Задачи к семинарам по курсу "Методы машинного обучения для обработки данных" кафедры физики частиц и космологии физического факультета МГУ.

Задание 2 по теме "Деревья решений."

2.1 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, [1]). 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 4FGL.

В файле `4fgl_full.dat` [2] содержится информация о 7195 объектах 3FGL, 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 для пульсаров. Более полное описание кодов приведено в Таблице 7 в работе [3].

Задача посвящена построению классификатора на основе деревьев решений.

а) Выделить из каталога набор известных пульсаров и блазаров и случайно разбить его на две равные части: набор для тренировки и набор для тестирования. Построить дерево решений для классификации, используя набор для тренировки. Оценить точность классификации с помощью набора для тестирования.

б) Реализовать бустинг оригинальным методом Р. Шапире [4]. Оценить точность классификации.

в) Реализовать метод AdaBoost [5]. Оценить точность классификации.

г) Выполнить классификацию неидентифицированных объектов 3FGL.

2.2 Определение красного смещения по фотометрии. SDSS (Sloan Digital Sky Survey) — один из крупнейших современных каталогов звезд и галактик, построенный с помощью 2.5-метрового широкоугольного телескопа в штате Нью-Мехико в США. Измерения проводятся в пяти спектральных полосах

фотометрической системы u'g'r'i'z со средними длинами волн от 355.1, 468.6, 616.5, 748.1 и 893.1 нм. Точное спектроскопическое измерение красного смещения (идентификация спектральных линий поглощения) доступно для нескольких миллионов из 500 миллионов объектов. Красное смещение остальных объектов предлагается оценить приближенно на основе данных SDSS.

В файле `sdss_train.dat` [6] содержится информация о 1999998 объектах с известным красным смещением. Данные о каждом объекте записаны в 37 колонках. В колонке 1 — идентификатор объекта, во 2 и 3 колонках экваториальные координаты (прямое восхождение и склонение) в градусах, в 4-13 колонках — звездные величины в 5 фотометрических полосах и их ошибки, 14-23 — радиусы по Петросяну и их ошибки в 5 фотометрических полосах, 24-33 — отношения полуосей эллипса и их ошибки в 5 фотометрических полосах, 34 — идентификатор в базе данных спектрометрии, 35 — класс объекта. В 36 колонке содержится красное смещение по спектроскопическим данным, а в 37 колонке его ошибка. Описание каждой колонки приведено в файле `sdss_train.colnames` [6], а описание каталога в работе [7]. Информация об 760002 объектах, красное смещение которых предстоит определить — в файле `sdss_test.dat` [6]. Формат последнего файла совпадает с описанным выше форматом файла `sdss_train.dat`, но не содержит последних двух колонок.

Задача посвящена определению красного смещения методом регрессии с помощью деревьев решений.

а) Разбить каталог объектов с известным красным смещением на две равные части: набор для тренировки и набор для тестирования. Построить дерево решений для регрессии, используя набор для тренировки. Оценить точность регрессии с помощью набора для тестирования.

б) Реализовать бустинг оригинальным методом Р. Шапире [4]. Оценить точность регрессии.

в) Реализовать метод AdaBoost [5]. Оценить точность регрессии.

г) Определить красное смещение для объектов SDSS, для которых спектроскопические данные не приведены. Предоставить на проверку преподавателю для сравнения с ответом, хранящимся в закрытой части сервера [8].

Список литературы

[1] H. J. Pletsch *et al.*, "Discovery of Nine Gamma-Ray Pulsars in Fermi-LAT Data Using a New Blind Search Method," Astrophys. J. **744**, 105 (2012) doi:10.1088/0004-637X/744/2/105 [arXiv:1111.0523 [astro-ph.HE]].

[2] http://cluster.inr.ac.ru/ML/4FGL/

[3] S. Abdollahi *et al.* [Fermi-LAT], "*Fermi* Large Area Telescope Fourth Source Catalog," Astrophys. J. Suppl. **247** (2020) no.1, 33 [arXiv:1902.10045 [astro-ph.HE]].

[4] R. E. Schapire, "The Strength of Weak Learnability," Machine Learning **5** 197 (1990). http://www.cs.princeton.edu/ schapire/papers/strengthofweak.pdf

[5] Y. Freund, R. E. Schapire, "A Short Introduction to Boosting", Journal of Japanese Society for Artificial Intelligence, 14(5):771 (1999). http://www.site.uottawa.ca/ stan/csi5387/boost-tut-ppr.pdf

[6] http://cluster.inr.ac.ru/ML/SDSS/

[7] C. Stoughton *et al.* [SDSS], Astron. J. **123** (2002), 485-548 doi:10.1086/324741

[8] http://cluster.inr.ac.ru/ML/SDSS/for_teacher/