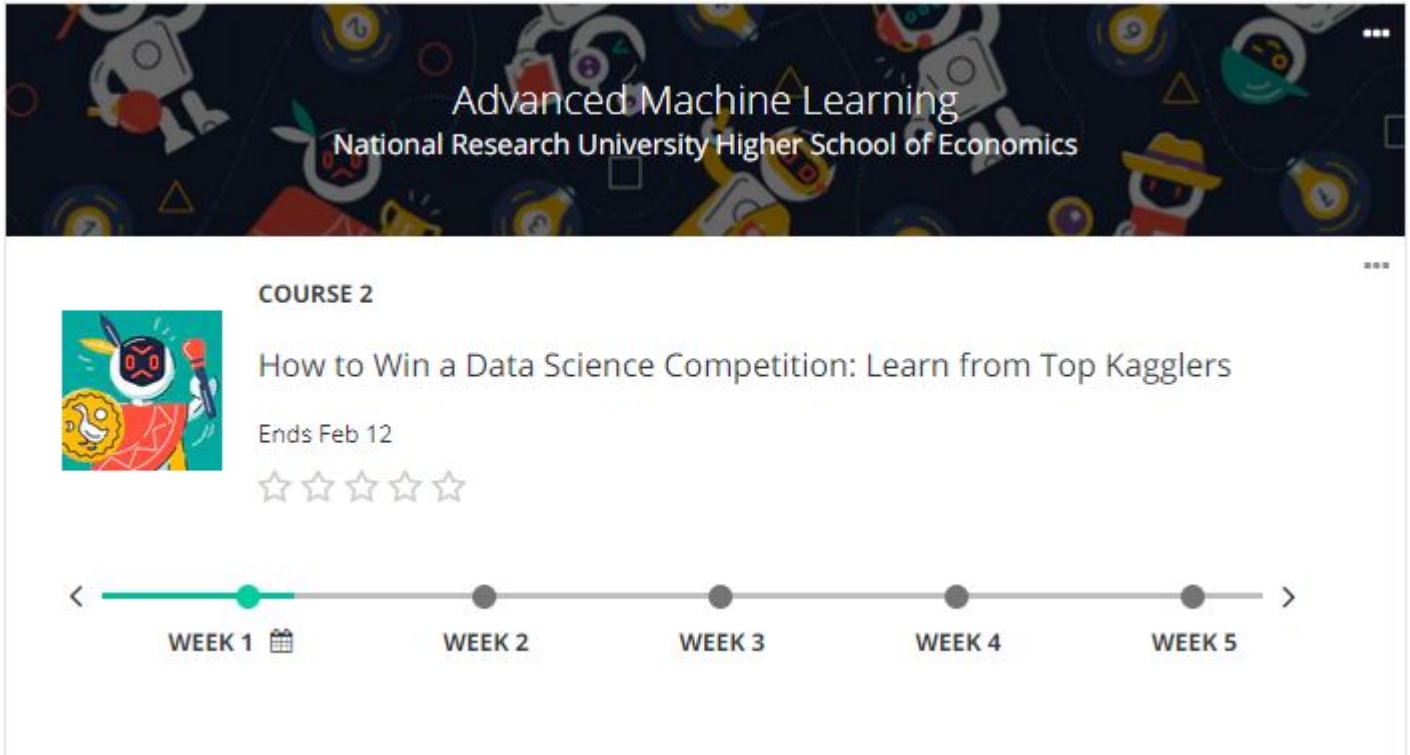
Feature Preprocessing and Generation Tips

Cinyoung Hur
Seoul AI
2018.01.06

I’m taking this course



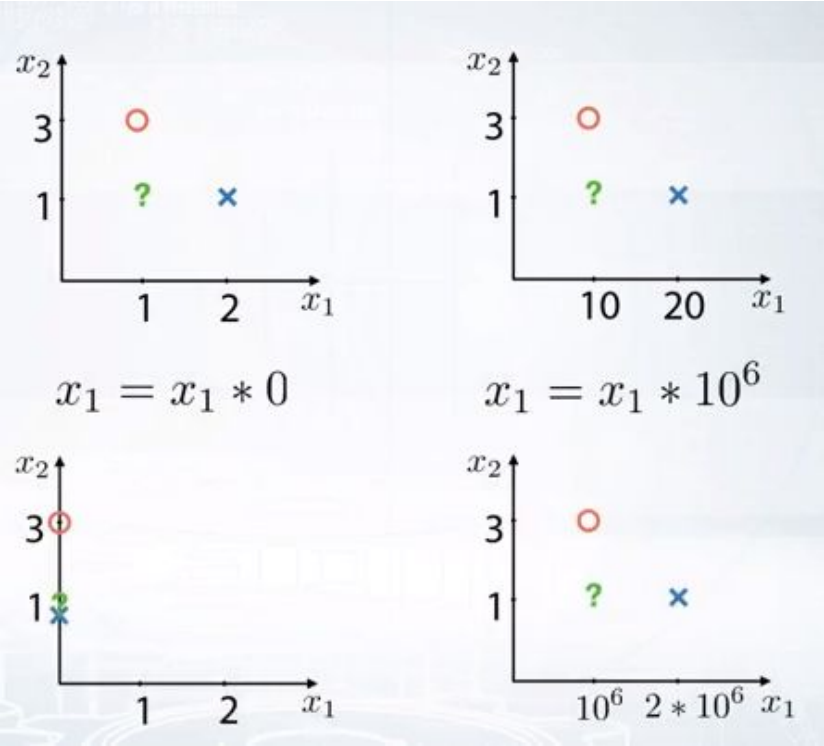
Preprocessing issues

Feature types

Feature types: Numeric features

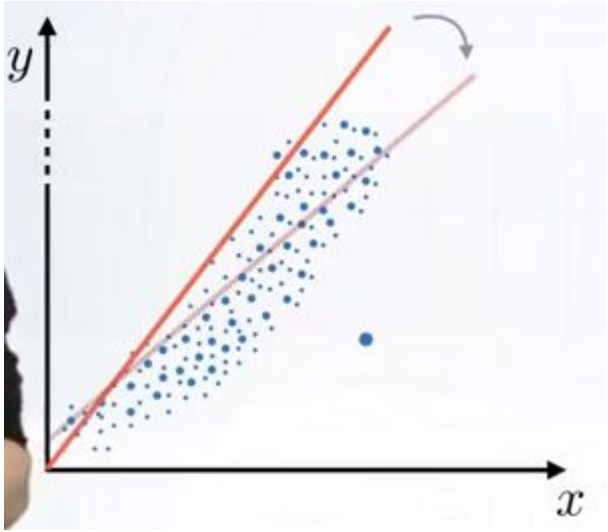
Scaling matters

- [0, 1]
  - MinMaxScaler
- mean=0, std=1
  - StandardScaler



Feature types: Numeric features

Outliers



Feature types: Categorical and ordinal features

Titanic dataset

	PassengerId	Survived	Pclass	Name
0	1	0	3	Braund, Mr. Owen Harris
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...
2	3	1	3	Heikkinen, Miss. Laina
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)
4	5	0	3	Allen, Mr. William Henry
5	6	0	3	Moran, Mr. James
6	7	0	1	McCarthy, Mr. Timothy J
7	8	0	3	Palsson, Master. Gosta Leonard

	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	male	22.000000	1	0	A/5 21171	7.2500	NaN	S
1	female	38.000000	1	0	PC 17599	71.2833	C85	C
2	female	26.000000	0	0	STON/O2. 3101282	7.9250	NaN	S
3	female	35.000000	1	0	113803	53.1000	C123	S
4	male	35.000000	0	0	373450	8.0500	NaN	S
5	male	29.699118	0	0	330877	8.4583	NaN	Q
6	male	54.000000	0	0	17463	51.8625	E46	S
7	male	2.000000	3	1	349909	21.0750	NaN	S

Feature types: Categorical and ordinal features

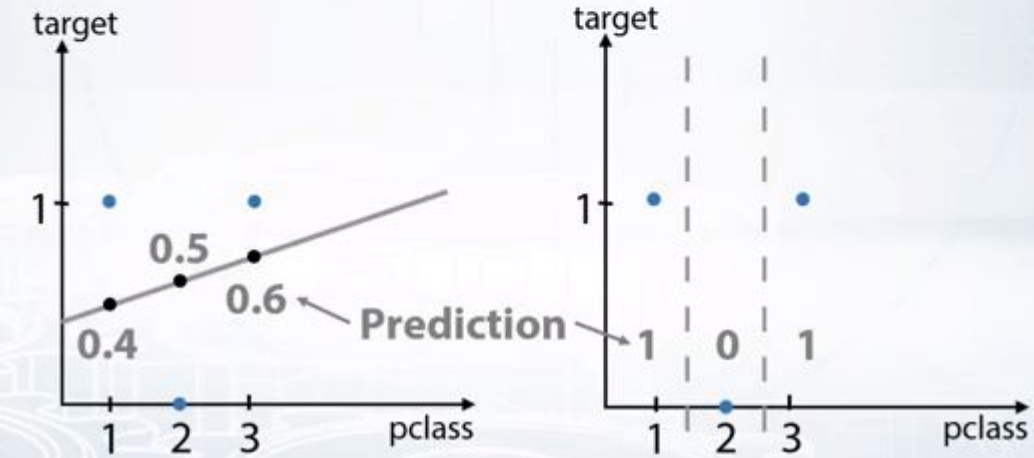
Quiz: Categorical features are beneficial to ( ) model.

1. Tree-based model (RandomForest, Decision tree)
2. Non-tree based model (Linear model, Neural Network)

Feature types: Categorical and ordinal features

Label encoding

pclass	1	2	3
target	1	0	1



Feature types: Categorical and ordinal features

Frequency encoding

K
embarked
S
C
S
S
S
Q
S
S
S
C
S
S

[S,C,Q] -> [0.5, 0.3, 0.2]

```
encoding = titanic.groupby('Embarked').size()
encoding = encoding/len(titanic)
titanic['enc'] = titanic.Embarked.map(encoding)
```

Feature types: Categorical and ordinal features

Quiz: Can frequency encoding be of help for non-tree based models?

1. Yes, it can
2. No, it can’t

Feature types: Categorical and ordinal features

One-hot encoding

One-hot encoding

pclass
1
2
1
3

pclass==1	pclass==2	pclass==3
1		
	1	
1		
		1

Feature generation: Categorical features

pclass	sex	pclass_sex
3	male	3male
1	female	1female
3	female	3female
1	female	1female

Pclass_sex==					
1male	1female	2male	2female	3male	3female
				1	
	1				
					1
	1				

References

- Feature preprocessing
  - Preprocessing in Sklearn
  - Andrew NG about gradient descent and feature scaling
  - Feature Scaling and the effect of standardization for machine learning algorithms
- Feature generation
  - Discover Feature Engineering, How to Engineer Features and How to Get Good at It
  - Discussion of feature engineering on Quora