

Задание 1 к части 2 “Деревья решений.”

1.1 Определение типа источника гамма-излучения. 3FGL (LAT 4-year Point 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 3FGL.

В файле `3fgl_full.dat` [2] содержится информация о 3034 объектах 3FGL, 1010 из которых неидентифицированы. В первой колонке — имя объекта (общий для всех префикс “3FGL” опущен), во 2 и 3 колонках экваториальные координаты (прямое восхождение и склонение) в градусах. В колонках 4-6 — параметры эллипса, описывающего ошибку измерения положения источника, 7-28 параметры спектра, 29 — индекс переменности. В колонке 30 содержится код типа источника или NULL для неидентифицированных источников. Код типа источника принимает значения `bll`, `BLL`, `bci`, `BCU`, `fsrq`, `FSRQ` для блазаров и значения `psr`, `PSR` для пульсаров. Более полное описание кодов приведено в Таблице 6 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.

1.2 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 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 примерно миллиона из 500 миллионов объектов. Красное смещение остальных объектов предлагается оценить приближенно на основе данных SDSS.

В файле `sdss_known_z.dat` [6] содержится информация о 1244498 объектах с известным красным смещением. Данные о каждом объекте записаны в 29 колонках. В колонке 1 — идентификатор объекта, во 2 и 3 колонках экваториальные координаты (прямое восхождение и склонение) в градусах, в 4-9 колонках — данные фотометрии в 5 полосах, 9-13 — радиусы, 14-23 — цвета, 24-28 — эллиптичности. В 29 колонке содержится красное смещение по спектроскопическим данным. Информация об объектах, красное смещение которых предстоит определить — в файле `sdss_unknown_z.dat` [6].

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 отсутствуют.

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] <ftp://cluster.inr.ac.ru/pub/ML/3FGL/>
- [3] F. Acero *et al.* [Fermi-LAT Collaboration], “Fermi Large Area Telescope Third Source Catalog,” *Astrophys. J. Suppl.* **218**, no. 2, 23 (2015) doi:10.1088/0067-0049/218/2/23 [arXiv:1501.02003 [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] <ftp://cluster.inr.ac.ru/pub/ML/SDSS/>