Практическое задание: применение линейных классификаторов к выявлению фамилий в текстах

крайний срок сдачи - 22.04.2013, 23.59

В данном практическом задании рассматривается задача бинарной классификации слов. Первый класс содержит фамилии людей, второй класс содержит «обычные» имена существительные, употребляемые не в качестве фамилий. Одно и тоже слово может принадлежать обоим классам. Вам предоставляются два файла в кодировке utf-8:

- train.txt — содержит ровно 101408 строк, в каждой из которых записано слово и тип этого слова, разделенные запятой. Тип слова — число, где 1 означает фамилию, а 0 означает обычное существительное.
- test.txt — содержит ровно 188920 строк. В каждой строке записано слово, тип которого надо определить.

Требуется:

1. придумать и реализовать информативные признаки.
2. обучить на данных train.txt линейный классификатор.
3. построить прогноз данных test.txt и записать полученный результат в файл result.txt. Файл result.txt должен содержать ровно 188920 строк, хранящих число от нуля до единицы — оценку вероятности того, что данное слово является фамилией.
4. оформить отчет, в котором описать использованные вами признаки, построенный линейный классификатор, способ настройки структурных параметров.

Оценка которую вы получите за задание складывается из двух состовляющих: 25% — отчет, 75% — качество, посчитанное по test.txt с помощью метрики AUC.

Ограничений на язык программирования/библиотеки/пакеты нет.

линейные классификаторы в R:
Для обучения логистической регрессии используйте функцию glm() с параметром

family=binomial(logit):
> model <- glm(formula=$Y \sim .$, data=training_dataset, family=binomial(logit))
Для получения оценок вероятностей при прогнозировании указывайте параметр type='response':
> predicted <- predict(model, newdata=test_dataset, type='response')

Используйте функцию svm() из пакета e1071 для метода опорных векторов.
> library(e1071)
> model <- svm(formula=$Y \sim .$, data=training_dataset, probability=TRUE)
> predicted_matrix <- predict(model, newdata=test_dataset, probability=TRUE)
> predicted <- attr(predicted_matrix, 'probabilities')[[1]]