Машинное обучение. DS-поток.

Задание 3.

- Дедлайн **06 марта 02:00**. После дедлайна работы не принимаются кроме случаев наличия уважительной причины.
- Выполненную работу нужно отправить на почту `mipt.stats@yandex.ru`, указав тему письма "`[ml] Фамилия Имя - задание 3`". Квадратные скобки обязательны. Если письмо дошло, придет ответ от автоответчика.
- Задачи необходимо оформить в tex'е и прислать `pdf` или же прислать фотку в правильной ориентации рукописного решения, где все четко видно.
- Решения, размещенные на каких-либо интернет-ресурсах не принимаются. Кроме того, публикация решения в открытом доступе может быть приравнена к предоставлении возможности списать.
- Не забывайте делать пояснения и выводы.

1. **(2 балла)** Докажите, что критерий Джини равен вероятности ошибки случайного классификатора, который выдаёт предсказания с вероятностями пропорционально доле классов в выборке.

2. **(2 балла)** Пусть $X = (x_1, ..., x_n)$ — выборка объектов и $Y = (Y_1, ..., Y_n)$ — соответствующие значения вещественного отклика. Критерий информативности для набора объектов вычисляется на основе того, насколько хорошо их отклик предсказывается константой:

$$H(X) = \min_{c \in Y} \frac{1}{n} \sum_{i=1}^{n} L(Y_i, c),$$

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 $H(X)$.

Выведите критерии информативности для следующих функций потерь:

(a) $L(y, c) = (y - c)^2$;

(b) $L(y, c) = |y - c|$.

Найдите также оптимальное предсказание в листьях дерева.

3. **(4 балла)** В случае задачи классификации рассматривается вероятностное предсказание, и критерий информативности имеет вид

$$H(X) = \min_{\substack{p_1,...,p_K \in [0,1] \\ p_1+...+p_K=1}} \frac{1}{n} \sum_{i=1}^{n} L\left(Y_i, \{p_k\}\right).$$

Выведите критерии информативности и найдите оптимальные оценки вероятностей в листах для следующих функций потерь:

(a) $L(y, \{p_k\}) = \sum_{k=1}^{K} (p_k - I\{y = k\})^2$;

(b) $L(y, \{p_k\}) = -\sum_{k=1}^{K} I\{y = k\} \log p_k$.

4. **(3 балла)** Запишите оценку сложности построения одного решающего дерева в зависимости от размера обучающей выборки n, числа признаков d, максимальной глубины дерева D. В качестве правил используются пороговые функции $I\{x_j > t\}$. При выборе правил в каждой вершине перебираются все признаки, а в качестве порогов рассматриваются величины t, равные значениям этого признака на объектах, попавших в текущую вершину.

5. **Скоро будет.**