Задачи командного тура

В командной части заключительного этапа участники решают реальную задачу анализа данных [1]. Участникам олимпиады предоставлен набор обезличенных данных пациентов клиники “Гамма-Нож” НИИ им. Бурденко. Часть пациентов скрыта и на ней производится проверка. Участникам олимпиады предложено предсказать две характеристики:

- вероятность факта ремиссии;
- среднюю продолжительность жизни.

Набор данных по обоим задачам предоставлен вместе с обезличенными медицинскими показателями. На основе алгоритмов машинного обучения участники делали предсказания для повышения качества медицинской диагностики.

10.1. Описание предметной области

В конце 2016 года в России зарегистрировано 3 518 842 онкологических пациентов. Из них у 599 348 пациентов диагноз поставлен впервые.

Одним из серьезных осложнений онкологического заболевания является метастазирование в головной мозг, что снижает продолжительность жизни. Лечение пациентов с метастатическим поражением головного мозга требует комплексного подхода: хирургии, лекарственной терапии, воздействия на метастатические очаги ионизирующим излучением с помощью высокоточных установок, кобальтовых аппаратов и линейных ускорителей заряженных частиц.

Увеличение продолжительности жизни этой категории пациентов и улучшение ее качества связаны как с совершенствованием различных методов лечения и нейровизуализации, так и с поиском благоприятных прогностических факторов, то есть тех факторов, от которых в большей степени зависит вероятность выздоровления от болезни.

При анализе полученных результатов необходимо учесть множество параметров (клинических, морфологических и т.д.). Способность машинного обучения находить ключевые признаки в сложных наборах данных, устанавливать скрытые закономерности позволит выявить благоприятные прогностические факторы и улучшить наше понимание в оценке эффективности лечебных методик и создать новые инструменты для прогнозирования результатов лечения.

[1] Командная часть заключительного этапа была разработана Николаем Князевым

10.2. Описание актуальной задачи

Участникам олимпиады предоставлен набор обезличенных данных пациентов клиники «Гамма-Нож» НИИ им. Бурденко.

«Гамма-нож» (Leksell Gamma Knife Perfexion) – аппарат для проведения высокоточного одномоментного облучения различных патологических образований головного мозга. Пучки гамма-излучения, порождаемого источниками, с точностью в несколько десятых долей миллиметра сходятся в фиксированной точке внутри радиационного блока – изоцентре установки. Ионизирующее излучение приводит к повреждению ДНК патологических клеток и клеточных мембран, вследствие чего нарушается рост опухоли. Доза облучения достаточно велика для того, чтобы достичь необходимого эффекта после однократной процедуры . Поэтому данный вид лучевого лечения называется радиохирургией, в отличие от радиотерапии – когда больному проводится до 30-40 сеансов небольшими дозами.

Школьникам будет предложено спрогнозировать две характеристики:

- Среднюю продолжительность жизни (Задача №1)
- Вероятность факта ремиссии (Задача №2)

Часть пациентов будет скрыта, и на ней будет производится проверка (валидационная выборка, тестирующий сет). На основании списка характеристик пациентов, таких как «Онкологический диагноз», «Активирующие мутации», «Дата рождения», «Число радиохирургий (РХ) на аппарате Гамма Нож», «Локальный рецидив после радиохирургии / гипофракционирования» и другие (всего около 30 характеристик) предлагается решать предложенные задачи. Не у всех пациентов заполнены все поля, как такие данные использовать корректно - вопрос для участников.

10.3. Задача 1

Задача предсказания средней продолжительности жизни является задачей регрессии и результат оценивается по коэффициенту детерминации. Коэффициент детерминации – это доля дисперсии зависимой переменной, объясняемая рассматриваемой моделью зависимости, то есть объясняющими переменными. Более точно — это единица минус доля не объясненной дисперсии (дисперсии случайной ошибки модели, или условной по факторам дисперсии зависимой переменной) в дисперсии зависимой переменной. Его рассматривают как универсальную меру зависимости одной случайной величины от множества других. В датасете продолжительности жизни содержатся как умершие, так и живые пациенты, как эти данные использовать корректно – вопрос для участников.

Типология данных:

- **HID** – идентификационный номер;
- **Год** - год, в котором лечился пациент
- **Онкологический диагноз (основное заболевание)** – меланома (онкологическое заболевание, развивающееся из пигментных клеток кожи), НМРЛ – немелкоклеточный рак легкого, РП – рак почки, РМЖ – рак молочной железы, КРР – колоректальный рак. Наиболее трудно поддаются лечению – КРР, НМРЛ и меланома.

- **Дата постановки онкологического лечения** – та дата, когда пациент обратился к врачу впервые и ему поставили диагноз.
- **Первичный очаг (дата удаления)** – дата, когда пациенту удалили первичный очаг. Первичный очаг – тот очаг в другом органе, который метастазировал (его клетки по кровеносной и лимфатической системе распространились в головной мозг, где сформировались метастазы)
- **Активирующие мутации** – мутации в гене. Онкологическое заболевание развивается вследствие наличия мутаций в генах. Активирующие мутации – это те мутации, которые специфичны именно для данного вида опухоли.
- **Дата проведения ОВГМ** – дата, когда пациенту проводили облучения всего головного мозга (другой вид лучевого лечения)
- **Дата операции на ГМ** – дата, когда пациенту провели облучение (радиохирургия) на аппарате Гамма – нож
- **Число радиохирургий на аппарате ГМ** – количество проведенных операций (радиохирургий) на Гамма ноже
- **Дата 1 -й РХ** – дата первой радиохирургии на аппарате Гамма -нож (в некоторых случаях пациент несколько раз проходит лечение на аппарате гамма – нож. В дата сете мы учитываем только первую радиохирургию)
- **Индекс Карновского на момент первой РХ** – функциональный статус пациента на момент радиохирургии на гамма ноже. (100 -состояние нормальное, жалоб нет, 10 – умирающий больной). http://www.bionco.ru/tables/carnovskyindex_scal
- **Число очагов в ГМ, подвергнутых РХ** – количество метастазов в головном мозге человека, которые были пролечены за одну радиохирургию на гамма ноже
- **Суммарный объем всех очагов** – сумма объемов всех очагов
- **Объем максимального очага** – объем самого максимального очага
- **Экстракраниальные метастазы на момент радиохирургии/ гипофракционирования** – метастазы, которые есть у пациента в других органах, помимо головного мозга. Гипофракционирование – другой режим лучевого лечения, подведение высокой дозы за несколько фракций.
- **Лекарственное лечение** – лекарственная терапия, которая проводилась пациенту (два варианта – химиотерапия (общая лекарственная терапия) и таргетная терапия (таргетная терапия подобрана в соответствии конкретному типу опухоли. Ее возможно подобрать благодаря наличию активирующих мутаций).
- **Локальный рецидив** – очаг был пролечен на гамма-ноже и возник (рецидивировал) в том же месте вновь
- **Дистантный метастаз** – этого очага раньше не было, он возник независимо от проводимого лечения
- **Интракраниальная (внутримозговая) прогрессия** (лок. Рецидивы и дист. Метастазы) – этот столбец сумма двух предыдущих.
- **Лечение ИК рецидива** – лечение интракраниального рецидива (операция – оп, таргетная терапия, химиотерапия, РХ – вновь провели радиохирургию на этот очаг)
- **OS (overall survival)** – общая выживаемость (1- пациент погиб вследствие

заболевания, 0 – пациент жив)

Набор данных

- База живущих пациентов:
  http://nti-contest.ru/wp-content/uploads/bd/1.1.%20data_pacientAlive.pdf
- База умерших пациентов (тренировочная выборка):
  http://nti-contest.ru/wp-content/uploads/bd/1.2.%20X_train.pdf
- Целевая переменная тренировочной выборки (прогноз по количеству дней жизни):
  http://nti-contest.ru/wp-content/uploads/bd/1.3.%20y_train.pdf
- Бейзлайн:
  https://drive.google.com/open?id=1VbMI3zSaLCEx7Wvk2oQcw9C9MwI2bhES
- Пример корректно отформатированной выдачи под условия проверочного сервера:
  http://nti-contest.ru/wp-content/uploads/bd/1.5.%20submit_example.pdf
- Тестовая выборка:
  http://nti-contest.ru/wp-content/uploads/bd/1.6.%20X_test.pdf

Код проверки и генерации правильного решения

```
import csv
import math
import random
import math
import requests
def csv_reader(file_obj):
        """
        Read a csv file
        """
        res = list()
        reader = csv.reader(file_obj)
        for row in reader:
            res.append(float(" ".join(row)))
          #  print(" ".join(row))
        return res
def mse(true,pred):
        totalError = 0
        if len(pred)!= len(true):
            print("problem array len", len(pred))
            return "ERROR!"
        for i in range(min(len(pred), len(true))):
            totalError+= (true[i]-pred[i])**2
        return totalError
def r2(true,pred):
        mean_a = [sum(true)/len(true)]*len(true)
        r2 = 1 - mse(true,pred)/mse(true,mean_a)
        return r2

#This is a sample Data Challenge
#Learn more: https://stepik.org/lesson/9172

def generate():

        return "Смотри тестовый датасет!"
```

```
35
36  def solve(dataset):
37          #https://stepik.org/media/attachments/lesson/75102/y_test_new_101.csv
38
39          import requests
40          url = "https://stepik.org/media/attachments/lesson/75102/y_test_new_101.csv"
41          r = requests.get(url)
42          task_arr = [str(int(i)) for i in r.content.split()]
43
44          return " ".join(task_arr)
45
46
47  def check(reply, clue):
48          reply = [float(i) for i in reply.split()]
49          url = "https://stepik.org/media/attachments/lesson/75102/y_test_new_101.csv"
50          r = requests.get(url)
51          clue = [int(i) for i in r.content.split()]
52          res = r2(list(clue), list(reply))
53          if res<0:
54              res=0
55          balls = res*100
56          resT = "Вы набрали " + str(balls) +" баллов"
57          return res,resT
```

Пример решения задачи №1 командой школьников Олимпиады НТИ
файл .ipynb:
https://drive.google.com/open?id=1GwJ7I5ff9WTjZW0xvRQ5Iw82BQMoH7ep

файл .html исключительно для визуального просмотра кода:
https://drive.google.com/open?id=12auUuOhYMKUzb0Ipny4Dmibn6joNmybP

10.4. Задача 2

Задача предсказания факта ремиссии является задачей классификации, и результирующее количество баллов равно мере F1. Мера F1 равна

$$F1 = \frac{2 \cdot \text{точность} \cdot \text{полнота}}{\text{точность}+\text{полнота}}$$

Считается относительно класса положительной ремиссии. При этом

$$\text{точность} = \frac{\text{количество верно предсказанных ремиссий}}{\text{общее количество предсказанных ремиссий}}$$

$$\text{полнота} = \frac{\text{количество верно предсказанных ремиссий}}{\text{общее количество ремиссий в тестовой выборке}}$$

Типология данных:

- **Sex** – пол,
- **Localization** – локализация,
- **Cancer histology** – гистология,
- **RS1** – дата радиохирургии,

- **V1** – объем очага,
- **PD1** – предписанная доза,
- **PI1** – предписанная изодоза
- **MV12GY** – объем матрицы, облученный дозой 12 Гр,
- **MV10GY** – объем матрицы, облученный дозой 10 Гр,
- **NV12GY** – объем нормальных тканей, облученный дозой 12 Гр
- **NV10GY** – объем нормальных тканей облученных дозой 10 Гр
- **Рецидив** – облученный очаг возник вновь
- **Время развития рецидива** – время, прошедшее с момента радиохирургии до последнего наблюдения

Локализация

- **BS** – ствол головного мозга
- **Ce** - мозжечок
- **Eye** – глаз
- **Fr** – лобная доля
- **FrOc** – лобно - затылочная доля
- **FrPa** - лобно - теменная
- **GB** – базальные ганглии
- **Oc** – затылочная доля
- **Op** – оптический нерв
- **Pa** – теменная доля
- **PaTe** - теменно - височная
- **PC** – мосто - мозжечковый угол
- **Pi** – пинеальная область
- **SC** – хиазмально - селлярная
- **Te** - височная область
- **V1** – первый желудочек
- **V3** – третий желудочек
- **V4** – четвертый желудочек
- **VL** – боковые желудочки
- **н/д** - нет данных

Гистология

- **РМЖ** – рак молочной железы
- **РП** – рак почки
- **РШМ** – рак шейки матки
- **РЯ** – рак яичников

Меланома

- **МРЛ** – мелкоклеточный рак легкого
- **НМРЛ** – немелкоклеточный рак легкого
- **КРР** – колоректальный рак
- **Другой** – более редкие виды рака

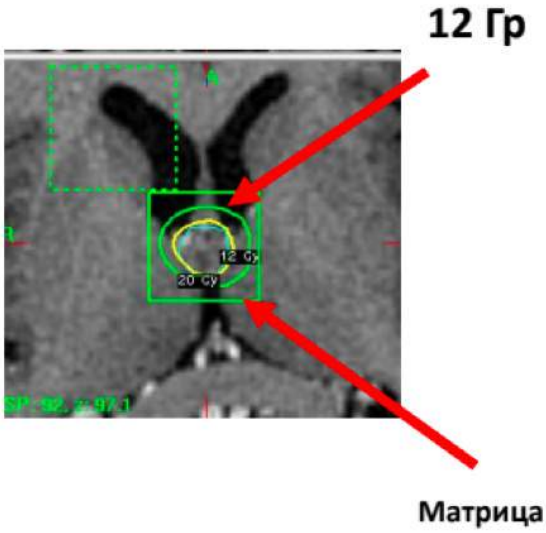
Предписанная доза/изодоза

D = E/m (Гр = Дж/кг)

От 10 Гр (крайне редко) до 24 Гр

Предписанная изодоза – отношение предписанной дозы к максимальной дозе в мишени (от 40 до 98 %)

Объем матрицы



12 Гр

10/11. NV12/10 GY – объем нормальных тканей, облученный дозой 12/10 Гр

NV12GY = MV12GY – V очага

NV10GY = MV10GY – V очага

12/13. Рецидив и время рецидива

Набор данных

- База пациентов (тренировочная выборка):
  http://nti-contest.ru/wp-content/uploads/bd/2.1.%20X_train.pdf
- Целевая переменная тренировочной выборки (где 0=отсутствует вероятность ремиссии, 1=присутствует вероятность ремиссии):
  http://nti-contest.ru/wp-content/uploads/bd/2.2%20y_train.pdf
- Бейзлайн:
  https://drive.google.com/open?id=1d3VqiRJlczU0XIJAU740V8G7Fx3WDrPk
- Пример корректно отформатированной выдачи под условия проверочного сервера:
  http://nti-contest.ru/wp-content/uploads/bd/2.4.%20example.pdf
- Тестовая выборка:
  http://nti-contest.ru/wp-content/uploads/bd/2.5.%20X_test.pdf

Код проверки и генерации правильного решения

```
import csv
import math
import random
import math
import requests
def csv_reader(file_obj):
        """
        Read a csv file
        """
        res = list()
        reader = csv.reader(file_obj)
        for row in reader:
            res.append(float(" ".join(row)))
          #  print(" ".join(row))
        return res
def f1(true,reply, label):
        if len(true) != len(reply):
            return 0
        totalTrue = 0
```

```
        predTrue = 0
        truepredTrue = 0
        for i in range(len(true)):
            if true[i]==label:
                totalTrue+=1
            if reply[i]==label:
                predTrue+=1
            if true[i]==label and reply[i]==label:
                truepredTrue+=1
        pr = truepredTrue/predTrue
        rc = truepredTrue/totalTrue
        return 2*pr*rc/(pr+rc)



#This is a sample Data Challenge
#Learn more: https://stepik.org/lesson/9172

def generate():

        return "Смотри тестовый датасет!"

def solve(dataset):

        url = "https://stepik.org/media/attachments/lesson/75231/y_test_212.csv"
        r = requests.get(url)
        clue = [str(int(i)) for i in r.content.split()]

        return " ".join(clue)



def check(reply, clue):
        reply = [float(i) for i in reply.split()]
        url = "https://stepik.org/media/attachments/lesson/75231/y_test_212.csv"
        r = requests.get(url)
        clue = [int(i) for i in r.content.split()]
        res = f1(list(clue), list(reply),1)
        if res<0:
            res=0
        balls = res*100
        resT = "Вы набрали " + str(balls) +" баллов"
        return res,resT
```

Пример решения задачи №2 командой школьников Олимпиады НТИ

файл .ipynb:
https://drive.google.com/open?id=1lCr0DA_ml-xQG8vU2KzS_tEdkzboBPmd

файл .html исключительно для визуального просмотра кода:
https://drive.google.com/open?id=1cfcyYURA1wiBTVbzWIyYG4sA0OgifNXq