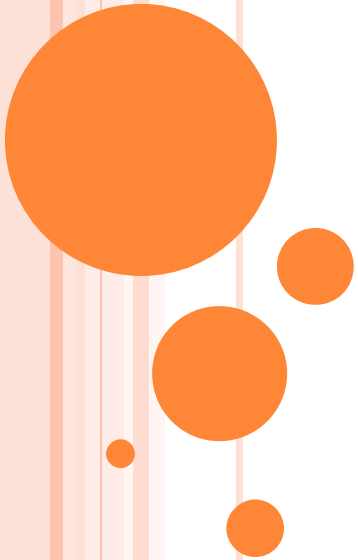


How to Win a Data Science Competition

～WEEK1～

嶋野　友也（2018.08.10）

１．WEEK１の講座内容

●Recap of main Machine Learning Algorithm

(1)Linear：separate 2 parts （ex. Logistic Regression、Support Vector Machine）
(2)Tree-Based：（ex. Decision Tree、Random Forest、GBDT）
(3)kNN
(4)Neaural Net
GBDT,Neaural Net is most powerful methods

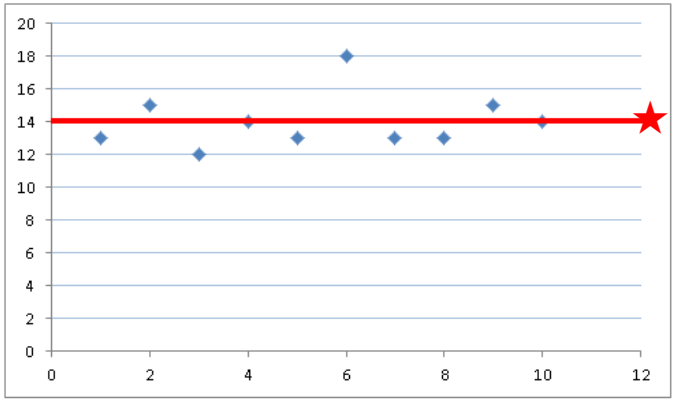
●Feature Processing and generation with respect to model

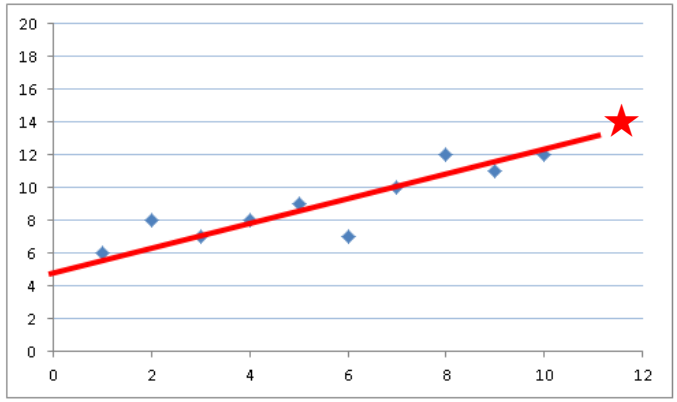
(1)numeric（数値）
　(a)Scaling：違う幅だと意味をなさないから揃える
　　・MinMax Scaler：最小最大値基準
　　・Standard Scaler：平均値基準
　(b)OutLier：異常値を排除することでズレを防止
　(c)Rank
　(d)Log、Sqrt
(2)categorical（カテゴリー）
(3)ordinal（順序付カテゴリー）　順序はあるが差分に意味がない⇒数値とは別（ex.小学校、中学校、高校）
　→Non Tree Basedの場合：One-Hot-Encoding
　→Tree Based の場合：Label Encoding、Frequency Encoding
(4)datetime（日付）
(5)coordinate（座標）

２．売上予測のイメージ

●売上予測のイメージ

- 「線形回帰」のイメージ





木とかイメージがわかないし、線形回帰モデルで分析しよう！

※実は違ったんですが・・・

３．分析

●分析（線形回帰）

- 店舗、商品、月ごとの売上数を集計
- 「0」～「32」の売上数から「33」の売上を予測するモデル分析

※売上実績がない月は「0」埋め

		0	1	2	3	4	5	6	7	8	9	...	24	25	26	27	28	29	30	31	32	33
shop_id	item_id																					
2	31	0	4	1	1	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	1
	486	0	0	0	0	0	0	0	0	0	0	...	0	3	2	1	0	2	0	0	1	3
	787	0	0	0	0	0	0	0	0	0	0	...	0	0	1	0	0		0	0	0	1
	794	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	1
	968	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	1
	988	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	1
	1075	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	1	0	0	1
	1121	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	1
	1377	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	1
	1387	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	1

```
clf = linear_model.LinearRegression()
clf.fit(X, Y)
```

４．テスト

●テスト

- テストデータとトレーニングデータを結合（実績値を追加）
- 「１」～「３３」の売上数から次月度の売上を予測（前述のモデルを使用）

  ※実績値が不明なデータは「0」埋め（雑！）

	ID	shop_id	item_id	0	1	2	3	4	5	6	...	24	25	26	27	28	29	30	31	32	33	34
0	0	5	5037	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	2.0	0.0	0.0	0.0	1.0	1.0	1.0	3.0	1.0	0.0	
1	1	5	5320	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
2	2	5	5233	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	3.0	2.0	0.0	1.0	3.0	1.0	
3	3	5	5232	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1	0.0	0.0	
4	4	5	5268	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
5	5	5	5039	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	1.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	1.0	1.0	
6	6	5	5041	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.0	2.0	
7	7	5	5046	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	
8	8	5	5319	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	9.0	2.0	3.0	2.0	2.0	4.0	3.0	2.0	3.0	0.0	
9	9	5	5003	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	

result = clf.predict(X)

- 結果　score 2.93116 （761／784 位） ※7/26時点
- なんか違う！これって過去数値の加重平均じゃん。イメージしたのは時系列分析

５．悪戦苦闘の日々

●課題事項

- ①相関係数

  >print(clf.coef_.round(2))
  [0.01 -0.03 0.02 0.00 -0.06 0.09 ・・・ 0.07 -0.00 0.50 0.01 0.76 -0.11 0.01]

  ３ヶ月前の売上数と５か月前の売上数に大きく依存します。*⇒そんなわけないだろ！！*

  直近３ヶ月平均とか、そういう方が自然でしょ・・・

  [0.00 -0.00 0.00 0.00 -0.00 0.00 ・・・ 0.00 0.10 0.10 0.10 0.25 0.25 0.25]

- ②テストデータの補完（実績なし商品）

  （1）実績なしは「0」補完→「0」でなく全商品の平均値で補完しよう！

  （2）実績なしは「0」補完→「0」ではなく、商品ごとの平均値で補完しよう！

- ③その他

  売上数以外の項目使用、Scaling未実施、トレーニングデータの分析など課題多数

★チャレンジ結果

	date	title	score
第１回	７月２７日	線形回帰	2.93116
第２回	７月２７日	３ヶ月平均値	3.69924
第３回	８月０４日	商品平均値補完	5.90242
第４回	８月０４日	商品ごと平均値補完	8.84502

やればやるほど点数が下がり、テンション下がったので、WEEK2以降の講座を受けてから、がんばります