ЛР 3-4 (продвинутый вариант)
Построение фреймворка для сравнительной оценки методов машинного обучения в задачах распознавания

Для бинарной классификации могут использоваться различные методы машинного обучения, перечень и достаточно детальное описание которых можно найти, например, в курсе «Машинное обучение» (курс лекций, К.В.Воронцов):

http://www.machinelearning.ru/wiki/index.php?title=%D0%9C%D0%B0%D1%88%D0%B8%D0%BD%D0%BD%D0%BE%D0%B5_%D0%BE%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D0%B5_(%D0%BA%D1%83%D1%80%D1%81_%D0%BB%D0%B5%D0%BA%D1%86%D0%B8%D0%B9%2C_%D0%9A.%D0%92.%D0%92%D0%BE%D1%80%D0%BE%D0%BD%D1%86%D0%BE%D0%B2)

Основным методом оценки качества методов бинарной классификации является **ROC-кривая** – график, который отображает соотношение между долей объектов от общего количества носителей признака, верно классифицированных как несущих признак (англ. true positive rate, TPR, называемой чувствительностью алгоритма классификации), и долей объектов от общего количества объектов, не несущих признака, ошибочно классифицированных как несущих признак (англ. false positive rate, FPR, величина 1-FPR называется специфичностью алгоритма классификации) при варьировании порога решающего правила.

Задача лабораторной работы – выявить методы машинного обучения, наиболее подходящие для классификации рака легкого (Lung Cancer).

В качестве экспериментальных данных можно использовать любой открытый датасет по теме.

Ниже приводится перечень некоторых ссылок на датасеты по теме "Lung Cancer Dataset". Надо поискать по ссылкам именно табличные данные, которые там прикреплены в разных форматах - в архивах, pdf, xls и т.д.

http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi
5
https://icbp.nci.nih.gov/resources/list-of-datasets-and-databases

http://www.rcpa.edu.au/Library/Practising-pathology/ICCR/Cancer-Datasets

http://www.isdscotland.org/Health-Topics/Cancer/Cancer-Audit/
https://biometry.nci.nih.gov/cdas/nlst/datasets/comprehensive/

http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/dataset.cgi?study_id=phs000753.v1.p1&phv=201872&phd=&pha=&pht=3876&phvf=&phdf=&phaf=&phtf=&dssp=1&consent=&temp=1

http://www.camda.duke.edu/camda03/datasets/

Задание на ЛР:

выбрать датасет по теме Lung Cancer

выбрать не менее 5 методов машинного обучения (список не должен не полностью совпадать с использованными в референтной статье)

найти в открытых источниках программные средства для реализации этих методов;

для каждого метода по выбранному датасету построить ROC-кривую;

провести сравнение полученных ROC-кривых; результат представить графической форме (по образцу Fig.1 референтной статьи)

провести валидацию полученных результатов (по формуле 10 референтной статьи).