МГУ им.Ломоносова
Факультет Вычислительной Математики и Кибернетики
Кафедра Математических Методов Прогнозирования

Отчёт по проделанной работе «Topical Classification of Biomedical Research Papers»

Шаймарданов Ильдар
317 группа

Апрель 2012 года

1) Постановка задачи.

Авторы задачи собрали коллекцию из 20000 научных статей на медицинскую тему. Каждому документу поставлено в соответствие некоторое подмножество терминов из словаря MeSH. Это подмножество ключевых фраз взятое с весами, обозначающими их выраженность в документе, описывает документ. Каждый документ отнесен к нескольким темам(числом 83). Эти темы надо научиться предсказывать по обучающей выборке, состоящей из 10000 статей. Контрольная выборка состоит тоже из 10000 статей. Оценка обобщающей способности ведётся с помощью F-меры.

Формат данных следующий: один файл для матрицы объект-признак и один файл для матрицы объект-классы.

Файл для матрицы объект-признак:

Каждая строка отвечает одному документу. В строке содержится 25640 чисел от 0 до 1000 разделенных пробелом(25640 — количество терминов MeSH)

Всего 10000 строк(для обучающей выборки)

Файл для матрицы объект-классы:

Каждый строка отвечает одному документу. В строке содержится 83 бинарных чисел(0 или 1), разделённых пробелом

Всего 10000 строк(для обучающей выборки)

Кроме того дан отдельный файл с тестовыми данными(контрольная выборка), содержащим объекты, темы которых нужно предсказать, и по которым будет вычисляться оценка качества.

2) Решение задачи.

Каждый документ может быть отнесён к нескольким темам, то есть мы имеем задачу мягкой кластеризации. Поэтому имеет смысл решать задачу классификации по каждой теме отдельно, то есть приняв гипотезу о независимости встречаемости тем (что не факт) решить 83 задачи классификации.

В качестве инструмента классификации была выбрана библиотека liblinear, в прошлом хорошо себя показавшая в задаче распознавания пешеходов курса машинной графики. Библиотека умеет эффективно работать с разреженными матрицами больших размеров (наша мтрица — 10000x25640), а также имеет простой интерфейс и проста в установке.

Классификаторы библиотеки имеют следующие параметры:

1) Параметр регуляризации C
2) Параметр E — эпсилон, отвечает за скорость сходимости
3) B — смещение(bias)
4) S — тип классификатора(характеризуется используемыми метриками

Подбор параметров производился кросс-валидацией — обучение на первой половине, оценка — на второй половине обучающей выборки.

Предварительно были удалены признаки, ни разу не встречающиеся в обучающей выборке.

Таким образом был выбран классификатор
L1-regularized logistic regression

Предварительный результат на сайте соревнования — 0.361

3) Выводы.

Чтобы я сделал если бы у меня было больше времени :

1) Начал бы решать задачу сразу и продолжать работать над ней регулярно.
2) Читать больше статей
3) Тратить больше времени на эксперименты

Советы новичкам:

1) Все предыдущие пункты, касающиеся нехватки времени
2) Не стесняться спрашивать советы у преподавателей
3) Больше практиковаться

4) Итоги командной работы

Из своей группы я хочу поблагодарить Петра Ромова за создание адекватной страницы обсуждения и удачный выбор библиотеки линейной классификации, а также за его разъяснения сути самой задачи и некоторый код. Мне также пригодился код Дмитрия Кондрашкина по подсчету f-меры, которые я мог написать сам, но, тем не менее, воспользовался готовым вариантом.

Мой скромный вклад в обсуждение задачи состоит в процедуре записи ответов к тестовой выборке в файл, готовый к отправке на сайт соревнования. Код содержал ошибку, которую нашла Мария Любимцева, за что ей большое спасибо. Также я опубликовал небольшой совет по установке liblinear, касающийся проблемы, возникающей у пользователей MSVS.