Отчет по Competetion 3

Yandex SHAD & MIPT FIVT, ML, Spring 2015 [Kaggle.com]

Вихрева Мария, ВМК МГУ

13 мая 2015

Формулировка задачи

Yandex SHAD & MIPT FIVT, ML, Spring 2015 [Kaggle.com]

Дано: "важности" входящих в статью терминов

Задача: предсказать категории статьи

Функционал качества: $\mathrm{F1} = 2\frac{\mathrm{pr}}{\mathrm{p+r}}$, где $\mathrm{p} = \frac{\mathrm{tp}}{\mathrm{tp+fp}}$, $\mathrm{r} = \frac{\mathrm{tp}}{\mathrm{tp+fn}}$

- 25640 безымянных признаков
- train – 10000 объектов, test – 10000 объектов
- матрица признаков – разреженная
- 83 категории

Baseline-решения

kNN (F1=0.30)

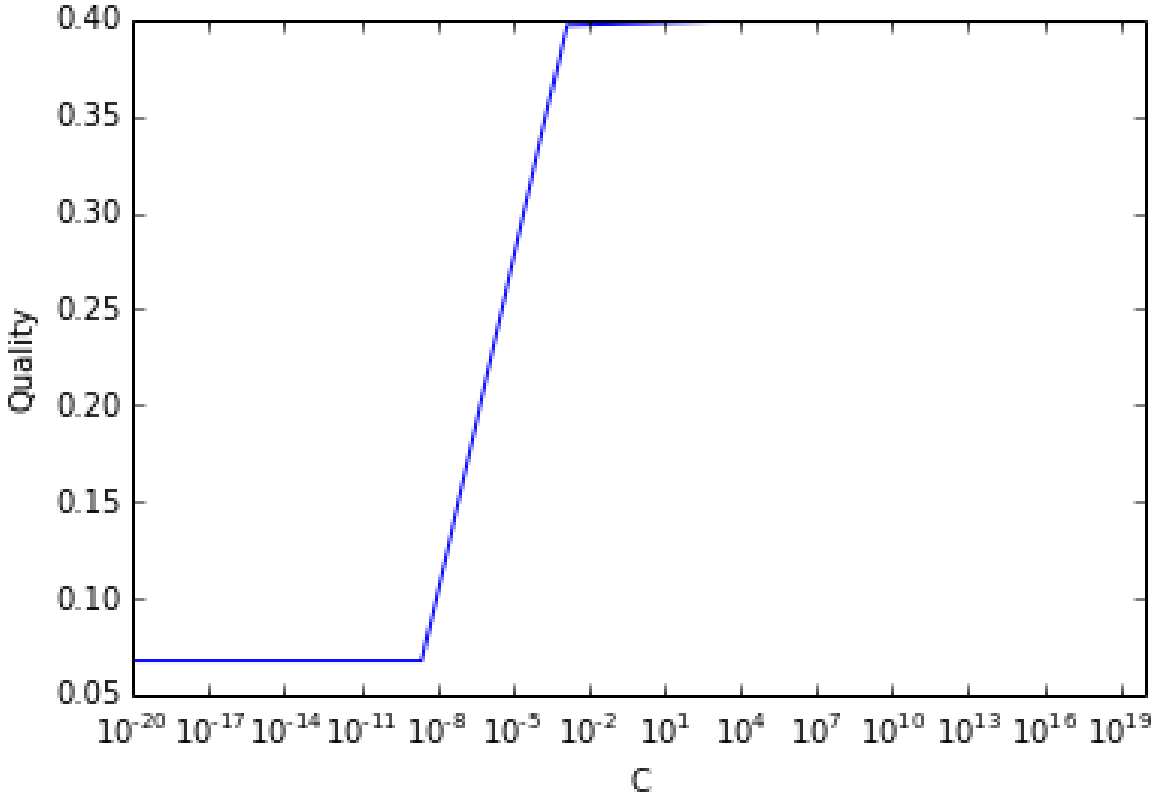
- обернут в OneVsRest классификатор
- веса соседов в соответствии с расстоянием до них

SVM (F1=0.39)

- обернут в OneVsRest классификатор
- на базе libsvm
- ядро - линейное

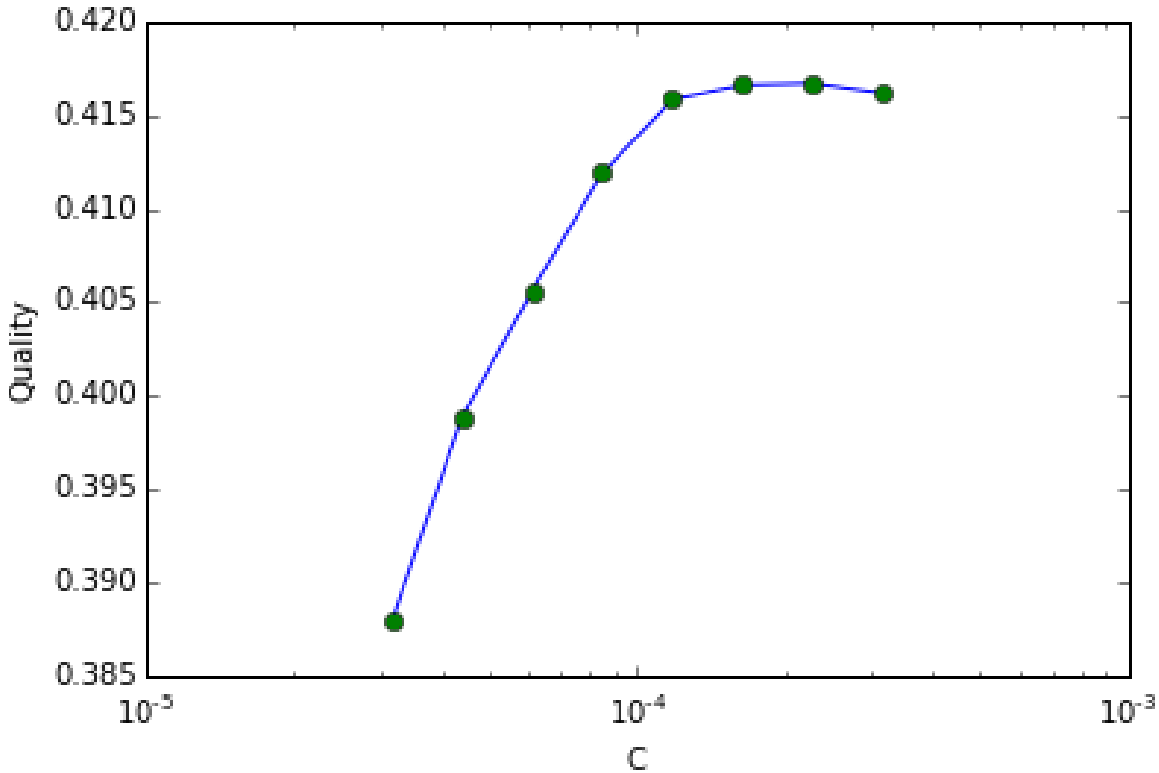
Настройка параметра C

Quality – средний F1-score на нескольких фолдах (фолдов 3)



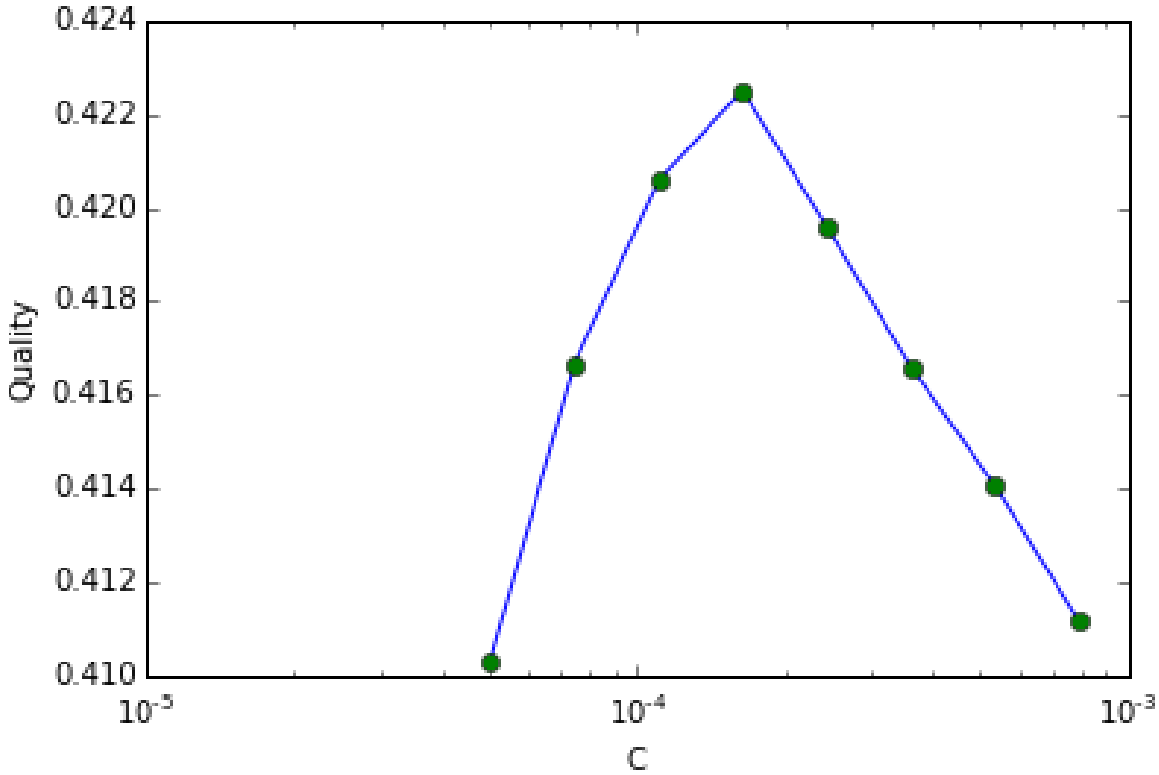
Настройка параметра C

Quality – средний F1-score на нескольких фолдах (фолдов 3)



Настройка параметра C

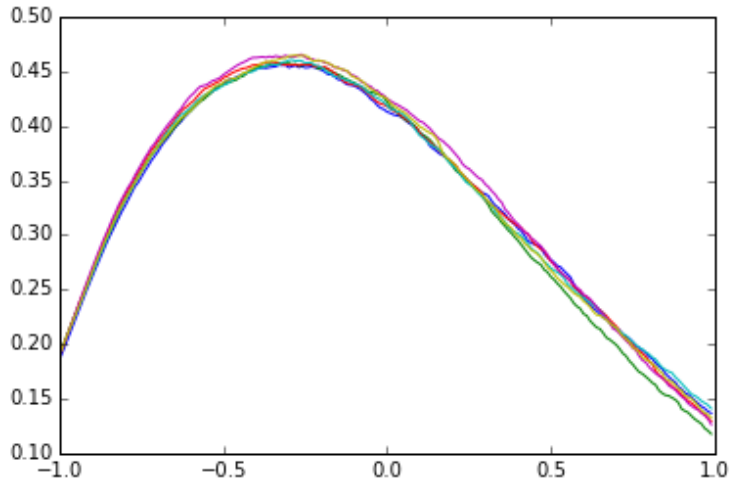
Quality – средний F1-score на нескольких фолдах (фолдов 7)

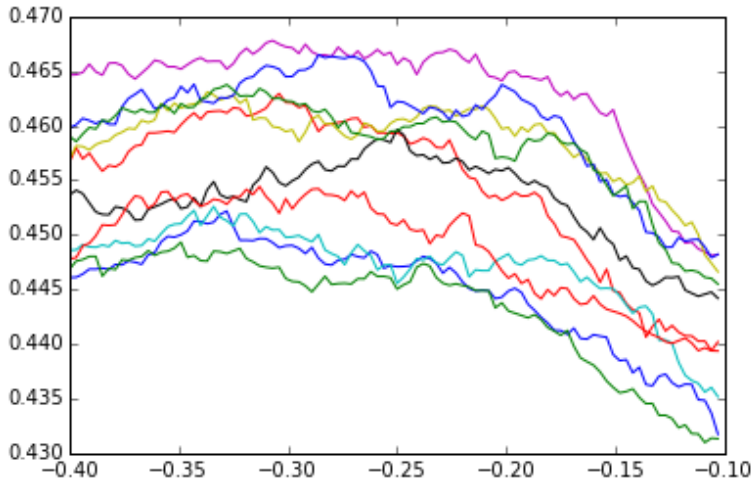


C=0.0017, CrossValidation=0.42, Leaderboard=0.43676

Настройка порога отсечения

Отступы объектов – SVM.decision_function
Quality – F1-score на нескольких фолдах (фолдов 7)

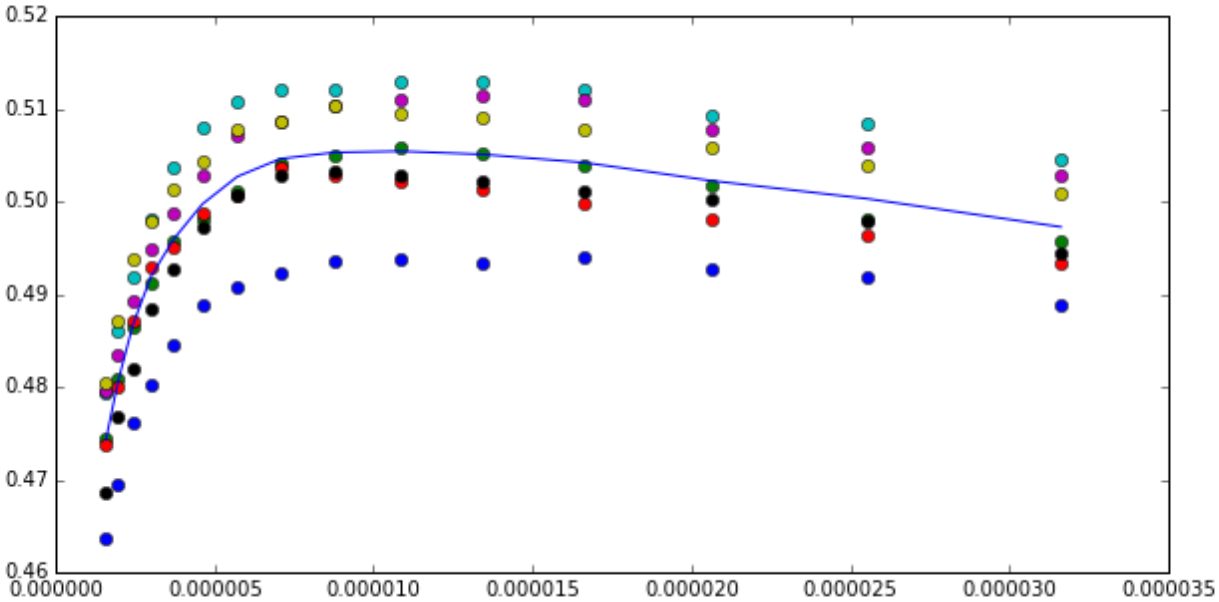




C=0.0017, threshold=-0.32, CrossValidation=0.46,
Leaderboard=0.46773

Настройка C и порога одновременно

Quality – F1-score на нескольких фолдах (фолдов 7)



C=1e-05, threshold=-0.4, CrossValidation=0.505, Leaderboard=0.51376

Привет!