МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ

(НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)

Институт №8 «Информационные технологии и прикладная математика»

Кафедра 810Б «Информационные технологии в моделировании и управлении»

**Практическое задание №1**

**по курсу «Интелектуальный анализ данных»**

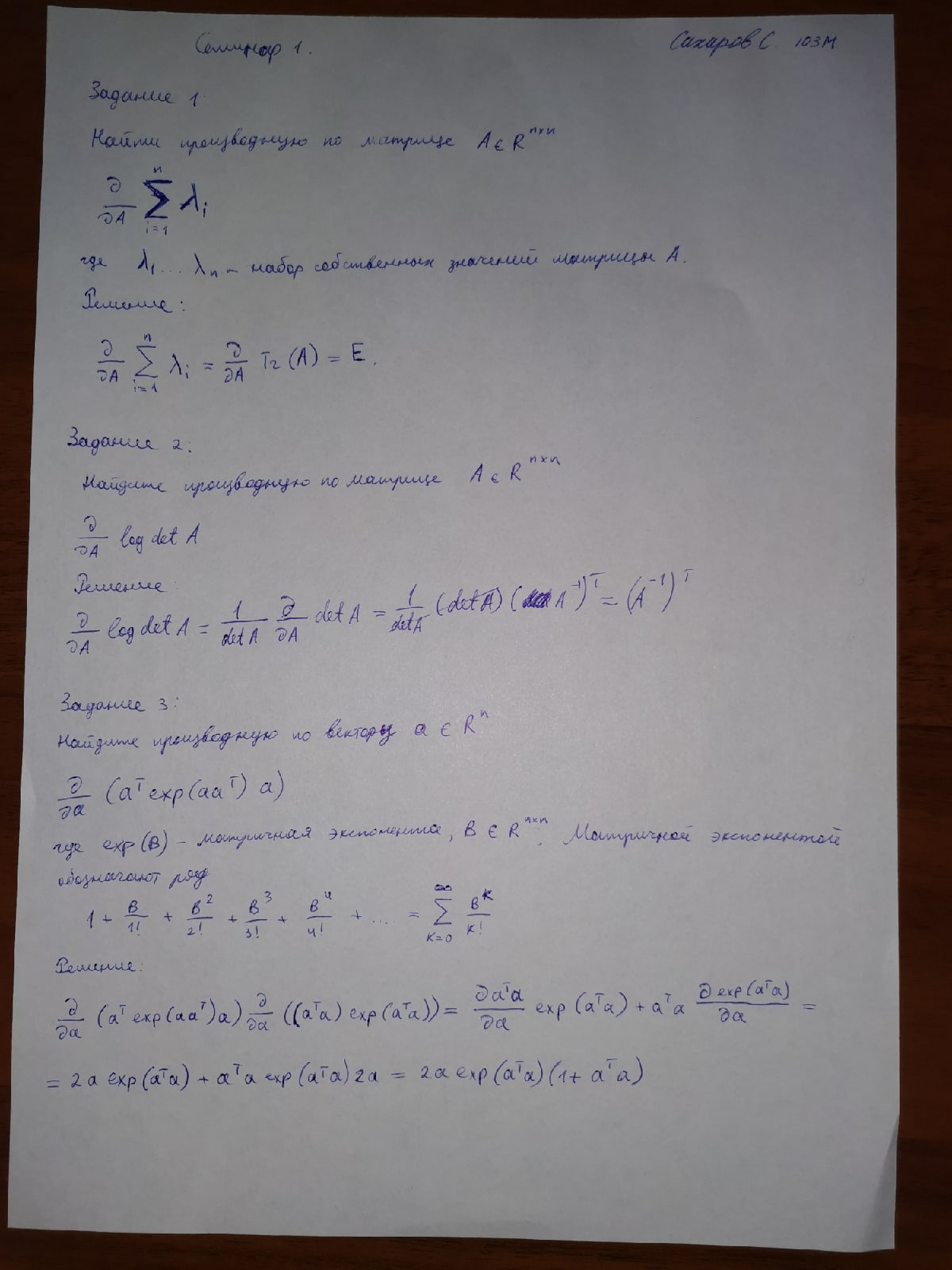
**Матрично-векторное дифференцирование. Условия оптимальности. Линейная регрессия.**

Выполнил: С.В. Сахаров

Группа: М8O-103М-19

Преподаватель: Абгарян К.К.

Москва, 2020



Задание 7.

1) Почему L1 регуляризация производит отбор признаков?

Потребность в отборе признаков может быть обусловлена следующими факторами:

* Наличие в обучающей выборке признаков, не относящихся к задаче;
* На предсказывающую модель могут накладываться ограничения по скорости работы. В такой ситуации следуют отбирать наиболее важные признаки.

2) Почему коэффициент регуляризации нельзя подбирать по обучающей выборке?

Коэффициент регуляризации нельзя подбирать по обучающей выборке потому что он контролирует баланс между подгонкой под обучающую выборку и штрафом за излишнюю сложность. Данный коэффициент подбирается под конкретную задачу.

3) Почему накладывать регуляризатор на свободный коэффициент w0 может быть плохой идеей?

Накладывать регуляризатор на свободный коэффициент w0 является плохой идеей, так как в таком случае будут учитываться априорные представления о близости целевой переменной к нулю и отсутствии необходимости в учете ее смещения.

4) Что такое кросс-валидация, чем она лучше использования отложенной выборки?

Кросс-валидация это метод оценивания точности модели, при котором для теста используется каждый раз разная часть обучающей выборки, которая не использовалась в последней эпохе обучения. Кросс-валидация - это способ оценить способность модели работать на гипотетическом тестовом наборе, когда такой набор в явном виде получить невозможно.

5) Почему категориальные признаки нельзя закодировать натуральными числами? Что такое one-hot encoding?

Категориальные признаки нельзя закодировать натуральными числами потому что так теряется весь смысл категориальности, а также появляются ложные интерпретации. one-hot encoding лишен этих недостатков. Для кодируемого категориального признака создаются N новых признаков, где N — число категорий. Каждый i-й новый признак — бинарный характеристический признак i-й категории.

6) Для чего нужно масштабировать матрицу объекты-признаки перед обучением моделей машинного обучения?

Масштабирование необходимо для приведения всех признаков к 1 масштабу, что бы случайные выбросы не сильно сказывались на процессе обучения моделей машинного обучения.

7) Почему MSE чувствительно к выбросам?

MSE достаточно сильно штрафует за большие отклонения, поэтому и возникает большая чувствительность к выбросам.