«Практические задачи анализа данных»

3 курс, майнор «Интеллектуальный анализ данных»

Групповой или индивидуальный проект

Авторы: Д.И. Игнатов, Е.Л. Черняк

Срок сдачи итогового отчета: неделя перед зачетной, 4 модуль 2017

Постановка задачи и техническое задание: до 13.04.2017

Постановка задачи

Под групповым проектом подразумевается коллективное выполнение задания, связанного с применением методов разработки данных и машинного обучения. Перед тем как приступить к выполнению проекта необходимо:

1. Сформировать группы, состоящие из не более трех человек. Можно работать над индивидуальным проектом.
2. Найти (выбрать) набор данных для анализа.
3. Сформулировать постановку задачи, описать данные (число объектов, признаков, число классов, можно привести и другие статистики) и составить план решения в виде технического задания по проекту. Объем ТЗ – примерно 1-2 страницы текста.

Необходимо проинформировать проверяющих о выборе задачи до дедлайна (13.04.2017), а затем можно перейти к выполнению проекта. Подходящие наборы данных, например, можно найти на сайтах:

UC Irvine Machine Learning Repository
http://www.kaggle.com/competitions
http://www.openml.org/
http://www-stat.stanford.edu/~tibs/ElemStatLearn/
http://lib.stat.cmu.edu/datasets
http://www.statsci.org/datasets.html
http://www.amstat.org/publications/jse/jse_data_archive.htm
http://www.physionet.org/physiobank/database
http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/DataSets.

Приветствуется работа с текстовыми (лингвистическими) данными, например:

http://romip.ru/ru/collections/index.html
http://universaldependencies.org/
http://trec.nist.gov/
http://pan.webis.de/data.html
http://www.clef-initiative.eu/dataset/test-collection
http://www.dialog-21.ru/evaluation/
http://alt.qcri.org/semeval2017/index.php?id=tasks

Примерное содержание отчета по проекту следующее:

1. Формулировка задачи
2. Описание данных
3. Обоснование выбора методов
4. Постановка / результаты экспериментов
5. Сравнение методов
6. Выводы

Q.: *Есть ли ограничения снизу/сверху по размеру данных, текста отчета и набору применяемых алгоритмов машинного обучения?*

A.: *Ограничения снизу есть. Данные не должны быть слишком маленькими, например, размером не менее 50 объектов × 10 признаков (признаков может быть меньше, при достаточно большом числе примеров). Текст должен быть похож на подробный и понятный преподавателю или сокурснику рассказ о том, что Вы сделали, с таблицами, графиками, скриншотами и прочими вспомогательными иллюстрациями. В принципе, чем больше методов применено, тем лучше. Сравнение и интерпретация результатов обязательны. Если Вы, например, решили применить кластеризацию, то необходимо сравнить результаты работы несколько методов. Применение своего оригинального метода весьма приветствуется.*
Необходимо продемонстрировать всю цепочку работы с данными, включающую в себя их сбор, предобработку (шкалирование, удаление выбросов, отбор или извлечение признаков и т.п.), применение методов, сравнение, анализ ошибок и интерпретацию результатов.

Q.: *Какие методы анализа данных и машинного обучения стоит использовать?*

A.: *Можно использовать любые методы машинного обучения и анализа данных, в том числе, те, которые обсуждались на предыдущих курсах. Методы стоит выбирать в зависимости от сформулированной задачи: классификация, регрессия, кластерный анализ, ранжирование, рекомендательные системы, анализ последовательностей и др.*

Защита проектов состоится на зачетной неделе. Максимальная возможная оценка по проекту – 10. В ходе защиты могут быть заданы вопросы, касающиеся любой темы, изученной в курса.

Приветствуется использование следующих библиотек:

- Orange `http://orange.biolab.si/`;
- Weka `www.cs.waikato.ac.nz/ml/weka`;
- Scikit-learn `http://scikit-learn.org/stable/`

- Matlab или R.

Приветствуется работа с применением частых множеств признаков, см., например, пакет SPMF `http://www.philippe-fournier-viger.com/spmf/`.

Техническое задание и отчет по проекту следует отправлять по адресу <iad.hse@yandex.ru>. Тема письма должна быть оформлена следующим образом:
[ИАД-X]-[ТЗ]-Фамилия(-ии),
[ИАД-X]-[Отчет]-Фамилия(-ии), где X – номер вашей группы.

Если средняя оценка с учетом проверки ДЗ, ТЗ и отчета по проекту 8 баллов и выше, то оценка по данной части курса может быть выставлена автоматически как среднее арифметическое (округление арифметическое или в пользу студента на усмотрение проверящего).