Машинное обучение – 2. Теоретические задачи

Каргальцев Степан

09.03.2017

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)?*

   Заметим, что $E(MSE) = E(E(MSE \mid Leaf)$, где $Leaf$ – случайная величина, говорящая что текущий объект попадает в лист $Leaf$ при прохождении дерева.

   Тогда достаточно показать требуемое утверждение для условного матожидания внутри. В дальнейшем условие я буду его опускать, но подразумевать.

   Обозначим первый алгоритм (отвечать средним) за $a_1(x)$, а второй (отвечать случайным) за $a_2(x, \omega)$ (ω имеет равномерное распределение на $\{1, \dots, n\}$ (где n – количество объектов в листе) и ни от чего не зависит).

   $E(y - a_2(x))^2 = E(E((y - a_2(x, \omega))^2 \mid \omega)) = \frac{1}{n}\sum\limits_{i=1}^{n} E(y - y_i)^2 = E(\sum\limits_{i=1}^{n}(y - y_i)^2)$. ($y_i$ — ответ на i-м объекте)

   То есть достаточно показать, что $(y - \overline{y})^2 \leqslant \frac{1}{n}\sum\limits_{i=1}^{n}(y - y_i)^2$

   $$\Leftrightarrow$$

   $y^2 - 2\sum y\overline{y} + \overline{y}^2 \leqslant y^2 - \frac{2}{n}y\sum y_i + \frac{\sum y_i^2}{n} \Leftrightarrow \overline{y}^2 \leqslant \overline{y^2}$

   А последнее это неравенство Йенсена

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?*

   Видимо, проблема в том, что мы строим дерево, уменьшая ошибку ответа средним. Таким образом, если строить линейную модель в листьях, то мы будем ~~работать не по профессии~~ отвечать не тем, на что обучались.

   Модифицировать построение разбиений, видимо, стоит так, чтобы вместо минимизации вариации минимизировать матожидание ошибки ответом линейной модели.

   К сожалению, препятствием к содержательным и формальным рассуждениям в этой задаче стал мой poor time management skill, так что на этом океане рассуждений я и закончу повествование посвященное второй задаче.

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:*

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

   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

$$H=-\int\left\{\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}\ln\left[\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}\right]\right\}dx=$$

$$-\int\left\{\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}e^{-\frac{1}{2}x^T\Sigma^{-1}x}\ln\left[\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}e^{-\frac{1}{2}x^T\Sigma^{-1}x}\right]\right\}dx=$$

$$-\int\left\{\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}e^{-\frac{1}{2}x^T\Sigma^{-1}x}\cdot(-\frac{1}{2}x^T\Sigma^{-1}x)\right\}dx-\ln\left[\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}\right]\cdot\int\left\{\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}e^{-\frac{1}{2}x^T\Sigma^{-1}x}\right\}dx=$$

$$\frac{1}{2}\ln\left[(2\pi)^n|\Sigma|\right]-\int\left\{\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}e^{-\frac{1}{2}x^T\Sigma^{-1}x}\cdot(-\frac{1}{2}x^T\Sigma^{-1}x)\right\}dx$$

Осталось показать, что

$$\int\left\{\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}e^{-\frac{1}{2}x^T\Sigma^{-1}x}\cdot(x^T\Sigma^{-1}x)\right\}dx=n$$

Приступим.

$$x^T\Sigma^{-1}x=\sum_{i,j}x_i\cdot x_j(\Sigma^{-1})_{i,j}$$

$$\int\frac{x_ix_j}{(2\pi)^{n/2}|\Sigma|^{1/2}}e^{-\frac{1}{2}x^T\Sigma^{-1}x}dx=\Sigma_{i,j}$$

$$\Sigma=\Sigma^T\Rightarrow\Sigma\cdot(\Sigma^{-1})^T=E$$

$$n=\sum_i E_{i,i}=\sum_i\sum_k\Sigma_{i,k}((\Sigma)^{-1})^T_{k,i}=\sum_i\sum_k\Sigma_{i,k}(\Sigma^{-1})_{i,k}=\sum_{i,j}\Sigma_{i,j}(\Sigma^{-1})_{i,j}$$

Собираем:

$$\int\left\{\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}e^{-\frac{1}{2}x^T\Sigma^{-1}x}\cdot(x^T\Sigma^{-1}x)\right\}dx=\sum_{i,j}\Sigma_{i,j}(\Sigma^{-1})_{i,j}=n\Rightarrow$$

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

Что и требовалось.