# Отчет по Competetion 3

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

Морозов Алексей, ВМК МГУ

12 мая 2015

# Формулировка задачи

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

Задача: Категоризация научных статей

Функционал качества: Mean F Score= $\frac{2*\text{precision}*\text{recall}}{\text{precision}+\text{recall}}$

# Формулировка задачи

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

Данные:

Дана матрица важности слов в научных статьях, 10 000 статей, 25 560 слов.

Ответы:

Бинарный вектор из 83 меток.

# Основные проблемы задачи

Проблемы данных:

- Данные даны просто в виде разреженной матрицы, конкретный смысл значений "важности" слов неизвестен
- Не дано ни мета-информации, ни сырых данных

Проблемы ответов:

- О настоящем смысле меток также нет никакой информации
- Между ответами существуют зависимости, истинный характер которых неизвестен

# Предобработка исходной выборки

Данные нормализованы и приведены к интервалу [0; 1]

# Обработка ответов

## Свойства задачи категоризации

Стандартный подход - классификатор One Vs Rest.

Проблемы:

- При обучении каждого классификатора получаются очень несбалансированные классы
- Предположение о независимости меток неверно

Главная проблема - зависимости между метками.

- От зависимости в общем случае не избавиться никак
- Однако, можно попытаться избавиться от коррелированности

# Обработка ответов

## PCA - Метод главных компонент

Метод главных компонент - метод сжатия размерности пространства. Его основой является SVD - Singular Value Decomposition.

- Одно из основных свойств PCA - некоррелированность столбцов преобразованной матрицы
- Преобразуем ответы с помощью метода главных компонент с числом компонент, равным исходному
- Используем результат в качестве новых признаков

# Обработка ответов

## Гребневая регрессия

- Разложение ответов известно только для обучающей выборки
- Восстановим их для тестовой с помощью гребневой регрессии
- Параметр $\lambda$ для каждого столбца будем подбирать отдельно

# Классификация

Обучим несколько One Vs Rest классификаторов

- Nearest Neighbors:
  - Косинусная метрика, выборка - исходная матрица
  - Евклидова метрика, выборка - только pca-признаки
- Support Vector Machine:
  - RBF ядро, выборка - только pca-признаки
  - Линейный, выборка - объединение исходной матрицы и pca-признаков
- Во случаях pca-признаки можно нормализовать к [0, 1], а можно оставить как есть. Нормализованные показывают лучший результат.

# Лучшие модели

- Linear SVM: C = 1000, class_weight = auto
  Качество: отдельно не проверялось
- SVM RBF: C = 1000, $\gamma = 3$, class_weight = auto
  Качество: 0.52246 public, 0.51561 private.
- Nearest Neighbors: k = 60 или k = 600
  Качество: отдельно не проверялось

# Итоговый классификатор

Итоговый классификатор получен объединением ответов нескольких моделей

- Nearest Neighbors:
  - Косинусная метрика, k = 60 и k = 600;
  - Евклидова метрика, k = 60 и k = 600;
- Linear SVM:
  C = 1000, нормализованные и ненормализованные pca-признаки
- SVM RBF:
  Нормализованные признаки:
  - C = 1000
  - $\gamma$ = 1, 1.5, 2, 2.7-3.3
  - class_weight = auto

  Ненормализованные признаки:
  - C = 1000
  - $\gamma$ = 0.3, 0.4, 0.5, 1, 1.2, 2
  - class_weight = auto

# Возможные улучшения

- Улучшение качества регрессии приведет к значительному улушению качества классификации
  RBF SVM с настоящими PCA-признаками и в обучении, и в контроле дает идеальное качество на кросс-валидации
- Использование метода, описанного в статье: общая оптимизационная задача для восстановления новых признаков и для классификации документов:

$$\min_{W,M} \sum_{i=1}^{n} \sum_{l=1}^{m} l(x_i, \tanh(Mx_i), y_{il}, f) + \lambda_1 \sum_{l=1}^{m} \|w_l\|^2 + \lambda_2 \sum_{i=1}^{n} \|\tanh(Mx_i) - h_i\|^2$$

Привет!
Спасибо Маше за шаблон!