딥러닝 기반 핵심 산업별 빅데이터 분석

<머신러닝&딥러닝 파일럿 프로젝트>

주 제	House Prices: Advanced Regression Techniques	링 크	https://www.kaggle.com/c/house-prices-advanced-regression-techniques
팀 명	D팀	일 자	2018년 11월 29일
팀 장	조민정<m1nch0@naver.com>	팀 원	고준형, 김효신

1. 과제 개요

사람들이 일반적으로 생각하는 집값에 영향을 미치는 요소(평수, 침실 수, 위치 등) 외에도 어떠한 요소들이 실제 집값에 영향을 끼치는지 알아보려고 한다. 해당 분석 결과는 부동산 구입을 앞둔 사람들이 가격 협상에 나설 때 매도가 조정을 유도할 수 있는 핵심 요인들을 발견하고 이용할 수 있도록 도와줄 것이다.

1) 주제

Ames, Iowa에 있는 주거용 건물들의 79개의 변수들을 분석하고 향후 집값 예측하기

2) 배울 수 있는 내용

고급 선형회귀 스킬 및 텐서플로우를 이용한 예측 모델 구성, 예측 오차율을 최소화하기.

3) 주제 선정 이유

- ✓ 조원들의 수준에서 배운 내용을 전부 응용, 실습해보기에 가장 적절한 데이터
- ✓ 향후 프로젝트에서 필요한 예측 개념을 이해하고 머신러닝/딥러닝 기초를 복습할 수 있다.
- ✓ 본 분석의 데이터는 선형회귀모델을 공부하는 데이터 사이언티스트들을 위해 주어졌다,
- ✓ 동일한 집값 예측 데이터인 Boston Housing(<모두의 딥러닝> 예제) 에서는 데이터가 이미 충분히 정제되고 가공되어 쉽게 분석이 가능했다면, 본 데이터는 전처리 과정을 실습하고 예측모델에 맞게 가공하는 훈련까지 할 수 있어 선정했다.
- ✓ Boston Housing (https://www.kaggle.com/schirmerchad/bostonhoustingmlnd)

2. 데이터 설명

1) train.csv

1461 x 81

주요 데이터 필드

SalesPrice(예측해야 하는 것, 집값), LotShape(평수), Kitchen(부엌 수), bedroom(침실 수), Yearbuilt(연식), Foundation(지반), Electronial(전기타입), Neiborhood(시내와의 물리적 거리), HouseStyle(주거타입), SaleCondition(판매 상태) 등 총 81개 칼럼

2) test.csv

3) data_description.txt

각 칼럼에 대한 세부 설명

4) sample_submission.csv

침실 수와 평방미터, 연간 판매량을 통해 도출한 선형 회귀 모델의 결과 샘플

3. 과제 수행 내역

1. 데이터 탐색
   - 컬럼간의 상관관계 및 필요한 데이터 지표 파악
     - 상관계수를 통해 SalePrice(집값)과 81개 지표간의 상관관계 파악, 상위 10가지 피쳐와 하위 10가지 피쳐 파악
2. 데이터 시각화
   - 주요 데이터 시각화 및 그래프를 통해 데이터 상관관계 파악
     - 비슷한 피쳐들끼리 묶어서 SalePrice(집값)과의 관계 파악 및 이상치 제거
3. 데이터 전처리(정제)
   - 데이터 품질 높이기
     - 결측치 대치, 불필요한 피처 제거
   - 데이터 타입에 따른 분류 및 수치화
     - 카테고리 데이터: Unique Value를 숫자로 변경
     - 서수 데이터: 0과 1 사이의 값으로 표준화
   - 데이터 표준화
     - Standardization: (요소값 - 평균)/표준편차

   전처리 전

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	...	PoolArea	PoolQC	Fence	MiscFeature	MiscVal	MoSold	YrSold	SaleType	SaleCondition	SalePrice
0	1	60	RL	65.0	8450	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0	2	2008	WD	Normal	208500
1	2	20	RL	80.0	9600	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0	5	2007	WD	Normal	181500
2	3	60	RL	68.0	11250	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	9	2008	WD	Normal	223500
3	4	70	RL	60.0	9550	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	2	2006	WD	Abnorml	140000
4	5	60	RL	84.0	14260	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	12	2008	WD	Normal	250000
5	6	50	RL	85.0	14115	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	MnPrv	Shed	700	10	2009	WD	Normal	143000
6	7	20	RL	75.0	10084	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0	8	2007	WD	Normal	307000
7	8	60	RL	NaN	10382	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	Shed	350	11	2009	WD	Normal	200000
8	9	50	RM	51.0	6120	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0	4	2008	WD	Abnorml	129900
9	10	190	RL	50.0	7420	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0	1	2008	WD	Normal	118000

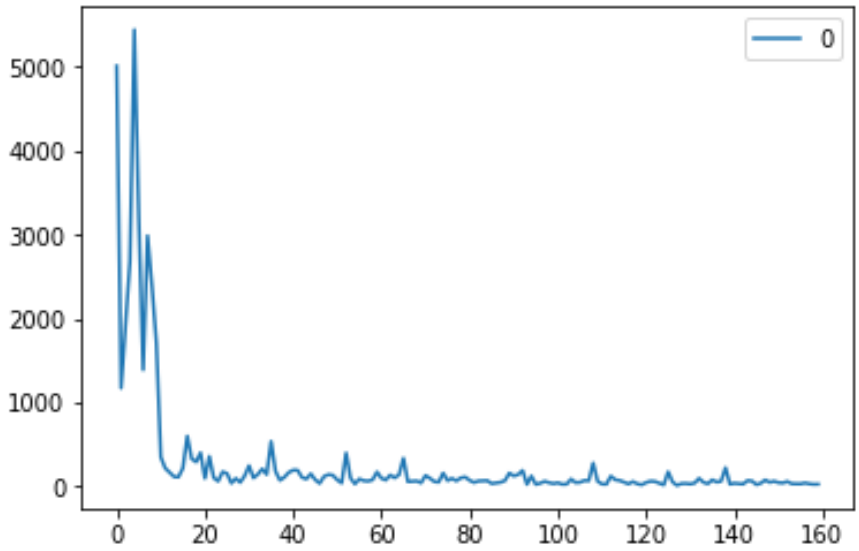
   전처리 후

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	LotShape	LandContour	Utilities	LotConfig	...	GarageQual	GarageCond	PavedDrive	WoodDeckSF	OpenPorchSF	EnclosedPorch	ScreenPorch	SaleType	SaleCondition	SalePrice
0	1	0.073350	1	65.0	-0.207071	0.75	0.875	1	0.75	1	...	0.1	0.5	0.833333	-0.751918	0.216429	-0.359202	-0.270116	1	1	208500
1	2	-0.872264	1	80.0	-0.091855	0.75	0.875	1	0.75	2	...	0.1	0.5	0.833333	1.625638	-0.704242	-0.359202	-0.270116	1	1	181500
2	3	0.073350	1	68.0	0.073455	0.75	0.625	1	0.75	1	...	0.1	0.5	0.833333	-0.751918	-0.070337	-0.359202	-0.270116	1	1	223500
3	4	0.309753	1	60.0	-0.096864	0.75	0.625	1	0.75	3	...	0.1	0.5	0.833333	-0.751918	-0.175988	4.091122	-0.270116	1	2	140000
4	5	0.073350	1	84.0	0.375020	0.75	0.625	1	0.75	2	...	0.1	0.5	0.833333	0.779930	0.563567	-0.359202	-0.270116	1	1	250000
5	6	-0.163054	1	85.0	0.360493	0.75	0.625	1	0.75	1	...	0.1	0.5	0.833333	-0.432783	-0.251453	-0.359202	-0.270116	1	1	143000
6	7	-0.872264	1	75.0	-0.043364	0.75	0.875	1	0.75	1	...	0.1	0.5	0.833333	1.282568	0.156057	-0.359202	-0.270116	1	1	307000
7	8	0.073350	1	0.0	-0.013508	0.75	0.625	1	0.75	3	...	0.1	0.5	0.833333	1.123000	2.374723	3.371217	-0.270116	1	1	200000
8	9	-0.163054	2	51.0	-0.440508	0.75	0.875	1	0.75	1	...	0.3	0.5	0.833333	-0.033864	-0.704242	2.994902	-0.270116	1	2	129900
9	10	3.146594	1	50.0	-0.310264	0.75	0.875	1	0.75	3	...	0.5	0.5	0.833333	-0.751918	-0.643870	-0.359202	-0.270116	1	1	118000

4. 모델링
   - Keras를 이용한 예측모델
   - Tensorflow를 이용한 예측모델
     - Gradient Descent Optimizer
     - Adam Optimizer
     - batch, epoch, learning rate를 조절하여 cost가 최저값으로 수렴하도록 학습
     - cost값이 Inf와 NaN값으로 Overshooting되어 SalePrice를 1/10000으로 조정.(W의 범위가 지나치게 커지지 않도록 만드는 효과) 조정 후 Cost가 점차 감소하는 학습 그래프를 보임.

4. 결과 보고

- 다중선형회귀 모델로 학습한 결과



<Cost fuction 그래프>

- Cost가 변동폭을 가지나 점차 줄어드는 양상으로 보임.
- 정확한 결과값은 첨부파일(submission_GD_v1.0.csv)로 확인
- AdamOptimizer와 GradientDescentOptimizer를 비교해봤을 때, GradientDescentOprimizer 가 좀 더 정확한 값을 출력하여 GradientDescentOprimizer 채택

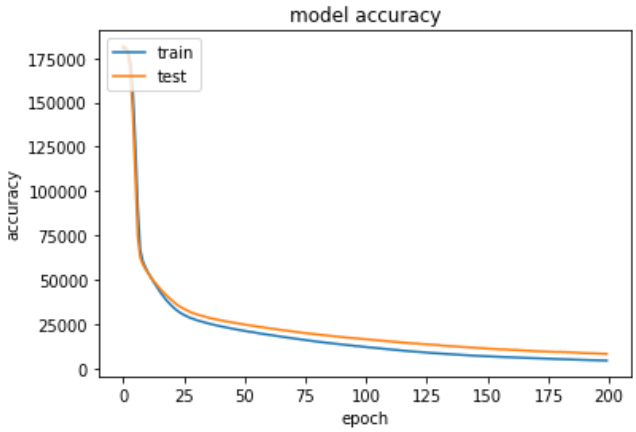
	Y	Result
0	208500.0	[205072.66]
1	181500.0	[194097.31]
2	223500.0	[196610.03]
3	140000.0	[182586.8]
4	250000.0	[268644.16]
5	143000.0	[177048.28]
6	307000.0	[261628.89]
7	200000.0	[256488.42]
8	129900.0	[151728.52]
9	118000.0	[79692.95]

GradientDescent

	Y	Result
0	208500.0	[206972.2]
1	181500.0	[196102.25]
2	223500.0	[221228.16]
3	140000.0	[186386.58]
4	250000.0	[276685.06]
5	143000.0	[204966.02]
6	307000.0	[273158.56]
7	200000.0	[260371.55]
8	129900.0	[148413.58]
9	118000.0	[62225.617]

Adam

- Keras를 이용한 모델로 예측한 결과



- Train 학습값과 Test 예측값이 유사하게 진행되며 예측값을 내는 것을 확인

- 이 학습을 통해 배운 것
  - 머신러닝에 있어서 데이터 전처리의 중요성과 어려움.
  - 종속변수가 너무 크면 W(가중치)도 너무 커지고 가중치간 편차가 심해짐. 전처리 과정에서 표준화(Normalization)하는 것이 중요.
  - Learning rate에 따라 Cost function의 진폭이 점차 커지거나 점차 줄어들게 됨.
  - Keras 모델 사용법
- 프로젝트에서 아쉬운 점
  - cost의 변동폭을 줄일 수 있는 방법(이상치 제거, feature 선택 다양화)들을 충분히 시도해보지 못함.
  - Validation을 통해 overfitting을 체크하지 못함
  - Accuracy를 측정할 수 있는 classification case도 도전해보고자 함

감사합니다