Московский государственный университет имени М. В. Ломоносова
Факультет вычислительной математики и кибернетики

---

# Отчет о проделанной работе на соревновании на Kaggle: Предсказание зарплаты по тексту объявления

*Студент 3 курса ВМК (317 группа):*
Оспанов А.М.

Москва, 2015

# Содержание

# 1 Выделение признаков

Как и в любой задаче машинного обучения, нам нужно выделить признаки из исходных данных и представить их в некотором виде для дальнейшей работы. При решении данной задачи были перепробованы разные способы представления данных для обучения. Но хорошо заработал только самый первый придуманный вариант (ирония?).

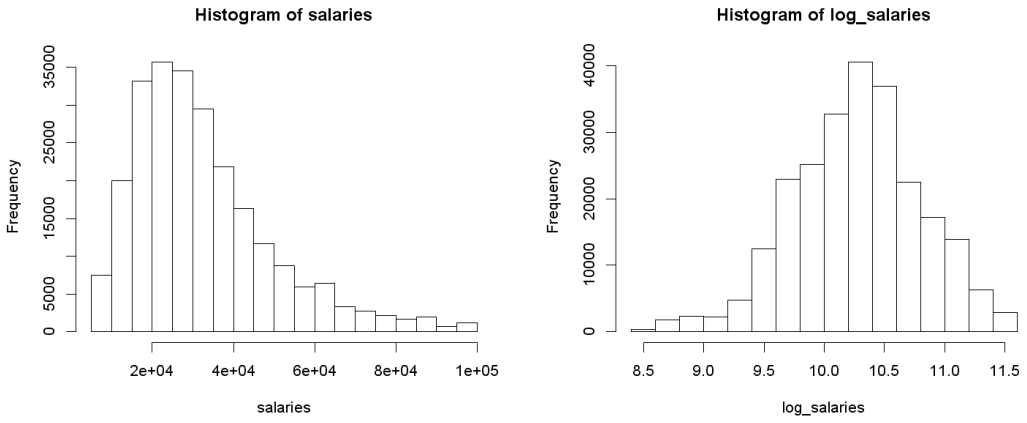
Для обработки данных использовалась библиотека nltk (`http://www.nltk.org/`), а именно: модуль выделения токенов (nltk.tokenize), список стоп-слов (nltk.corpus.stopwords), стеммер (nltk.stem)

Рассмотрим перепробованные варианты представления данных (самые интересные):

1. Title, FullDescription, LocationNormalized, Category
2. Title (делился на три класса: Senior, Junior, Default), FullDescription (топ 50 слов с их количеством), LocationNormalized, Category
3. Title, FullDescription, LocationRaw, Category
4. Все признаки
5. Вместо зарплаты, берется логарифм зарплаты

На всех кроме 1-го варианта не удалось получить хороший результат, что привело к использованию первого варианта и подстройка параметров под эти данные.

Взятие логарифма в пятом примере объясняется следующими рисунками:




т.е. логарифм преобразовывает распределение зарплат к нормальному. Но к сожалению этот вариант не получилось правильно настроить.

## 2 Подбор параметров для Vowpal Wabbit

Функция потерь: --loss_function quantile

n-грамм: --ngram: лучший результат показал 2-грамм без пропусков

Далее настраивались параметры шага градиентного спуска. Лучшие результаты:

--initial_t 0.5: параметр улучшил скорость сходимости

--power_t 0.3: параметр улучшил точность

-l 600: параметр улучшил точность

L1-регуляризатор: --l1 0.00000005: увеличил точность метода

Количество итераций: 200: сошелся за 188 шагов

Количество битов в таблице признаков: -b 27

Квадратичная зависимость: -q tc (t - title, c - category)

## 3 Обучение и тестирование

Обучение велось на первых 100000 элементах обучающей выборки и тестировалось на 60000 последних элементах. После каждого подбора параметров считалась ошибка и подбор происходил, пока ошибка уменьшается. Как только получили хороший результат, обучаемся на полных данных и получаем ответы для контрольной выборки. Убеждаемся, что в системе Kaggle все отлично и радуемся или увидев результат хуже, огорчаемся.

## 4 Обработка ответов

Из исходных данных можно заметить, что зарплаты заданы целыми числами, кратными 100. Следовательно ответ нужно так же нормировать. Концы зарплат, близких к 500, например принадлежащих интервалу 250-750, округлить до 500 и т.д. Также можно заметить, что нет зарплат очень больших (>100000) и очень маленьких (<13000). Их можно отсечь, т.е. приравнять порогу.

P.S. Эту идею мне подсказали =). На самом деле, очень хорошая идея и более того рабочая!

# 5 Заключение

В итоге получаем подобные следующему данные:

50000 |title charter senior quantiti surveyor |fulldescr award win multi disciplinari consult look experienc mric fric senior quantiti surveyor join success grow quantiti survey team this excel opportun experienc charter quantiti surveyor take team manag respons whilst maintain hand involv provid grow quantiti survey team south west area you high involv busi develop work alongsid director cost relat project you manag small team quantiti surveyor mentor apc junior quantiti surveyor guid charter qs varieti project pre contract post contract the success candid need charter mric fric senior quantiti surveyor consult experi you need client face senior quantiti surveyor proven track record busi develop experi grow qs divis candid must dynam ambiti enthusiast good interperson busi develop skill you need broad rang sector experi proven consult experi deliv project educ offic retail commerci sector some experi work hous associ sector would advantag essenti if feel relev experi would like discuss fantast opportun detail pleas call lee faux appli onlin email |loc somerset |category consult job

следующую команду для обучения:

vw -d train.txt --passes 200 --ngram 2 -k -c -f model.vw --loss_function quantile -l 600 -q tc -b 27 --initial_t 1 --power_t 0.3

и следующую команду для получения ответов:

vw -d test.txt -i model.vw -t -p predictions.txt

В системе получаем 5067.83620 баллов на 31% выборке и 5158.47037 баллов на 69% выборке, выходим на 7 место и радуемся :)