

МАШИННОЕ ОБУЧЕНИЕ.

ЗАДАНИЕ 2

Студент: Ивахненко Максим Дмитриевич

Группа: 491

Задача 1. Какая стратегия поведения в листьях регрессионного дерева приводит к меньшему матожиданию ошибки по MSE : отвечать средним значением таргета на объектах обучающей выборки, попавших в лист, или отвечать таргетом для случайного объекта из листа (считая все объекты равновероятными)?

Доказательство. Рассмотрим лист дерева.

Пусть в нем находится n объектов.

Обозначим случайно взятый элемент в листе y^{rand} .

Тогда:

$$E \sum_{i=1}^n (y_i - y^{rand})^2 - E \sum_{i=1}^n (y_i - \bar{y})^2 = n[(Ey_1^2 - 2E[y_1 y^{rand}] + E[y^{rand}]^2) - (Ey_1^2 - 2E[y_1 \bar{y}] + E[\bar{y}]^2)] \quad (1)$$

воспользовались линейностью и одинаковораспределенностью величин.

Из равновероятности следует:

$$E[y_1 y^{rand}] = E[y_1 \frac{1}{n} \sum_{i=1}^n y_i] = E[y_1 \bar{y}] \quad (2)$$

Из 1 и 2 следует:

$$E \sum_{i=1}^n (y_i - y^{rand})^2 - E \sum_{i=1}^n (y_i - \bar{y})^2 = n[E[y^{rand}]^2 - E[\bar{y}]^2] =$$

$$E \sum_{i=1}^n y_i^2 - \frac{1}{n} E[\sum_{i=1}^n y_i^2 + \sum_{i \neq j} y_i y_j] =$$

$$[Ey_1]^2 - Ey_1^2 - n([Ey_1]^2 - Ey_1^2) = Dy_1(n-1)$$

Т.е. ошибка при выборе среднего \leq ошибки при выборе случайного. \square

Задача 2. Одна из частых идей – попытаться улучшить регрессионное дерево, выдавая вместо константных ответов в листьях ответ линейной регрессии, обученной на объектах из этого листа. Как правило такая стратегия не дает никакого ощутимого выигрыша. Попробуйте объяснить, почему? Как стоит модифицировать построение разбиений в дереве по MSE , чтобы при разбиении получались множества, на которых линейные модели должны работать неплохо?

Доказательство. При разбиении листьев дерева мы минимизируем значение функции

$$G(L, R) = \frac{|L|}{|Q|} H(L) + \frac{|R|}{|Q|} H(R) \rightarrow \min$$

$$H(M) = \frac{1}{|M|} \sum_{i \in M} (\bar{y}_M - y_i)^2.$$

Таким образом разбиение подбирается так, чтобы максимально приблизить все значения y_i к \bar{y}_M . Чтобы работала регрессионная модель, надо \bar{y}_M в функции $H(M)$ заменить на уравнение прямой $y = kx + b$, подбирая параметры k и b . \square

Задача 3. *Unsupervised* решающие деревья можно было бы применить для кластеризации выборки или оценки плотности, но проблема построения таких деревьев заключается в введении меры информативности. В одной статье предлагался следующий подход – оценивать энтропию множества S по формуле:

$$H(S) = \frac{1}{2} \ln((2\pi e)^n |\Sigma|)$$

Здесь Σ – оцененная по множеству матрица ковариаций. Т.е. не имея других сведений, в предложенном подходе мы по умолчанию считаем, что скопления точек можно приближенно считать распределенными нормально. Убедитесь, что это выражение в самом деле задает энтропию многомерного нормального распределения.

Доказательство. Энтропия многомерного нормального распределения:

$$\int_{R^n} f(x) \ln f(x) = \frac{1}{2} E((x - \mu)^T (x - \mu)) + \ln(|\Sigma|^{\frac{1}{2}} (2\pi)^{\frac{n}{2}}) =$$

$$\frac{1}{2} E(\Sigma_{i,j} (x - \mu)_i (x - \mu)_j \Sigma_{i,j}^{-1}) + \ln(|\Sigma|^{\frac{1}{2}} (2\pi)^{\frac{n}{2}}) = \frac{n}{2} + \frac{1}{2} \ln(|\Sigma| (2\pi)^n) = H(S)$$

так как $\Sigma_{i,j} = (x - \mu)_i (x - \mu)_j$.

□