Задача классификации пола на данных социальной сети «ВКонтакте»

Василий Советов, Максим Новиков,
Михаил Кольцов, Павел Швец

Факультет ВМК МГУ им. М.В.Ломоносова

Техносфера@mail.ru

12 апреля 2014 г.

Описание задачи

- Данные 1800 человек

Описание задачи

- Данные 1800 человек
- Сильно смещенная выборка: 1300 мужчин против 500 женщин

Описание задачи

- Данные 1800 человек
- Сильно смещенная выборка: 1300 мужчин против 500 женщин
- Для оценки качества классификаторов использовалась кросс-валидация по 3 частям с усреднением по метрике

$$\left| \frac{TP}{TP + FN} + \frac{TN}{TN + FP} - 1 \right|$$

Процесс обучения

1. Кросс-валидация каждого алгоритма в отдельности

1. Кросс-валидация каждого алгоритма в отдельности
2. Нормировка результатов α_i по метрике

$$\sum_i \alpha_i = 1$$

Процесс обучения

1. Кросс-валидация каждого алгоритма в отдельности
2. Нормировка результатов α_i по метрике

$$\sum_i \alpha_i = 1$$

3. Обучение каждого из классификаторов отдельно

Процесс тестирования

1. Предсказание классов каждым алгоритмов в отдельности

Процесс тестирования

1. Предсказание классов каждым алгоритмов в отдельности
2. Вычисление результирующей метки класса

$$\left[\sum_i \alpha_i \cdot \text{out}(\text{clf}_i)\right]$$

Градиентный спуск

Используемый признак: подписки пользователей

Градиентный спуск

Используемый признак: подписки пользователей

- Открыты у всех пользователей

Используемый признак: подписки пользователей

- Открыты у всех пользователей
- Есть популярные тематические паблики, в которых преобладает определенный пол

Градиентный спуск

Используемый признак: подписки пользователей

- Открыты у всех пользователей
- Есть популярные тематические паблики, в которых преобладает определенный пол
- Научный интерес - насколько этот признак может линейно разделить пользователей по полам

$$Q = \sum_{i=1}^{l} L\left(\langle w, x_i \rangle y_i\right)$$

$$w^{(t+1)} = w^{(t+1)} - \eta \bigtriangledown Q\left(w^{(t)}\right)$$

$$X = \mathbb{R}^{n+1},\ y = \{-1, +1\}$$

Правило Хэбба

$$L(a, y) = (-\langle w, x \rangle \cdot y)_+$$

$$a(x, w) = sign\,\langle w, x \rangle$$

$$Q = \sum_{i=1}^{l} L\left(\langle w, x_i\rangle\, y_i\right)$$

$$w^{(t+1)} = w^{(t+1)} - \eta \bigtriangledown Q\left(w^{(t)}\right)$$

$$X = \mathbb{R}^{n+1},\ y = \{-1, +1\}$$

Правило Хэбба

$$L(a, y) = (-\langle w, x\rangle \cdot y)_+$$

$$a(x, w) = sign\,\langle w, x\rangle$$

Градиентный шаг:

$$\langle w, x_i\rangle \cdot y_i < 0 \Rightarrow w^{(t+1)} = w^{(t)} + \eta x_i y_i$$

Реализация:

- Остановка роста дерева по: глубине, количеству элементов выборки в вершине, достигаемому изменению impurity
- Отсутствие признака - отдельная категория
- Деление происходит всегда на 2 потомка
- Варианты impurity: misclassification, gini, informational entropy

Реализация:

- Остановка роста дерева по: глубине, количеству элементов выборки в вершине, достигаемому изменению impurity
- Отсутствие признака - отдельная категория
- Деление происходит всегда на 2 потомка
- Варианты impurity: misclassification, gini, informational entropy

Параметры и признаки, не оправдавшие себя:

- Отсечение по глубине или изменению impurity
- Возможность использовать категориальные признаки

Деревья решений

- Количество друзей
- Количество фотографий
- Количество групп
- Средняя частота выкладывания контента на стену

Деревья решений

- Количество друзей
- Количество фотографий
- Количество групп
- Средняя частота выкладывания контента на стену

Оптимальные параметры:

Деревья решений

- Количество друзей
- Количество фотографий
- Количество групп
- Средняя частота выкладывания контента на стену

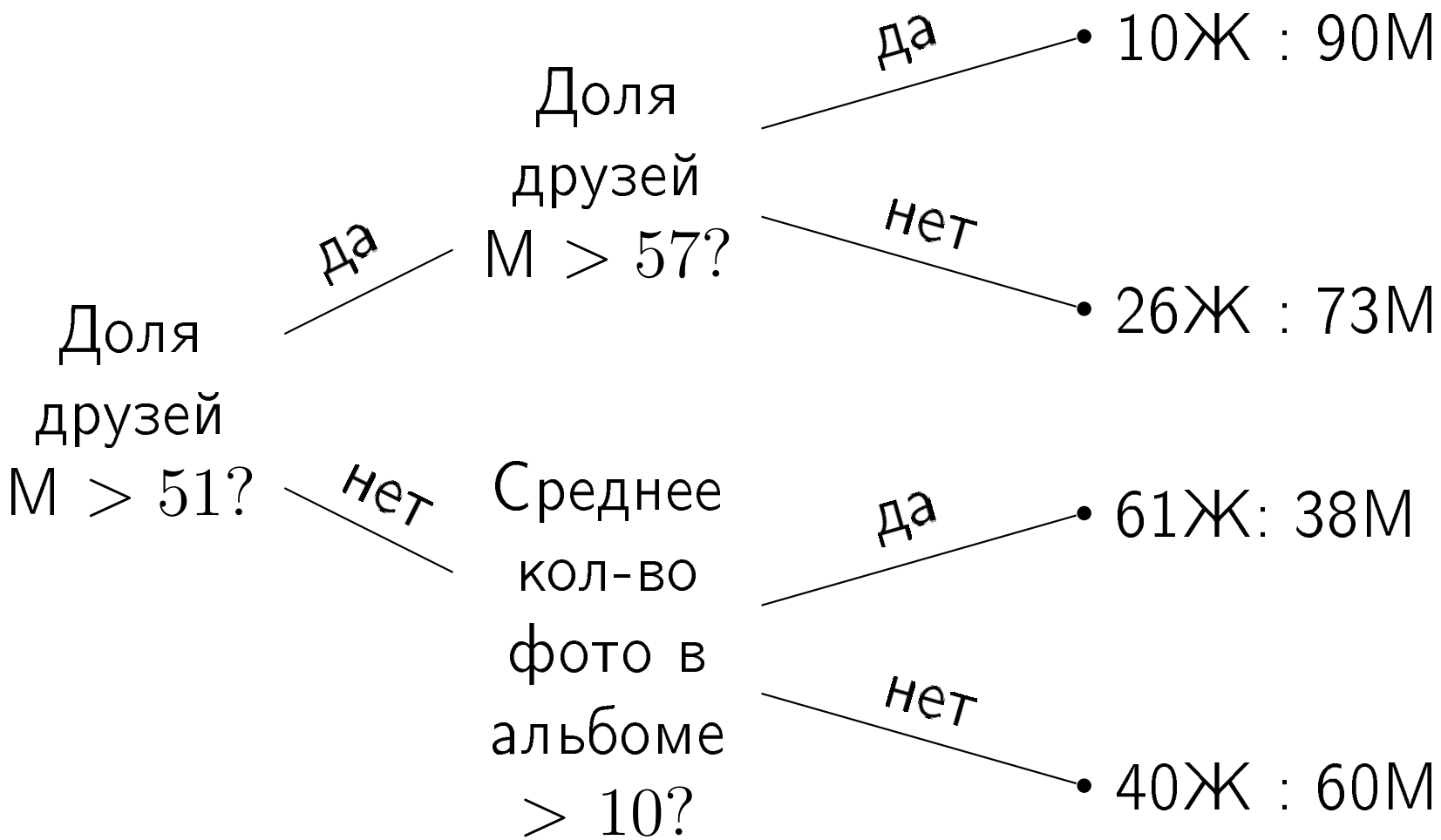
Оптимальные параметры:

- impurity - informational entropy $= \sum_i p_i \log p_i$
- Максимальное количество элементов в вершине - 1% от выборки

Используемые признаки:

- Процент друзей мужского пола
- Среднее количество фотографий в альбоме, количество альбомов
- количество друзей / количество "интересных страниц"
- количество аудиозаписей / количество разных исполнителей
- Среднее количество лайков, количество комментов и репостов на 20 последних сообщениях со стены

Деревья решений



Наивный Байес

Используемый признак: исполнители из аудиозаписей пользлователя

Используемый признак: исполнители из аудиозаписей пользлователя

- Открыты у большого количества пользователей

Используемый признак: исполнители из аудиозаписей пользлователя

- Открыты у большого количества пользователей
- Научный интерес - проверить, насколько точно можно предсказать пол исключительно по слушаемым исполнителям

Нормализация данных:

Нормализация данных:

- Stopwords: 'Offspring' или 'The Offspring'

Нормализация данных:

- Stopwords: 'Offspring' или 'The Offspring'
- Понижение регистра, удаление пробелов, пунктуации ...

Можно было бы еще...

Можно было бы еще...НО:

Можно было бы еще...НО:

- Аббревиатуры:

Можно было бы еще...НО:

- Аббревиатуры: 30STM, БГ...
  а если аббревиатуры пересекаются?

Можно было бы еще...НО:

- Аббревиатуры: 30STM, БГ...
  а если аббревиатуры пересекаются?
- Совместные записи:

Можно было бы еще...НО:

- Аббревиатуры: 30STM, БГ...
  а если аббревиатуры пересекаются?
- Совместные записи:
  Drake feat. Kanye West, Lil Wayne, & Eminem (Untz Remix)

Можно было бы еще...НО:

- Аббревиатуры: 30STM, БГ...
  а если аббревиатуры пересекаются?
- Совместные записи:
  Drake feat. Kanye West, Lil Wayne, & Eminem (Untz Remix)
- Опечатки

Можно было бы еще...НО:

- Аббревиатуры: 30STM, БГ...
  а если аббревиатуры пересекаются?
- Совместные записи:
  Drake feat. Kanye West, Lil Wayne, & Eminem (Untz Remix)
- Опечатки

Требуется что-то большее чем автоматическая обработка

Используемые признаки:

- Распределение пола друзей
- Количество аудиозаписей
- Количество подписок
- Количество друзей

Метод опорных векторов

Используемые признаки:

- Распределение пола друзей
- Количество аудиозаписей
- Количество подписок
- Количество друзей

Используется решение недвойственной задачи

Метод опорных векторов

Проблемы:

- Долгое время работы нелинейных ядер

Проблемы:

- Долгое время работы нелинейных ядер
- Разная гиперплоскость для $libLINEAR$ и линейного ядра $libSVM$

Проблемы:

- Долгое время работы нелинейных ядер
- Разная гиперплоскость для $libLINEAR$ и линейного ядра $libSVM$
- Работа только с количественными признаками

Проблемы:

- Долгое время работы нелинейных ядер
- Разная гиперплоскость для $libLINEAR$ и линейного ядра $libSVM$
- Работа только с количественными признаками

Метод не попал в итоговый классификатор

Трудности

- Наличие ограничения на количество запросов к API

Трудности

- Наличие ограничения на количество запросов к API
- Итеративность скачивания данных

Трудности

- Наличие ограничения на количество запросов к API
- Итеративность скачивания данных
- Унификация форматов хранения файлов и всех интерфейсов функций

Трудности

- Наличие ограничения на количество запросов к API
- Итеративность скачивания данных
- Унификация форматов хранения файлов и всех интерфейсов функций
- Неочевидный выбор эвристик

Трудности

- Наличие ограничения на количество запросов к API
- Итеративность скачивания данных
- Унификация форматов хранения файлов и всех интерфейсов функций
- Неочевидный выбор эвристик
- Корреляция различных классификаторов

Итоговый выбор

- Наивный Байесовский классификатор показывает лучший результат

Итоговый выбор

- Наивный Байесовский классификатор показывает лучший результат
- «Бустинг» слабых классификаторов

Итоговый выбор

- Наивный Байесовский классификатор показывает лучший результат
- «Бустинг» слабых классификаторов
- Итоговая метрика — f_score

Результаты

Метод	*Точность*	*Полнота*	*F-мера*
Gradient Descent	0.85	0.76	0.72
Decision Trees	0.80	0.80	0.80
Naive Bayes	0.85	0.85	0.85
SVM	0.78	0.97	0.86
Mix	0.85	0.86	0.86