Бонусная домашка с задачами

Для самых поразительных студентов совбака

2 ноября 2020 г.

Дедлайн: 21 ноября в 21:00.

Все задачки — бонусные. Каждая весит 1 балл! Максимум — 4.

Решения расписывать подробно! Очень! А то не поверим. Формулы без комментариев не засчитаем.

Решения сдавать в Tex или Markdown.

Веселитесь :)

1. Рассмотрим задачу обучения линейной регрессии. Все обозначения можно вспомнить из материалов лекций :)

$$Q(w) = (y - Xw)^T(y - Xw) \to \min_w$$

Будем решать её с помощью градиентного спуска. Допустим, мы находимся на некоторой итерации k, и хотим выполнить очередной шаг

$$w^{(k)} = w^{(k-1)} - \alpha \nabla_w Q(w^{(k-1)}).$$

При известных y, X, $w^{(k-1)}$ найдите длину шага α, минимизирующую функционал ошибки на шаге k:

$$Q(w^{(k-1)} - \alpha \nabla_w Q(w^{(k-1)})) \to \min_\alpha.$$

2. Найдите константу c, минимизирующую квантильную функцию потерь ($0 < \tau < 1$ фиксировано):

$$\sum_{i=1}^{\ell} \rho_\tau(y_i - c) \to \min_c,$$

$$\rho_\tau(x) = \begin{cases} \tau x, & x > 0, \\ (\tau - 1)x, & x \leqslant 0. \end{cases}$$

3. В анализе данных для сравнения среднего значения некоторой величины у объектов двух выборок часто используется критерий Манна–Уитни–Уилкоксона[1], основанный на вычислении U-статистики.

[1] https://en.wikipedia.org/wiki/Mann–Whitney_U_test

Пусть у нас имеется выборка X и классификатор $b(x)$, возвращающий оценку принадлежности объекта x положительному классу. Тогда вычисление U-статистики для подвыборки X, состоящей из объектов положительного класса, производится следующим образом: объекты обеих выборок сортируются по неубыванию значения $b(x)$, после чего каждому объекту в полученном упорядоченном ряду $x_{(1)}, \ldots, x_{(\ell)}$ присваивается ранг — номер позиции $r_{(i)}$ в ряду (начиная с 1, при этом для объектов с одинаковыми значением $b(x)$ в качестве ранга присваивается среднее значение ранга для таких объектов). Тогда U-статистика для объектов положительного класса равна:

$$U_+ = \sum_{\substack{i=1 \\ y_{(i)}=+1}}^{\ell} r_{(i)} - \frac{\ell_+(\ell_+ + 1)}{2}.$$

Здесь ℓ_-, ℓ_+ — количество объектов. отрицательного и положительного классов соответственно

Покажите, что для значения AUC-ROC классификатора $b(x)$ на выборке X и U-статистики верно следующее соотношение:

$$\mathsf{AUC} = \frac{U_+}{\ell_-\ell_+}.$$

4. При построении деревьев критерий информативности для набора объектов R вычисляется на основе того, насколько хорошо их целевые переменные предсказываются константой (при оптимальном выборе этой константы):

$$H(R) = \min_{c \in Y} \frac{1}{|R|} \sum_{(x_i, y_i) \in R} L(y_i, c),$$

где $L(y, c)$ — некоторая функция потерь. Соответственно, чтобы получить вид критерия при конкретной функции потерь, необходимо аналитически найти оптимальное значение константы и подставить его в формулу для $H(R)$.

Выведите критерий информативности для следующей функции потерь:

$$L(y, c) = \sum_{k=1}^{K} (c_k - [y = k])^2$$