Модуль 1
Задачи классификации

Группа №3

Безденежных Константин
Куликов Владислав
Мельников Андрей
Николаев Евгений
Чиркова Надежда

Определение пола пользователей социальной сети ВКонтакте

Обучающая выборка: 398 человек

- 156 женского пола (≈39%)
- 242 мужского пола (≈61%)

Метод оценки качества: метрика accuracy

cross_val_score (from sklearn.cross_validation)

Байесовский классификатор

Текстовая информация:

- Записи на стене (+комментарии)
- Названия аудиозаписей
- Исполнители аудиозаписей
- Названия групп

MultinomialNB

Обучение

Вход: тексты $\mathbf{T_i}$, классы $\mathbf{c_k}$

- Разбиваем $\mathbf{T_i}$ на токены $\mathbf{t_{ij}}$
- $\forall$ класса $\mathbf{c_k}$ вычисляем $\mathbf{p(c_k)}$
- $\forall(\mathbf{t_{ij}}, \mathbf{c_k})$ вычисляем $\mathbf{p(t_{ij}|c_k)}$

Выход: $\mathbf{p(c_k)}$, $\mathbf{p(t_{ij}|c_k)}$

MultinomialNB

Применение

Вход: текст **T**, $\mathbf{p(c_k)}$, $\mathbf{p(t_{ij}|c_k)}$

- Разбиваем **T** на токены $\mathbf{t_i}$
- $\forall$ класса $\mathbf{c_k}$**:**
  - $\mathbf{p(c_k|T) = p(c_k)}$
  - $\forall$ токена $\mathbf{t_i}$: $\mathbf{p(c_k|T) *= p(t_i|c_k)}$

Выход: $\mathbf{p(c_k|T)}$

Байесовский классификатор

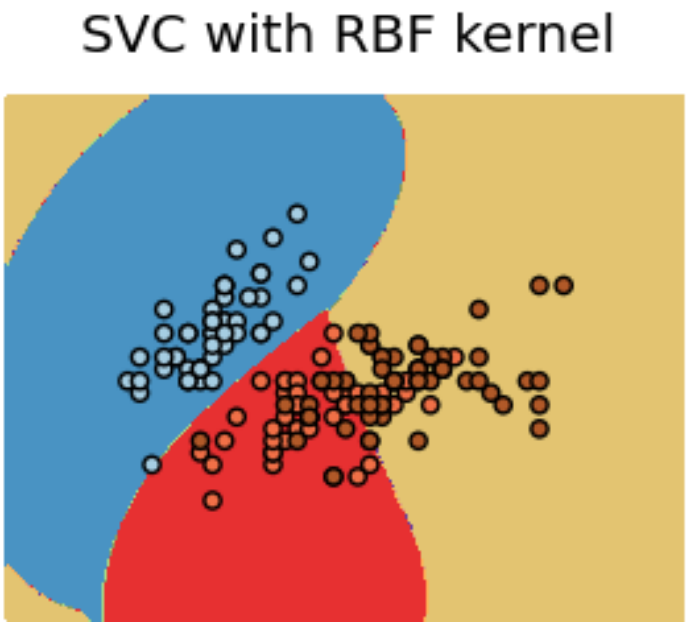
Обработка данных:

- ~~Стоп-слова~~
- ~~Стемминг~~
- ~~Лемматизация~~
- n-граммы

Метод опорных векторов (SVM)

Использована реализация sklearn.svm

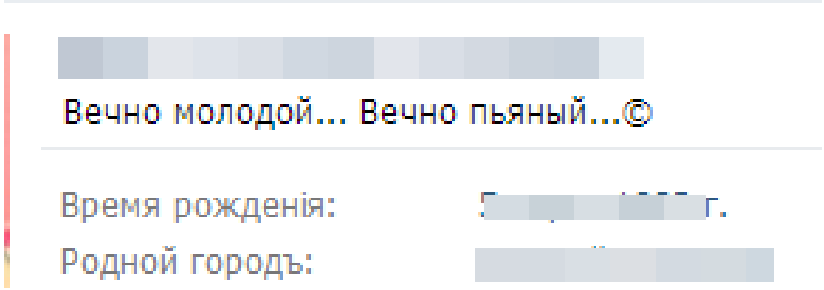
- Ядро: RBF
- Gamma: 0.7



Метод опорных векторов (SVM)

Признаки для классификации:

- Число друзей
- Число аудиозаписей
- Число групп



Дерево решений

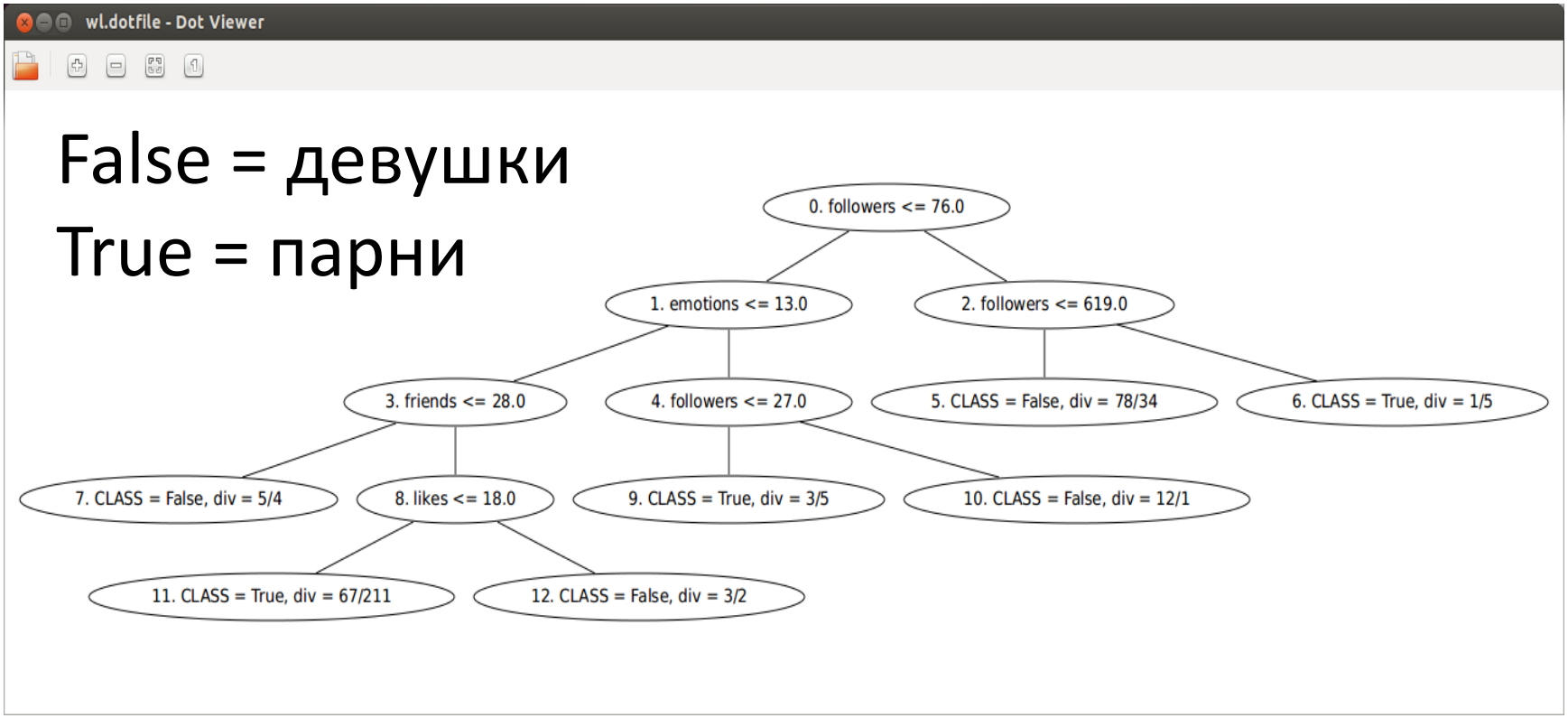
Признаки для классификации:

friends, followers, emotions, likes

Параметры обучения:

- min_delta_imp = 0.02
- min_leaf_size = 5
- Метод подсчета impurity: misclassification

Дерево решений



Accuracy = [0.701 0.709 0.686 0.698 0.663]

Дерево решений

Проблемы:

- Много отсутствующих признаков
- Дисбаланс классов
- Переобучение и отсутствие глобальной оптимальности

Итоги

- MultinomialNB: 0.61
- SVM: 0.70
- DT: 0.69