Высшая Школа Экономики, ФКН, ФТиАД

Введение в соревнования по машинному обучению, Никита Чуркин, 2021 г.

Домашнее задание №1 (теоретическая часть): метрики, валидация и признаки

Максимально возможное число баллов за работу – 10 баллов. В скобках после названия задачи указан максимальный балл за задачу.

Обозначения для данного документа не приводятся: считаем, что далее в тексте все обозначения полностью понятны из контекста (или указаны в самой задаче). Мы также используем нотацию Айверсона.

Задачи, в которых не требуется написания кода, можно оформлять любым способом (написать от руки и отсканировать/сфотографировать), использовать любой текстовый редактор или LaTeX. Остальные задачи принимаются в формате .ipynb (с комментариями в ячейках) или .py (с документом-описанием).

1. (A)symmetric MAPE (1 балл).

В задачах регрессии часто используется метрика MAPE (mean absolute percentage error). Данная метрика не является симметричной: она по-разному штрафует завышенные и заниженные предсказания. Чтобы (в том числе) исправить данную особенность, была придумана метрика

$$\text{sMAPE}(y, \hat{y}) := \frac{100}{n} \sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{|y_i| + |\hat{y}_i|}$$

Покажите, что метрика sMAPE не является симметричной в указанном выше смысле (достаточно числового примера); проиллюстрируйте ваши рассуждения с помощью соответствующего графика.

2. Повысьте мой ROC AUC, пожалуйста (1 балл).

Для задачи классификации некоторой моделью построены предсказания $\hat{y}$. Мы хотим увеличить метрику ROC AUC для данной задачи с помощью аффинного преобразования, т.е. найти коэффициенты $\alpha, \beta \in \mathbb{R}$ такие, что

$$\text{AUC}(y, \hat{y}) < \text{AUC}(y, \alpha\hat{y} + \beta)$$

Какие значения коэффициентов приводят к желаемому результату?

3. Равенство метрик (2 балла).

Для задачи бинарной классификации по предсказанию модели подсчитаны метрики Accuracy, F1 score, Precision, Recall. Так получилось, что все они в данной задаче оказались равными одному числу α.

Чему может равняться α? Какой баланс классов может быть в данной задаче?

4. KNN-признаки (3 балла).

Запрограммируйте алгоритм построения KNN-based признаков, рассмотренных на лекции. Можно вычислять любые подвиды KNN-признаков, например:

- значение j-го признака $x^j_{neighbor(i)}$ у ближайшего соседа (для рассматриваемого объекта x_i).
- разница $x^j_i - x^j_{neighbor(i)}$ для количественного признака или индикатор $[x^j_i = x^j_{neighbor(i)}]$ для категориального.
- расстояние до ближайшего соседа: $d(x_i, x_{neighbor(i)})$.

Для реализации можно пользоваться библиотеками *sklearn*, *annoy* и т.п.

Выберите любые 2 алгоритма из списка: логистическая регрессия, случайный лес, градиентный бустинг, MLP.

Проведите ряд экспериментов на любом *непопулярном* датасете (Титаник, UCI-датасеты запрещены). Какие из полученных признаков наиболее полезны в вашей задаче? Какие алгоритмы лучше всего реагируют на добавление KNN-based признаков?

5. Отбор признаков (3 балла).

Запрограммируйте любой из методов отборов признаков.

Выберите любой алгоритм из списка: случайный лес, градиентный бустинг, MLP.
Проведите ряд экспериментов на любом *непопулярном* датасете (Титаник и UCI-датасеты запрещены) с целью ответить на следующие вопросы:

- Помогает ли отбор признаков в данной задаче? Метрика качества увеличивается или снижается?
- Следует ли делать отбор признаков до или после тюнинга гиперпараметров?
- Добавьте в признаки шум (например, сгенерировав колонку из белого шума или перемешанную колонку другого признака).

  Справляется ли реализованный алгоритм с удалением шума?