Машинное обучение.
Задание 2

Пашментов Никита

1 Ответы в листьях регрессионного дерева

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)?

1.1 Решение

Пусть в лист попали объекты X_i с значениями y_i.
$\xi = \frac{1}{n}\sum\limits_{i=1}^{n} y_i$
$E\sum\limits_{i=1}^{n}(y_i-\xi)^2 = E\sum\limits_{i=1}^{n} y_i^2 - E\sum\limits_{i=1}^{n} 2y_i\xi + E\sum\limits_{i=1}^{n}\xi^2$
$\eta = y_\theta$, где θ - дискретное равномерное распределение.
$E\sum\limits_{i=1}^{n}(y_i-\eta)^2 = E\sum\limits_{i=1}^{n} y_i^2 - E\sum\limits_{i=1}^{n} 2y_i\eta + E\sum\limits_{i=1}^{n}\eta^2$
Утверждение: $E\sum\limits_{i=1}^{n}(y_i-\xi)^2 \leq E\sum\limits_{i=1}^{n}(y_i-\eta)^2$
Т.к. $E\xi = E\eta = \frac{1}{n}\sum\limits_{i=1}^{n} y_i$, то $E\sum\limits_{i=1}^{n} 2y_i\xi = E\sum\limits_{i=1}^{n} 2y_i\eta$. Следовательно, утверждение эквивалентно следующему: $E\xi^2 \leq E\eta^2$
$E\eta^2 \geq (E\eta)^2 = (E\xi)^2 = E\xi^2$
Ответ: Лучше отвечать средним значением таргета.

2 Линейные модели в деревьях

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?

2.1 Решение

Критерий построения разбиений в регрессионном дереве никак не учитывает то, насколько в каждой из дочерних ветвей зависимость близка к линейной(можно взять точки на прямой с большим расстоянием друг от друга). Тогда в качестве меры "хорошести"множества можно использовать, например, MSE модели $a(x)$, обученной на данном множестве, т.е.:

$$F(Q) = \frac{1}{|Q|} \sum_{x_j \in Q} (y_i - a(x_i))^2$$

И подставить в формулу для критерия разбиения $F(Q) - \frac{|L|}{|Q|} F(L) - \frac{|R|}{|Q|} F(R)$

3 Unsupervised decision tree

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:

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

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.

3.1 Решение

По определению
$H(f) = -\int\limits_{x \in \mathbb{R}^n} f(x) \ln(f(x)) dx = -E \ln(f(x))$
$\ln\left(\frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} e^{\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}\right) = -\ln\left((2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}\right) - \frac{1}{2}(x-\mu)^T |\Sigma|^{-1}(x-\mu)$

$E\ln(f(x)) = -\ln((2\pi)^{\frac{1}{2}}|\Sigma|^{\frac{1}{2}}) - \frac{1}{2}E(x-\mu)^T|\Sigma|^{-1}(x-\mu)$
Где $\frac{1}{2}E(x-\mu)^T|\Sigma|^{-1}(x-\mu) = -\frac{1}{2}tr(\Sigma\Sigma^{-1}) = -\frac{1}{2}n$
Итого $H(f) = -E\ln(f(x)) = \frac{n}{2} + \ln((2\pi)^{\frac{n}{2}})|\Sigma|^{\frac{1}{2}} = \frac{n}{2} + \frac{1}{2}\ln((2\pi)^n|\Sigma|) = \frac{1}{2}\ln((2\pi e)^n|\Sigma|)$