Машинное обучение

Семинар 5. Задачи на выбор метрик в практических кейсах.

30 января 2018

Задача 1

Классификация текстов на срочные и не срочные. Те, что вы объявите срочными, будет смотреть тех. поддержка. Тех. поддержка может смотреть 1000 текстов в сутки. Вам поступает 2000/10000/100000 текстов в сутки.

Задача 1. Решение

Количество текстов в сутки и размер тех. поддержки определяют какую долю текстов можно пометить как важные, то есть P(текст помечен важным) задаётся условием задачи.

Нам требуется задетектировать как можно больше действительно важных текстов. то есть максимизировать P(текст помечен важным | текст важный) при условии, что P(текст помечен важным) $= \alpha$, где α определяется условиями задачи (в нашем случае это 0.5, 0.1 и 0.01).

Итоговая метрика — recall при таком пороге, что доля помеченных текстов равна α.

Задача 2

Классификация текстов на срочные и не срочные. Те, что вы объявите срочными, будет смотреть тех. поддержка. Вам поступает 2000/10000/100000 текстов в сутки. Вы хотите максимально уменьшить количество нанятых людей, но при этом обрабатывать 99% всех важных случаев.

Задача 2. Решение

Обрабатывать 99% всех важных случаев означает, что $P(\text{текст помечен важным} \mid \text{текст важный}) \geqslant 0.99$.

Количество нанятых людей пропорционально $P(\text{текст помечен важным})$.

Таким образом, нам требуется минимизировать $P(\text{текст помечен важным})$ при условии, что $P(\text{текст помечен важным} \mid \text{текст важный}) \geqslant 0.99$.

Оптимизируемое выражение равно

$$\frac{P(\text{текст помечен важным} \mid \text{текст важный})}{P(\text{текст важный} \mid \text{текст помечен важным})} P(\text{текст важный}) =$$

$$= \frac{\text{recall}}{\text{precision}} P(\text{текст важный})$$

Задача 2. Решение

$$\frac{\text{recall}}{\text{precision}} \text{P(текст важный)} \to \mathsf{min} \text{ при условии, что recall} \geqslant 0.99$$

Так как обычно precision монотонно убывает от recall, то максимум достигается при recall $= 0.99$

$$\frac{\text{recall}}{\text{precision}} \text{P(текст важный)} \to \mathsf{min} \text{ при условии, что recall} = 0.99$$

P(текст важный) не зависит от выбора порога, а recall $= 0.99$, то задача становится:

$$\frac{0.99}{\text{precision}} \to \mathsf{min} \text{ при условии, что recall} = 0.99$$

или

$$\text{precision} \to \mathsf{max} \text{ при условии, что recall} = 0.99$$

Задача 2. Решение

$$\text{precision} \to \mathsf{max} \text{ при условии, что recall} = 0.99$$

Итоговая метрика — precision при таком пороге, что recall $= 0.99$.

Задача 3

Предсказание спроса на товар. В зависимости от вашего предсказания, на склад завезут разное количество товара. За продажу одной единицы товара вы получаете 200 рублей. Покупка и хранение одной единицы товара стоит 120 рублей. Товар скоропортящийся и если вы не продатите, то товар можно считать потерянным.

Задача 3. Решение

Нужно оценить, сколько мы потеряем денег в случаях, когда мы завезли меньше спроса, и когда больше спроса. Пусть завозят a(x) единиц продукции, а реальный спрос составляет y единиц продукции. Если завезено больше продукции, то она испортится и мы потратим лишние 120 рублей на каждую единицу товара. Если завозим меньше, то недопродаём товар и недозарабатываем 80 рублей с каждой единицы. Таким образом, функция потерь следующая:

$$\begin{cases} 80(y - a(x)), & \text{если } y > a(x) \\ 120(a(x) - y), & \text{если } y \leqslant a(x) \end{cases}$$

Задача 4

Предсказание минимального эффекта воздействия. Вам нужно гарантировать, что настоящий эффект будет не меньше предсказанного вами в 95% случаев.

Задача 4. Решение

При помощи оптимизации квантильной функции потерь $(\alpha\ (a(x_i) - y_i)\ I\{a(x_i) \geqslant y_i\} + (1 - \alpha)\ (y_i - a(x_i))\ I\{a(x_i) < y_i\})$ мы можем оценивать условные квантили целевой переменной, тогда

$$P(y > a(x)) \approx P(y > Z_{1-\alpha}(y \mid x)) = \alpha,$$

то есть, мы можем давать нижние оценки на целевые переменные, которые будут нарушаться с вероятностью α

Задача 5

Предсказание эффекта воздействия. Вам нужно предсказывать доверительный интервал для величины, в который она будет попадать с вероятностью 95%.

Задача 5. Решение

Аналогично предыдущей задаче, мы можем оценивать квантили условного распределения, но теперь обучим две модели a_1 и a_2 с параметрами α_1 и α_2, тогда

$$\mathrm{P}(\mathrm{a}_1(\mathrm{x}) < \mathrm{y} < \mathrm{a}_2(\mathrm{x})) \approx \mathrm{P}(\mathrm{Z}_{1-\alpha_1}(\mathrm{y} \mid \mathrm{x}) < \mathrm{y} < \mathrm{Z}_{1-\alpha_2}(\mathrm{y} \mid \mathrm{x})) = \alpha_1 - \alpha_2.$$

Какие технические проблемы могут возникнуть при таком подходе?

Задача 6

Предсказание среднего величины. Вам нужно прогнозировать математическое ожидание некой величины на следующий день

Задача 6. Решение

Как мы знаем,

$$\sum_{i=1}^{n} (a(x_i) - y_i)^2 \to \mathsf{min} \implies a(x_i) \approx E(y \mid x = x_i)$$

Поэтому нас интересует оптимизация MSE, а целевая переменная это значение требуемой величины в следующий день.