Задачи к экзамену по курсу кафедры физики частиц и космологии "Методы машинного обучения для обработки данных" для магистров и аспирантов физического факультета МГУ, осень 2023 г.

1. *Реконструкция направления прихода нейтрино в IceCube*

   Цель задачи: обучить модель определять направление прихода нейтрино в IceCube используя большой массив симулированных данных, который коллаборация подготовила для kaggle: `https://www.kaggle.com/competitions/icecube-neutrinos-in-deep-ice/overview`

   Данные представляют собой импульсы, зарегистрированные в оптических модулях эксперимента (DOM) Каждое событие содержит нейтрино, но также может содержать космические мюоны. Число импульсов в событиях: от 2 до $\sim 10^5$.

   Замечание: для удобства работы, часть данных и пример их использования загружены на кластер ОТФ ИЯИ РАН. Доступ к кластеру предоставляется по запросу. Путь к данным: `cluster55.inr.ac.ru:/storage/vol5/timiryasov/icecube_data`

   (a) Обработать данные используя алгоритм LineFit `https://docs.icecube.aq/icetray/main/projects/linefit/index.html`

   (b) Натренировать простую двунаправленную рекуррентную сетт (GRU) глубиной 2-3 слоя. В качестве функции потерь можно использовать метрику, предложенную организаторами. Попробовать разные варианты аггрегации: max pooling используя все конечные состояния, линейное преобразование из последнего скрытого состояния. Добиться средней угловой ошибки меньше чем у LineFit.

   (c) Модифицировать предыдущий пункт используя трансформер.
   Замечание: для тренировки архитектуры типа "трансформер" может потребоваться современная видеокарта. Доступ к таким видеокартам на кластере ОТФ ИЯИ РАН может быть предоставлен по запросу.

2. *Оптимизация построения дерева решений методом Монте-Карло марковских цепей*

   Для оптимальной классификации требуется построение дерева решений с минимальной энтропией. Прямой подход к этой задаче подразумевает перебор всех вариантов разбиения дерева. Число вариантов разбиения для n параметров и дерева глубины k в простейшем случае без повтрений параметров, равно числу размещений A_n^k. Трудоемкость перебора экспоненциально растет с увеличением числа параметров и глубины дерева.

   (a) Пусть разбиение задается упорядоченным набором k параметров. Определить операцию элементарного шага, переводящего от одного разбиения к другому. Реализовать алгоритм оптимизации разбиения методом Монте-Карло марковских цепей (алгоритм Метрополиса).

   (b) Обобщить алгоритм на случай деревьев с возможным повтором параметров разбиения.

(c) Обобщить алгоритм на случай деревьев, в которых разбиения в левой и правой ветвях не обязательно совпадают.

3. *Оптимизация архитектуры нейронной сети методом Монте-Карло марковских цепей*

   Задача выбора оптимальной архитектуры нейронной сети в большинстве случаев решается эмпирически. Прямой подход к этой задаче подразумевает перебор всех вариантов структуры сети. Число различных вариантов конфигурации оказывается непрактично большим. Если ограничиться полносвязными сетями глубины не более K+1, содержащих не более M нейронов на каждом слое, кроме последнего, количество конфигураций составит порядка M^K. Трудоемкость перебора экспоненциально растет с глубиной сети. В данной задаче предлагается выполнить оптимизацию методом случайного блуждания по пространству конфигураций.

   (a) Пусть архитектура полносвязной нейронной сети задается упорядоченным набором k целых чисел $k <= K$, определяющих число нейронов в каждом слое, кроме последнего. Определить операцию элементарного шага, переводящего от одной архитектуры к другой.

   (b) Реализовать алгоритм оптимизации архитектуры методом Монте-Карло марковских цепей (алгоритм Метрополиса-Гастингса).

   (c) Применить алгоритм для условий задачи 7.1, 7.2 или одной из экзаменационных задач.

   (d) Сравить результат при тренировке с фиксированным числом эпох и с фиксированным временем обучения.

   (e) Предусмотреть возможность шага, изменяющего функцию активации одного из слоев.

   (f) Предусмотреть возможность шага, изменяющего глубину нейронной сети.

4. *Предсказание числа солнечных пятен*

   На сайте Королевской обсерватории Бельгии `http://www.sidc.be/silso/datafiles` размещены исторические данные о числе пятен на Солнце с 1 января 1818 года по настоящее время. Известно, что солнечная активность циклична с периодом около 11 лет. Кроме того, в солнечной активности существуют закономерности на бóльших и меньших временных масштабах. Таким образом предсказание числа пятен на следующий день может опираться на данные за 22 прошедших года, то есть на более, чем 8000 известных значений. Построить нейронную сеть, решающую данную задачу. Используя ансамблевые методы, то есть сравнивая предсказание большого числа моделей (например, с помощью слоя dropout), оценить интервал достоверности предсказания.

5. *Бустинг нейронных сетей*

   Бустинг – алгоритм повышения точности классификатора за счет построения цепочки классификаторов, в которой последующие классификаторы фокусируются на событиях, неправильно классифицированных предыдущими. Один из наиболее эффективных алгоритмов был предложен Фройндом и Шапиро

в 1997 году. Применить алгоритм бустинга (AdaBoost) к классификатору, построенному на базе многослойной нейронной сети. Сравнить точность классификации гамма-всплесков (задача 7.2) или гамма-источников (задача 12) с бустингом и без бустинга

6. *Состав космических лучей ультравысоких энергий*

   Эксперимент Telescope Array регистрирует широкие атмосферные ливни, вызванные космическими лучами ультравысоких энергий. Предполагается, что первичные частицы — протоны или ядра химических элементов.

   Для каждого события в результате реконструкции определяются 16 наблюдаемых параметров (описание физического смысла параметров см. в работе arXiv:1808.03680). По адресу `http://cluster.inr.ac.ru/ML/TASD/` размещены результаты Монте-Карло моделирования эксперимента для ШАЛ, вызванных первичными протонами *p.dat*, ядрами гелия *he.dat*, азота *n.dat* и железа *fe.dat*. Кроме того, размещены три неизвестных смеси *unknown1.dat*, *unknown2.dat*, *unknown3.dat*. Известно, что в первом неизвестном наборе присутствуют только протоны и железо.

   Определить состав первичных частиц в каждом из трех неизвестных наборов.

7. *Поиск фотонов ультравысоких энергий*

   В эксперименте Telescope Array проводится поиск гамма-квантов ультравысоких энергий $E > 10^{18}$ эВ.

   Для каждого события в результате реконструкции определяются 16 наблюдаемых параметров (описание физического смысла параметров см. в работе arXiv:1811.03920). По адресу `http://cluster.inr.ac.ru/ML/TASD/` размещены результаты Монте-Карло моделирования эксперимента для ШАЛ, вызванных первичными протонами *task5_p.dat* и гамма-квантами *task5_gamma.dat*. Кроме того, размещен неизвестный набор *task5_unknown.dat*, в котором присутствует несколько событий, вызванных гамма-квантами.

   Оценить число гамма-квантов в неизвестном наборе.

   Замечание: физический смысл параметров в этой задаче и задаче по химическому составу один и тот же. Однако из-за применения другой реконструкции и условий отбора, в этой задаче необходимо использовать отдельный протонный набор.

8. *Искусственный интеллект в поисках подписи Стивена Хокинга*

   Известно, что на карте Planck видны инициалы “SH”. Цель задачи в том, чтобы установить насколько часто подобные подписи появляются на картах реликтового излучения в результате случайных флуктуаций. Для этого, в первую очередь, необходимо научиться объективно определять есть или нет на карте такая надпись.

   - Подготовка тренировочного набора. Сгенерируем 1000 случайных карт реликтового излучения, используя спектр мощности Planck (C_l) вместе с произвольными фазами сферических гармоник. Простой пример python-скрипта для генерации карт реликтового излучения в пикселизации HEALPix размещен по адресу `http://cluster.inr.ac.ru/ML/CMB/`.

Будем считать, что на них надписи SH, как правило, нет и назовем их фоновыми картами. Сгенерируем другие 1000 карт и после впишем на них буквы SH таким алгоритмом, чтобы они походили на наблюдаемую карту Planck. Для упрощения, сначала предлагается писать SH в том же месте и того же размера. Эти карты назовем сигнальными картами.

- Используя одну из реализаций сверточной нейронной сети на сфере (arXiv:1810.12186 или arXiv:1902.04083), разработать алгоритм классификации, определяющий присутствие надписи SH. На выходе нейронной сети будет величина α, характеризующая насколько уверенно идентифицируется надпись 'SH'.
- Натренировать сеть на сигнальных и фоновых картах. Проанализировать этим методом большое количество случайных карт и вычислить долю случаев, в которых α выше, чем в карте Planck. Таким образом будет установлена вероятность случайного появления такой надписи.

9. *Предсказание погоды в день экзамена*

   Используя данные о погоде за 7 дней перед экзаменом (зачетом), исключая сам день экзамена, предсказать температуру, силу ветра и влажность в 15:00 в день экзамена/зачета. Для тренировки использовать архивные данные метеостанции `http://goo.gl/WgzQ4D`. Используя ансамблевые методы, то есть сравнивая предсказание большого числа моделей (например, с помощью слоя dropout), оценить интервал достоверности предсказания.

10. *Распознавание изображений с помощью сверточных сетей*

    Один из конкурсов на популярном портале Data Science kaggle, был посвящен задаче распознавания изображений галактик `https://www.kaggle.com/c/galaxy-zoo-the-galaxy-challenge`

    В качестве тренировочных данных предлагалось использовать более 60 тыс цветных снимков 424x424, для классификации которых привлекались волонтеры. Задачей алгоритма было предсказать, как люди классифицируют изображения. По условиям конкурсов kaggle победитель публикует описание алгоритма. Подробное описание лучшего решения этой задачи приведено на следующих ресурсах:

    - `http://benanne.github.io/2014/04/05/galaxy-zoo.html`
    - `https://github.com/benanne/kaggle-galaxies/blob/master/doc/documentation.pdf`

    Реализовать собственный алгоритм решения задачи и оценить его эффективность по приведенному критерию:

    `https://www.kaggle.com/c/galaxy-zoo-the-galaxy-challenge#evaluation`

11. *Автокодировщик параметров Planck*

    Построить автокодировщик: нейронную сеть, число выходных нейронов которой совпадает с числом входных и превышает число нейронов на одном из промежуточных слоев. Последний считается результатом автокодировщика.

Тренировка автокодировщика производится с функцией цены, имеющей минимум при попарном совпадении выходных сигналов с входными. Используя данные марковских цепочек Planck, описанные в условии задачи 5.1 построить автокодировщик для параметров вселенных Planck.

а) исследовать точность автокодировки в зависимости от числа нейронов в скрытом слое;

б) построить алгоритм регрессии для оценки функции правдоподобия Planck на базе нейронной сети, входной информацией для которой является результат автокодировщика;

в) сравнить точность при фиксированном времени обучения с результатом нейронной сети без использования автокодировщика.

12. *Определение типа источника гамма-излучения*

4FGL (LAT 10-year Source Catalog) — каталог источников гамма-излучения, зарегистрированных космической гамма-обсерваторией Fermi LAT. Среди таких источников есть внегалактические и галактические источники. К внегалактическим источникам относятся активные ядра галактик и их подкласс — блазары, среди галактических отдельно отметим пульсары. Часть гамма-пульсаров не излучают в гамма-диапазоне, поэтому их пульсации можно идентифицировать только по данным гамма-телескопов. Поиск подобных пульсаций — вычислительно трудоемкая задача (см., например, arXiv:1111.0523). Сложность связана с тем, что частота пульсаций изменяется во времени, а частота регистрации гамма-квантов от источника в миллионы раз ниже частоты пульсаций. Для оптимизации использования ресурсов предлагается провести предварительный отбор объектов на основании данных о спектре объекта 3FGL.

В файле *4fgl_full.dat*, размещенному по адресу `http://cluster.inr.ac.ru/ML/4FGL/`, содержится информация о 7195 бъектах 4FGL, 2577 из которых неидентифицированы. В первой колонке — имя объекта (общий для всех префикс “4FGL” опущен), во 2 и 3 колонках экваториальные координаты (прямое восхождение и склонение) в градусах. В колонках 4-9 — параметры эллипса, описывающего ошибку измерения положения источника, 10-46 параметры спектра и переменности (параметры переменности: 41-43). В колонке 47 содержится код типа источника или unk для неидентифицированных источников. Код типа источника принимает значения bll,BLL,bcu,BCU,fsrq,FSRQ для блазаров и значения psr, PSR, msp, MSP для пульсаров. Более полное описание кодов приведено в Таблице 7 в работе arXiv:1902.10045.

Задача посвящена построению классификатора на основе нейронной сети.

а) Выделить из каталога набор известных пульсаров и блазаров и случайно разбить его на две части: набор для тренировки и набор для тестирования.

б) Построить алгоритм классификации. Оценить точность классификации.

в) Выполнить классификацию неидентифицированных объектов 4FGL.

13. *Генерация искусственных гамма-всплесков*

Количество зарегистрированных гамма-всплесков ограничено. Многие задачи требуют использования Монте-Карло набора гамма-всплесков с большой статистикой. В рамках традиционных алгоритмов, создание такого набора подразумевает численное описание популяции гамма-всплесков. Задача посвящена альтернативному способу создания набора гамма-вспелсков в рамках методологии генеративно-состязательных сетей (GAN).

а) Разработать архитектуру генератора, на входе которого шум, а на выходе – вектор такой же размерности, как и информация о гамма-всплесках (в условиях задачи 4.2).

б) Разработать архитектуру дискриминатора, тренируемого формировать на выходе 1 для настоящих данных и 0 для сгенерированного набора.

в) Выполнить итерационную тренировку связки генератор-дискриминатор.

14. *Классификация переменных звезд по кривым блеска*

OGLE-III – каталог кривых блеска переменных звезд, в который входит более 400 000 переменных объектов. Каждый объект отнесен к одному из 8 типов, включающих в себя цефеиды, RR Лиры, звезды типа Дельты Щита и другие.

а) Используя данные каталога OGLE-III (интерактивный вариант `http://ogledb.astrouw.edu.pl/~ogle/CVS/`, файловый архив `http://www.astrouw.edu.pl/ogle/ogle3/OIII-CVS/`) построить сверточную нейронную сеть для классификации объектов по их кривым блеска. Разделить набор данных на тренировочные и тестовые, оценить точность работы нейронной сети.

б) Альтернативный подход к классификации переменных звезд состоит в выделении характеристик кривых блеска, таких как среднее значение амплитуды сигнала, период повторений, максимальная амплитуда и другие. Используя библиотеку FATS (arXiv:1506.00010, `https://github.com/isadoranun/FATS`) получить значения характеристик кривых блеска для тренировочного и тестового набора из данных OGLE-III. Построить полносвязную нейронную сеть, решающую задачу классификации переменных звезд на основе характеристик FATS.

в) Сравнить эффективность работы сверточной нейронной сети с эффективностью работы сети, основанной на признаках FATS.

15. *Поиск аномальных гамма-всплесков*

а) Разработать вариационный автокодировщик, преобразующий параметры гамма-всплесков (в условиях задачи 4.2) в вектор скрытых параметров меньшей размерности, подчиняющийся гауссовскому распределению.

б) Выделить 10 наиболее непохожих на другие гамма-всплесков, выбрав объекты с наибольшим отклонением от центра в скрытом пространстве. Насколько вероятны такие отклонения с точки зрения гауссовского распределения.

в) Какие физические параметры выделенных объектов аномальны?

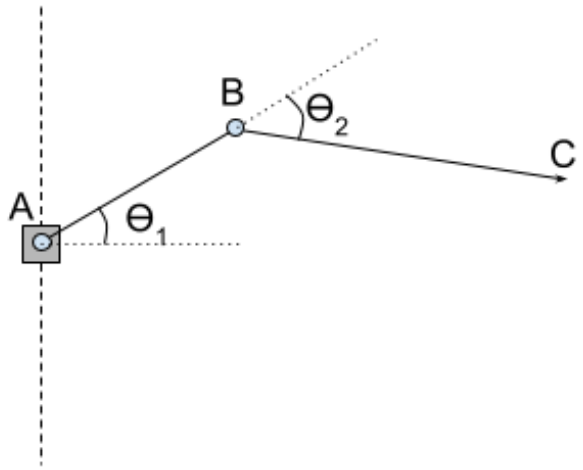

Рис. 1: Иллюстрация к обратной задаче с неединственным решением.

16. *Анализ обратных задач с неединственным решением* Часто обратные задачи имеют неединственное решение, и потому стандартные нейронные сети неприменимы для их решения. Рассмотрим в качестве примера следующую систему. Два жестких стержня соединены между собой шарниром и прикреплены через коробку к вертикальной рельсе (см. Рис. 1). Коробка может свободно перемещаться по рельсе, а стержни могут свободно вращаться вокруг точек A и B. Тогда заданному положению конечной точки C соответствует множество возможных конфигураций системы.

    Положим, что вертикальное положение коробки (точка A), углы поворотов θ_1 и θ_2 имеют нормальное распределение со средним 0 и дисперсиями 1 м, 30°, и 15°, соответственно. Такое поведение может обеспечиваться, например, случайной работой механизмов в точках соединений. Построить обратимую нейронную сеть (INN), которая по координатам точки C строит апостериорное распределение значений положения коробки и углов поворота. Длины стержней равны 1 м (AB) и 2 м (BC).

    Замечание: Необходимо самостоятельно сгенерировать данные для обучения нейронной сети. См. также статью arXiv:1808.04730.