях перерсапределения, скорее всего, происходить тоже не будет.

После некоторого числа итераций получим K разных моделей.

2.2 Предсказание

Пусть теперь нам надо найти ответ для объекта x. Для этого надо определить веса $w_i(x)$. Построим предварительный ответ для объекта, использовав единственную линейную модель f_0, обученную на всей выборке. Затем предскажем ответ с помощью каждой модели. Сравнивая предварительный ответ и ответы моделей, найдём веса. Для этого введём монотонно убывающую функцию $\mathfrak{E}(S_i(x))$, которая будет для ошибки $S_i(x) = \frac{||f_i(x)-f_0(x)||}{||f_0(x)||}$ вычислять меру правдоподобия. Вычислим правдоподобие для каждого класса, отнормируем - и получим веса моделей:

$$w_i(x) = \frac{\mathfrak{E}(S_i(x))}{\sum_{i=1}^{K} \mathfrak{E}(S_i(x))}$$

Примеры возможных функций $\mathfrak{E}(S)$: $\frac{1}{S+\varepsilon}$, e^{-S}, $1 - \frac{1}{1+e^{-S}}$.

3 Формальная запись

Algorithm 1 Кластерная регрессия

Вход:

$(X \in \mathbb{R}^{m \times n}, Y \in \mathbb{R}^{m \times r}, X^0 \in \mathbb{R}^n)$ — входные данные задачи авторегрессии;

K — число моделей;

N_{it} — число итераций;

$\mathfrak{E}$ - функция преобразования погрешности в правдоподобие.

Выход: Y^0 — ответ задачи авторегрессии

$\forall j \in \overline{1,m}\ C_j = rand(1,m)$ — начальное разбиение на классы
для $iter \in \overline{1,N_{it}}$
  для $j \in \overline{1,K}$ // *M-шаг*
    Обучаем алгоритм f_j на объектах с индексами C_j.
  для $i \in \overline{1,m}$ // *E-шаг*
    для $j \in \overline{1,K}$
      $S(i,j) = ||f_j(X_i) - Y_i||$ — вычисление ошибок
  для $i \in \overline{1,m}$ // *Перераспределение*
    для $j \in \overline{1,K}$
      $C_j = \{i \in \overline{1,m} | \underset{k}{argmin}(S(i,k)) = j\}$
  если $\exists i : C_i = \varnothing$ **то**
    $\{C_1, \ldots, C_K\} = \{C_i | C_i \neq \varnothing\}$
  если перераспределение не поменяло класс ни для одного объекта **то**
    выйти из цикла
Обучаем алгоритм f_0 на всей выборке.
$\tilde{Y^0} = f_0(X^0)$ — предварительный ответ.
$\forall\ j \in \overline{1,K} w'_j = \mathfrak{E}(\frac{||f_j(X^0) - Y^0||}{Y^0||})$ — вычисление вероятностей.
$\forall\ j \in \overline{1,K} w_j = \frac{w'_j}{\sum k=1^K w'_k}$ — нормировка.
$Y^0 = \sum_{j=1}^{K} w_j \cdot f_j(X^0)$ – ответ.