

ИНФОРМАТИКА И ИКТ

ПРОФИЛЬ «ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ»

120 минут

Задания выставлены на платформу Яндекс.Контест, настроена автоматическая проверка заданий, ссылка:

<https://contest.yandex.ru/contest/85721/enter/?retPage=standings>.

Требуется обязательный доступ к яндекс диску в месте проведения олимпиады.

Общая сумма баллов: 100.

Для выполнения заданий по программированию рекомендуется использовать следующие языки программирования: C++, Python, Java, Pascal, C#. Для решения задачи с табличными данными требуется табличный процессор.

Требуется ограничить в браузерах выход на другие сайты и ресурсы для обучающихся при прохождении олимпиады. Разрешен доступ только на платформу Яндекс.Контест и Яндекс-диск.

Разбор задач

Задача 1. Ансамбль решающих деревьев

В ансамбле (комитете) из 5 независимых моделей решающих деревьев проводится голосование по бинарному вопросу («да»/«нет»). После принятия решения каждое дерево даёт 1 балл за выбранный вариант. Каждое дерево с вероятностью 0,7 принимает правильное решение. После голосования, если сумма баллов $\geq T$ (порог) у варианта, у которого больше баллов, то принимается этот вариант, иначе ансамбль выдаёт ошибку. Какое целое T необходимо поставить, если нужна максимальная вероятность правильного решения ансамбля при условии, что вероятность неправильного решения не должна оказаться больше 0,1?

Баллы: 10

Задача 2. Байесовский классификатор спама

Пользователь хочет сделать свой фильтр спама по содержимому, имитирующий Байесовский. Он собрал информацию, что 20% всех приходящих ему писем — спам.

Слово «акция» встречается в 80% спам-писем и только в 10% не-спам-писем. Не разобравшись, пользователь решил просто поставить вероятность отправки в спам письма, если в нём есть слово «акция». Найдите, при каком пороговом значении вероятности $P(\text{в_спам} | \text{«акция»})$ цена ошибки «пропустить спам» (не заметить спам) в 3 раза выше, чем цена ошибки «ложное срабатывание» (пометить хорошее письмо как спам)

(Ценность оценивается в количестве тех или иных писем, т.е. пропустить 1 спам-письмо также плохо, как и отправить в спам 3 не-спам-письма).

Баллы: 12

Задача 3. Случайный лес и информативность признаков

В «Случайном лесу» для выбора лучшего признака при разбиении узла случайно выбирается подмножество из k признаков из общего числа n . Пусть существует ровно 3 «сильных» признака, которые идеально разделяют данные. Какова вероятность, что среди случайно выбранных k признаков окажется хотя бы один из этих трех, если $k=5$, $n=20$?

(Случайный лес — это популярный алгоритм машинного обучения, который используется для задач классификации (например, «отнести рекламу к запрещённой или допустимой») и регрессии по принципу использования различных вариантов сравнения / решения (признаков) и выбора общего. Узлы – группы данных для анализа лесом, после каждого разбиения на подгруппы (дочерние узлы) в них будут более чистые данные (в одном узле почти вся реклама разрешена, в другом почти вся запрещена))

Баллы: 10

Задача 4. Размещение данных на серверах

ИИ-систему хотят обучить распределять задач по серверам для параллельных вычислений. Для составления первой обучающей выборки необходимо просчитать все возможные варианты распределения 10 идентичных задач по 4 различным серверам. Также было сразу решено ввести ограничение на вместимость серверов: первый сервер

может принять не более 3 задач, второй — не более 4, третий — не более 5, четвертый — не более 6 задач.

а) Сколькими способами можно это сделать, если каждый сервер может получить любое приемлемое количество задач (включая 0)?

б) Сколько существует способов, если каждый сервер должен получить хотя бы одну задачу?

Вследствие идентичности задач для первой обучающей выборки варианты отличаются только количеством задач на том или ином сервере, например, один из вариантов для пункта а) $3 - 4 - 3 - 0$, т.е. первые 2 сервера заполнены, а все остальные задачи попали на третий сервер.

Баллы: 16

Задача 5. Векторное позиционирование автономного робота-строителя

Автономная система ИИ для строительного робота-сборщика анализирует геометрию кубического модуля. Из одной вершины куба исходят три взаимно перпендикулярных вектора (ребра).

Системе идентифицированы два из них:

$$\vec{a} = 7\vec{i} + 6\vec{j} - 6\vec{k}$$

$$\vec{b} = 6\vec{i} + 2\vec{j} + 9\vec{k}$$

Формальная постановка: известно, что векторы $\vec{a} = 7\vec{i} + 6\vec{j} - 6\vec{k}$, $\vec{b} = 6\vec{i} + 2\vec{j} + 9\vec{k}$ могут быть ребрами куба. Найдите координаты третьего вектора.

Решение:

Векторное позиционирование автономного робота-строителя

Найдем третье ребро $\vec{c} = (x; y; z)$ куба. Так как $\vec{a} \perp \vec{c}$, то $\vec{a} * \vec{c} = 0$ т.е, $7x + 6y - 6z = 0$; Так как $\vec{b} \perp \vec{c}$, то $\vec{b} * \vec{c} = 0$ т.е, $6x + 2y + 9z = 0$; из равенств $|\vec{c}| = |\vec{a}| = |\vec{b}| = 11$ вытекает, что $\sqrt{x^2 + y^2 + z^2} = 11$. Для нахождения координат вектора решим систему уравнений

$$\begin{cases} 7x + 6y - 6z = 0 \\ 6x + 2y + 9z = 0 \\ x^2 + y^2 + z^2 = 121 \end{cases}$$

Из первых двух уравнений выражаем x и y через z ($x = -3z, y = \frac{9}{2}z$) и подставляем их значения в третье уравнение системы: $9z^2 + \frac{81}{4}z^2 + z^2 = 121$. Отсюда находим, что: $z = \pm 1$, тогда $x = \pm 6$ и $y = \pm 9$. Следовательно $\vec{c} = \pm(6\vec{i} - 9\vec{j} - 2\vec{k})$.

Ответ: 692

Баллы: 10

Задача 6. Данные для выполнения этого задания находятся в файле электронной таблицы. Вы можете скачать файл в одном из трёх форматов (XLSX, ODS, CSV) по ссылке https://disk.yandex.ru/d/MvnnrySVTQQG_w

Вы — инженер по ML. Ваша команда проводит предобработку данных и EDA (разведочный анализ) перед обучением модели, которая будет классифицировать клиентов банка по кредитоспособности. Этап предобработки данных включает проверку на выбросы, которые могут повлиять на качество модели классификации.

Постановка задачи:

Вам предоставили файл с данными. В первом столбце указана заработная плата клиента банка, а во втором столбце — его категория (*на английском языке*), которая была предварительно присвоена на основе уровня дохода (базовый(C), стандартный(B), премиум(A)).

Определите общее количество выбросов по всем категориям клиентов. Значение считается выбросом, если выполняется неравенство $|x - \bar{x}| > \bar{x}$

$\bar{x}$ — среднее арифметическое значение зарплаты в категории.

x — заработная плата клиента.

Баллы: 14

Задача 7. Валидация цифрового следа

Привет, ИИ-стажёр! Сегодня мы обучаемся алгоритмической задаче на обработку текста: классифицировать строки на основе корректности расстановки скобок. Твоя цель — анализировать входную строку и выводить бинарный результат: "Корректно" (если все пары скобок правильно открыты/закрыты с учётом вложенности) или "Некорректно" (в противном случае). Скобки: круглые () и квадратные [] — ключи к балансу.

Пример 1

Ввод	Вывод
Nkldjdfs dfkjdfs	Корректно

Пример 2

Ввод	Вывод
Некорректно(	Некорректно

Баллы: 14

Задача 8. Модель-классификатор сделала предсказание на валидационной выборке. Популярным способом собрать показатели *precision* и *recall* в одну метрику является метрика *F1*. Напишите программу, для оценки качества модели, используя метрику *F1*.

Метрика *F1* – это среднее гармоническое метрик *precision* и *recall*

TP (true positive)

FP (false positive)

TN (true negative)

FN (false negative)

$F_1 = \frac{2}{\frac{1}{Recall} + \frac{1}{Precision}}$	$Precision = \frac{TP}{TP + FP}$	$Recall = \frac{TP}{TP + FN}$
--	----------------------------------	-------------------------------

Формат входных данных:

- первая строка содержит одно целое число n – количество предсказаний ($n \geq 6$);
- следующие строки содержат истинные и предсказанные значения, разделённые пробелом.

Формат выходных данных:

В ответе укажите значение метрики $F1$, округлив значение до 6 знаков после запятой.

ВВОД	ВЫВОД
7 1 1 1 0 0 0 0 1 1 0 1 1 1 1	0.666667
12 1 1 0 1 1 1 0 0 0 1 1 0 1 0 1 1 0 1 0 0 1 0 1 1	0.571429

Баллы: 14