1 Ответы в листьях регрессионного дерева

Рассмотрим матожидание среднеквадратичной ошибки алгоритма при условии равномерного независимого распределения объектов в тестовой выборке.
$E(MSE(\overline{X})) = \frac{1}{n}E\sum(X_i - P_i)^2)$, где P_i - предсказание на объекте X.
$E\sum(X_i - P_i) = \sum_i \sum_{P_i} \sum_{X_i} P(P_i)P(X_i)(X_i - P_i)^2$
Рассмотрим подвыборку, попавшую в один лист решающего дерева. Предсказание на них - константное. Следовательно,
Пусть дерево возвращает среднее значение объектов, попавших в лист.
$E\sum(X_i - \overline{X})^2 = \sum(X_i - \overline{X})^2$, так как ответ алгоритма $P = \overline{X}$
$\sum(X_i - \overline{X})^2 = \sum(X_i^2 - 2X_i\frac{\sum X_i}{n} + (\frac{\sum X_i}{n})^2)$
Теперь пусть дерево возвращает случайное значение объекта, попавшего в лист
$E\sum(X_i - X_j)^2 = \sum_j \frac{1}{n}\sum_i(X_i - X_j)^2 = \sum_i \frac{1}{n}\sum_j(X_i - X_j)^2 = \sum_i \frac{1}{n}(nX_i^2 - 2X_i\sum X_j + (\sum X_j^2))$
Итак,
$E\sum(X_i - X_j)^2 = \sum_i(X_i^2 - 2X_i\overline{X} + \overline{X^2})$
$E\sum(X_i - \overline{X})^2 = \sum_i(X_i^2 - 2X_i\overline{X} + \overline{X}^2)$

2 Линейные модели в деревьях

В процессе построения регрессионного дерева с константными ответами в листьях (например, со средним объектов в листе) разбиения выбираются на основании MSE-критерия. Другими словами, минимизируется средний квадратичный разброс объектов в листьях. Построенное таким образо дерево будет иметь в листьях объекты, слабо разбросанные относительно своего среднего, то есть, имеющие малую дисперсию.
Известно, что линейные модели плохо работают на обучающей выборке, в которой объекты имеют маленькую дисперсию. Чем меньше дисперсия значений в обучающей выборке, тем больше дисперсия коэффициентов линейной модели.
Чтобы получить хорошо работающие линейные модели, надо разбивать объекты так, чтобы подвыборки, попадающие в один лист, имели большую дисперсию.