Разбор решения

CMC MSU, Machine Learning (Fall 2015/2016): Competition 1

Николаев Владимир Владимирович

ВМК МГУ 317 группа

27 ноября 2015 г.

Постановка задачи

Имеются данные о вопросах и их авторах с сайта `stackoverflow.com`.
Для каждого вопроса заданы:

- PostId – ID вопроса (испорчен в тесте)
- PostCreationDate – Дата и время, когда тема была создана
- OwnerUserId – ID пользователя, открывшего тему
- OwnerCreationDate – Дата и время, когда был создан аккаунт автора темы
- ReputationAtPostCreation – Репутация автора темы на момент открытия темы
- OwnerUndeletedAnswerCountAtPostTime – Количество ответов на другие темы, данных автором на момент открытия текущей темы
- Title – Заголовок темы
- BodyMarkdown – Текст темы
- Tag1, . . . , Tag5 – Тэги темы

Обычные признаки

- ReputationAtPostCreation
- OwnerUndeletedAnswerCountAtPostTime
- PostCreationDate/OwnerCreationDate
- Разность PostCreationDate и OwnerCreationDate
- Счетчики по OwnerUserId

Признаки, извлеченные из текста

- Число строк кода/текста
- Длина заголовка/текста
- Число слов в заголовке/тексте
- Число знаков "?"
- Число ссылок
- Число вхождений "last_modified"
- Число предложений

Текст

- TF*IDF на Title+BodyMarkdown+3*Tags
- TF*IDF на Title
- В обоих случаях сделано по словам с 1,2,3 граммами

Архитектура

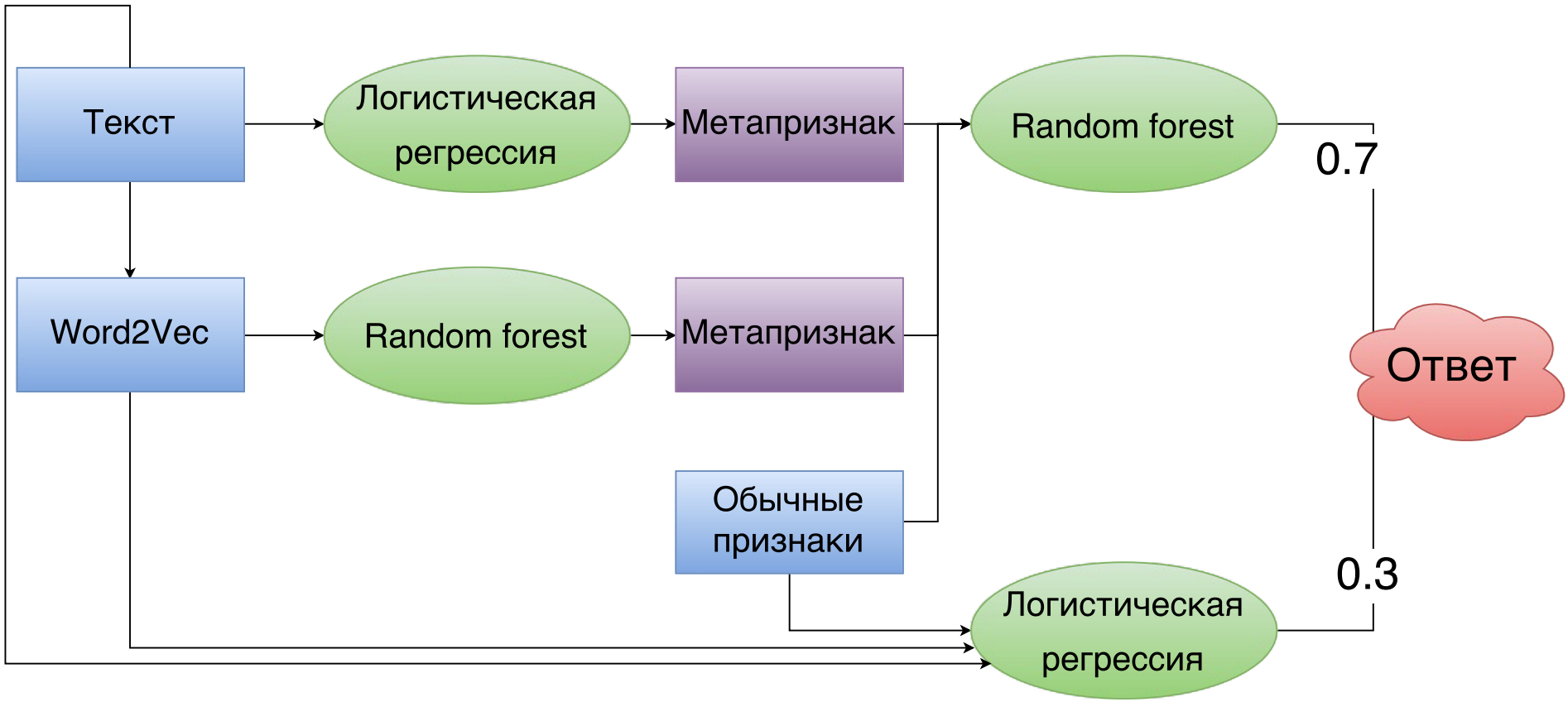

Рис. 1 : Архитектура классификатора