Высшая Школа Экономики, Введение в соревнования по машинному обучению

Домашнее задание №2 (теоретическая часть): Ансамбли, лотереи, паззлы и утечки

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

1. Ансамбли линейных моделей (1 балл).

Имеет ли смысл строить ансамбль из линейных моделей, если мета-модель нелинейная? А если линейная? (Под мета-моделью здесь понимается модель, строящаяся на out-of-fold предсказаниях базовых моделей).

2. ShakeUp-мера соревнования (2 балла).

Предложите несколько (не менее 3-х) *количественных* метрик для оценки shake up-а соревнований (очевидный пример: средняя разница между положением участников на Public LB и Private LB). Опишите сильные и слабые стороны предлагаемых метрик, а также вычислите ваши меры для нескольких конкурсов, рассмотренных на лекциях.

Прокомментируйте результаты и расскажите, как на ваш взгляд можно отличить случайный результат в соревновании от правильного решения?

3. Паззлы с матрицами (3 балла: 1а, 2б).

а) (А. И. Кострикин) Целые числа 1798, 2139, 3255, 4867 делятся на 31. Без всяких вычислений показать, что определитель

$$\begin{vmatrix} 1 & 7 & 9 & 8 \\ 2 & 1 & 3 & 9 \\ 3 & 2 & 5 & 5 \\ 4 & 8 & 6 & 7 \end{vmatrix}$$

также делится на число 31.

б) Докажите, не используя *компьютер* для расчетов, что определитель

$$\begin{vmatrix} 51237 & 79922 & 55538 & 39177 \\ 46152 & 16596 & 37189 & 82561 \\ 71489 & 23165 & 26563 & 61372 \\ 44350 & 42391 & 91185 & 64809 \end{vmatrix}$$

отличен от нуля.

4. Поиск похожих строк (4 балла).

Часто (в том числе для поиска data leak'а) требуется эффективный алгоритм мэтчинга похожих строк в таблице.

Дана матрица (таблица) размера $M \times N$ с элементами из любого множества и фиксированное натуральное число k, т.ч. $0 < k < N$.

На вход алгоритму подается индекс i строки из матрицы. На выходе требуется вывести все индексы j строк матрицы, которые отличаются от строки с индексом i не более, чем в k столбцах.

Эффективно реализуйте данный алгоритм, протестируйте его на сгенерированных данных и оцените временную сложность в терминах M, N, k.