딥러닝 기반 핵심 산업별 빅데이터 분석

<머신러닝&딥러닝 파일럿 프로젝트>

주 제	Otto Group Product Classification Challenge	**링 크**	https://www.kaggle.com/c/otto-group-product-classification-challenge
팀 명	하드캐리	**일 자**	2018년 11월 23일
팀 장	강호영<hoyoungk12@naver.com>	**팀 원**	이상훈, 유영재

1. 과제 개요

팀의 목표	- 수업에서 다뤘던 딥러닝의 다양한 기법들을 실제 실행하고 튜닝하면서 모든 팀원들이 딥러닝을 체득화 - 도전하는 모델 자체가 실용성을 가지고 있는 것
목표에 부합하는 데이터셋	- 딥러닝 기법을 바로 적용해볼 수 있기 위해 데이터 전처리는 최소화할 수 있는 데이터 *Ex) feature들이 잘 정리되어 있고 가급적 수치적으로 표현된 데이터* - 데이터가 분산되지 않고 train과 test로 형식으로 정리된 데이터 - 짧은 시간에 완성해야 하므로 도메인에 대한 전문 지식이 필요하지 않은 데이터 - 예측모델 자체가 기업이 현업에서 필요로 하는 실용성을 갖춘 데이터
프로젝트 주제	Otto Group Product Classification Challenge (제품 카테고리 분류 모델)
Otto Group 개요	- Otto Group은 세계적으로 유명한 e-commerce 회사 - 20개국이 넘는 곳에 자회사를 갖춤, 매일 전 세계에 수백만 개의 제품들을 판매
제품 카테고리 분류가 필요한 이유	- 제품의 성능에 대한 일관된 분석이 판매에 중요한 요소 > but 다양한 글로벌 인프라로 동일한 제품들이 다르게 분류되어 일관된 분석이 나오지 않는 문제 > 이를 해결하기 위해 유사한 제품들끼리 분류에 군집화하는 작업이 필요

2. 데이터 설명

<table>
<tr><td>데이터 종류</td><td colspan="2">train</td><td>Test</td></tr>
<tr><td>데이터 구성</td><td colspan="2">id (int형)
feat_1 ~ feat_93 (int형)
target (str형)</td><td>id (int형)
feat_1 ~ feat_93 (int형)</td></tr>
<tr><td rowspan="3">각 열의 특징</td><td>id</td><td colspan="2">각 제품에 대해 번호를 매김
<table>
<tr><th>id</th><th>feat_1</th><th>feat_2</th><th>feat_3</th><th>feat_4</th><th>feat_5</th></tr>
<tr><td>1</td><td>1</td><td>0</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td>2</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td>3</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td>4</td><td>1</td><td>0</td><td>0</td><td>1</td><td>6</td></tr>
<tr><td>5</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr>
</table></td></tr>
<tr><td>Feat</td><td colspan="2">93가지의 특징 중 각 제품에 해당하는 요소를 수치화한 데이터
<table>
<tr><th></th><th>id</th><th>feat_1</th><th>feat_2</th><th>feat_3</th></tr>
<tr><td>count</td><td>61878.000000</td><td>61878.00000</td><td>61878.000000</td><td>61878.000000</td></tr>
<tr><td>mean</td><td>30939.500000</td><td>0.38668</td><td>0.263066</td><td>0.901467</td></tr>
<tr><td>std</td><td>17862.784315</td><td>1.52533</td><td>1.252073</td><td>2.934818</td></tr>
<tr><td>min</td><td>1.000000</td><td>0.00000</td><td>0.000000</td><td>0.000000</td></tr>
<tr><td>25%</td><td>15470.250000</td><td>0.00000</td><td>0.000000</td><td>0.000000</td></tr>
<tr><td>50%</td><td>30939.500000</td><td>0.00000</td><td>0.000000</td><td>0.000000</td></tr>
<tr><td>75%</td><td>46408.750000</td><td>0.00000</td><td>0.000000</td><td>0.000000</td></tr>
<tr><td>max</td><td>61878.000000</td><td>61.00000</td><td>51.000000</td><td>64.000000</td></tr>
</table>
각 feature마다 최대값이 다른 것을 확인할 수 있음</td></tr>
<tr><td>target</td><td colspan="2">제품 카테고리
<table>
<tr><th>target</th></tr>
<tr><td>Class_1</td></tr>
<tr><td>Class_1</td></tr>
<tr><td>Class_1</td></tr>
<tr><td>Class_1</td></tr>
<tr><td>Class_1</td></tr>
</table></td></tr>
</table>

제출 형태	각 제품마다 분류군(9가지)에 대한 예상 확률 값을 담아 csv파일 제출
	`id,Class_1,Class_2,Class_3,Class_4,Class_5,`
	`1,0.0,0.0,0.0,0.0,1.0,0.0,0.0,0.0,0.0,0.0`
	`2,0.0,0.2,0.3,0.3,0.0,0.0,0.1,0.1,0.0`
	`...`
	`etc.`

평가 방식	다중 클래스 로그 손실을 이용하여 평가 $$logloss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log(p_{ij}),$$ 결론: 손실 값이 낮을수록 예상을 잘하는 모델
데이터 선택한 이유	- 데이터 종류가 2가지로 단순함 - 대부분의 데이터가 머신 러닝을 바로 적용할 수 있는 숫자형 데이터 - 실제 기업에서 필요로 하는 실용성 - 도메인에 대한 깊은 전문지식이 필요하지 않은 용이한 접근성

3. 과제 수행 내역

4. 결과 보고

.

감사합니다