Отчет по Competetion 1

CMC MSU, Machine Learning (Spring 14/15) [Kaggle.com]

Вихрева Мария, ВМК МГУ

8 апреля 2015

Формулировка задачи

CMC MSU, Machine Learning (Spring 14/15) [Kaggle.com]

Задача: предсказать зарплату по тексту объявления

Функционал качества: $\text{MAE} = \frac{1}{l}\sum_{i=1}^{l}|y_i - \hat{y}_i|$

Формулировка задачи

CMC MSU, Machine Learning (Spring 14/15) [Kaggle.com]

Признаки:

Id	–	идентификатор объявления
Title	–	название должности
FullDescription	–	текстовое описание вакансии. Из текста удалены все упоминания о размере заработной платы
LocationRaw	–	место работы в свободном формате
LocationNormalized	–	место работы после обработки, принимает одно из значений, описанных в Location_Tree.csv
ContractType	–	вид контракта, полная или частичная занятость
ContractTime	–	срок контракта, постоянный или срочный
Company	–	название работодателя
Category	–	одна из 30 категорий вида работы
SalaryRaw	–	текстовое описание размера зарплаты
SalaryNormalized	–	годовая зарплата, извлеченная из предыдущего поля (целевой признак)
SourceName	–	электронный адрес автора объявления

Предобработка

1. стандартный стемминг Snowball признаков 'Title', 'FullDescription', 'LocationRaw', 'Company', 'Category', 'SourceName'
2. объединение признаков в окружения

Признаки	vw окружение
'Title'	title
'Title', 'FullDescription'	description
'LocationRaw', 'LocationNormalized'	location
'ContractType', 'ContractTime', 'Company', 'Category', 'SourceName'	others;

3. для различия объектов проставляется метка с помощью его 'Id';
4. в качестве целевого значения каждого объекта берем 'SalaryNormalized' для обучающей выборки и 1 для тестовой.

Предобработка

Шаблон строки train.vw :

50000 ’23 |title ... |description ... |location ... |others ...

Типичная строка train.vw :

24000 ’6 |title measur surveyor |description measur surveyor our client a retail space plan survey compani have a 6month contract for a measur surveyor work on a larg nation project for a high street retail you will need to be happi to travel around the countri base from home and stay away for part of the week expens cover strong autocad or revit skill and experi in measur survey is essenti |location Raw^bristol Norm^Bristol |others Company^NaN ContractType^NaN ContractTime^contract Category^properti^job SourceName^hay^co^uk

Предобработка

Шаблон строки test.vw :

1 '1000023 |title ... |description ... |location ... |others ...

Типичная строка test.vw :

1 '1000000 |title html develop |description html develop html develop our client a communic leader has a requir for two html develop to join a larg six month contract base in stirl you will be involv in a web base applic for an extern client where you will be initi develop a prototyp and then creat the function websit as part of the project complet requir html css javascript jqueri twitter bootstrap or similar hay specialist recruit limit act as an employ agenc for perman recruit and employ busi for the suppli of temporari worker by appli for this job you accept the t c s privaci polici and disclaim which can be found on our websit |location Raw^stirlingshir Raw^stirl Raw^fk7 Raw^0 Norm^stirl |others Company^hay^it ContractType^NaN ContractTime^contract Category^it^job SourceName^jobserv^com

Настройка параметров

--holdout_after arg

– vw валидируется на всех объектах, начиная с arg-ого

– vw прекращает обучение, если качество на валидационной выборке снижается

– помогает настроить количество проходов (passes)

Настройка параметров

- b arg
  - размер хэш-функции
  - количество признаков 2^{arg}
  - на практике многих задач показано: чем больше arg, тем лучше

Настройка параметров

-q arg1arg2

– arg1, arg2 – первые символы имен окружений

--cubic arg1arg2arg3

– arg1, arg2, arg3 – первые символы имен окружений

--ignore arg

– добавляет окружение, первый символ которого arg

--keep arg

– удаляет окружение, первый символ которого arg

Пример:

-q a: --ignore b --ignore c

Регуляризация

--l1 arg
--l2 arg
--ftrl --trl_alpha arg1 --ftrl_beta arg2

– сочетание l1 и l2 регуляризации per-Coordinate FTRL-Proximal

– только для логистической регрессии!

Лучшая линейная модель

--passses 80 -b 27 --learning_rate 2000 --holdout_after 110000 --ngram 2

Private Leaderboard score = 5178

Настройка параметров

--nn arg

– нейронная сеть с arg нейронами на скрытом слое

vw-hypersearch

Пример:

vw-hypersearch -t valid.dat -L 1e-10 5e-4 vw --l1 % train.dat

-t file

считает качество на валидационной выборке file

-L

осуществляет log-поиск вместо обычного (использовать для небольших величин параметров)

Располагается : vowpal_wabbit/utl/vw-hypersearch

vw-varinfo

Пример:

vw-varinfo --l1 0.0005 -c --passes 40 train.dat

- train.dat – обучающая выборка
- все, что между vw-varinfo и train.dat – параметры команды vw

Располагается : vowpal_wabbit/utl/vw-varinfo

Вывод varinfo

ТОП-10 признаков с положительными весами:

FeatureName	Weight	RelScore
title^director	+6176.0700	100.00%
others^SourceName^theladd^co^uk	+4662.8900	75.50%
others^ContractType^NaN	+4530.4200	73.35%
others^ContractType^full_time	+4254.6200	68.89%
others^ContractTime^contract	+4234.1600	68.56%
others^ContractTime^NaN	+3867.0700	62.61%
others^ContractTime^permanent	+3739.0900	60.54%
title^manag	+3396.0900	54.99%
title^head	+3303.0300	53.48%
others^SourceName^careerbuild^com	+3256.8600	52.73%

Вывод varinfo

ТОП-10 признаков с отрицательными весами:

FeatureName	Weight	RelScore
description^dmb	-1434.4800	-23.23%
title^junior	-1439.4800	-23.31%
others^Company^ecm^select	-1445.6000	-23.41%
others^SourceName^inspir^intern	-1597.8000	-25.87%
others^Company^lipson^lloyd^jone^north	-1684.4000	-27.27%
title^labour	-1747.3500	-28.29%
title^assist	-1776.9500	-28.77%
title^apprentic	-1932.9200	-31.30%
others^Category^part^time^job	-2175.6700	-35.23%
others^SourceName^elanc	-2175.6700	-35.23%

Финальная команда

Обучение:

vw -d ./train.vw -c -k --passes 27 -b 27 --loss_function quantile --learning_rate 1500 --nn 300 --ngram 2 -f ./model.vw

Private Leaderboard score = 4459

Предсказание:

vw -t -d ./test.vw -i ./model.vw --loss_function quantile -p ./pred.txt

Создание submission-файла:

ipython notebook pred2sub.ipynb

Привет!