Модель Random Forest

Случайный лес – пример ансамблевого алгоритма, который основан на предсказаниях решающих деревьев. Идея ансамблевых моделей заключается в построении большого количества простых моделей, результаты которых затем агрегируются .

Решающее дерево $a(.)$ для задачи регрессии, строится по следующему алгоритму: Имеем обучающую выборку $D = \{x_i, y_i\}$, где x_i - вектор признаков для объекта с номером i, y_i - правильный ответ для объекта с номером i.

1. На каждом шаге текущее множество $C = \{x_i, y_i\}$ (изначально $C = D$) делится на два множества L и R, так чтобы минимизировать выражение вида: $\frac{N_l}{N_c}\sum_{i=1}^{N_l}\frac{(y_i-\overline{Y_l})^2}{N_l-1} + \frac{N_r}{N_c}\sum_{i=1}^{N_r}\frac{(y_i-\overline{Y_r})^2}{N_r-1}$, где слагаемые это несмещенные взвешенные оценки для дисперсий целевой переменной y в множествах L и R соответственно.
2. Выполняется шаг 1 для каждого из полученных на предыдущем шаге множеств.

Данный процесс продолжается до определенного момента. Разбиение происходит по значению одного из числовых признаков x_i. Те объекты, у которых значение выбранного признака меньше порога попадают в множество L; объекты у которых значение выбранного признака больше порога попадают в множество R. Полученный список порогов используется на новом объекте x_{new} для занесения его в какое-то множество Z, которое является одним из множеств на которые разбилась обучающая выборка D. Предсказание для нового объекта есть среднее целевой переменной y в множестве Z, т.е $a(x_{new}) = \overline{Y_z}$.

В задаче прогнозирования временных рядов для y_t признаками могут быть, например, $y_{t-1}, y_{t-2}, \dots, y_{t-12}$ и так далее.

Для объекта x алгоритм случайного леса делает прогноз следующим образом: $RF(x) = \sum_{i=1}^{N}\frac{a_i(x)}{N}$, где N – количество решающих деревьев (подбирается вручную), $a_i(x)$ – решающее дерево, которое осуществляет каждое разбиение из шага 1 не на всей выборке объектов, а лишь на случайной подвыборке, к тому же на случайно выбранных признаках. Особые правила при разбиении требуются для того, чтобы решающие деревья были простыми моделями.