Data Mining 2

Topic 02 : Feature Engineering

Lecture 01 : Introduction to Feature Engineering

Dr Kieran Murphy

Department of Computing and Mathematics, WIT.
(kmurphy@wit.ie)

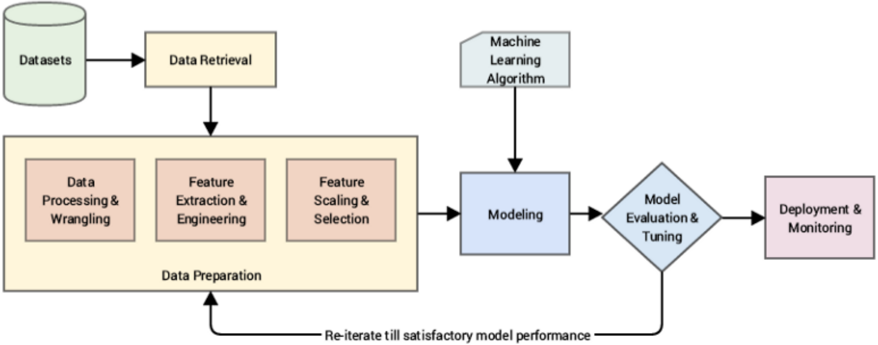
Spring Semester, 2021

Outline

- Need for feature engineering
- Generic feature engineering steps

Outline

Yet Another Data Mining Pipeline . . .



- Varying terminology.
  - Feature extraction and feature engineering are synonyms.
  - Feature transformation (or scaling) deals with scaling and transforming (to get symmetric (normal like) distributions.
  - Feature selection deal with picking subsets of given and generate features to use in the model.

Feature Engineering $\approx$ Extraction $\cup$ Engineering $\cup$ Transformations $\cup$ Selection

What is Feature Engineering?

Feature engineering
The process of transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data.

- is a representation problem*
  '*you have to turn your inputs into things the algorithm can understand*'
  – Shayne Miel
- is an Art
  '*Coming up with features is difficult, time-consuming, requires expert knowledge. "Applied machine learning" is basically feature engineering.*'
  – Andrew Ng
  '*...some machine learning projects succeed and some fail. What makes the difference? Easily the most important factor is the features used.*'
  – Pedro Domingos, (A Few Useful Things to Know about Machine Learning)"

*Recall the three components of a machine learning problem (Representation, Evaluation and Optimisation).

Why Feature Engineering?

- **Better representation of data**
    - Improved representation can be better understood by ML algorithms.
    - Better representations leads to improved visualisation — for example, compare the frequent word occurrences of a newspaper article as opposed to the raw text.
- **Better performing models**
    - The right features tend to give models that outperform other models no matter how complex the algorithm is.
    - In general if you have the right feature set, even a simple model will perform well and give desired results.

Better features make better models.

- **Essential for model building and evaluation**
    - Numerical data is slower on decision trees.
    - Categorical data not suitable for linear regression, recall one-hot encoding.
- **More flexibility on data types**
    - Want to build models on diverse data types (text, images, video).
- **Emphasis on the business and domain**
    - Better features are often motivated by domain experts (not data scientists), leading to more understandable (by humans) models.
    - Feature engineering emphasises to focus on the business and the domain of the problem when building features.

Example

At its simplest, feature engineering can be defined as the process of creating new features from the existing features in a dataset.

Consider a sample data that has details about a few items, such as their weight and price.

Item_ID	Item_Weight	Item_Price
FDA15	9.3	249.81
DRC01	5.9	48.27
FDN15	17.5	141.62
FDX07	19.2	182.10

`Price_per_Weight=Item_Price/Item_Weight`

Item_ID	Item_Weight	Item_Price	Price_per_Weight
FDA15	9.3	249.81	26.86
DRC01	5.9	48.27	8.15
FDN15	17.5	141.62	8.09
FDX07	19.2	182.10	9.48

If, for example, we were interested in a regression task, then, a similar effect could have been achieved by consider a suitable family of non-linear functions. However, generating features in advance allows us to stay with simpler and computationally less expensive families of functions.

Example: (Titanic dataset)

```
df = pd.read_csv('assets/train.csv')
df.head()
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2[illegible]	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.28[illegible]	[illegible]	[illegible]
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.925[illegible]	[illegible]	[illegible]
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

1 Have a (text) feature representing the passenger's name.

Example: (Titanic dataset)

```
df = pd.read_csv('assets/train.csv')
df.head()
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.28		
				Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.925		
				Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
				Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Title	Count
Capt	1
Col	4
Don	1
Dona	1
Dr	8
Jonkheer	1
Lady	1
Major	2
Master	61
Miss	260
Mlle	2
Mme	1
Mr	757
Mrs	197
Ms	2
Rev	8
Sir	1
the Countess	1

1 Have a (text) feature representing the passenger's name.

Example: (Titanic dataset)

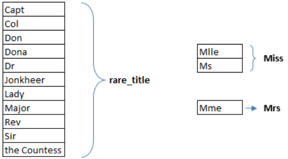
```
df = pd.read_csv('assets/train.csv')
df.head()
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.28		
				Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.925		
				Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
				Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

1 Have a (text) feature representing the passenger's name.

Title	Count
Capt	1
Col	4
Don	1
Dona	1
Dr	8
Jonkheer	1
Lady	1
Major	2
Master	61
Miss	260
Mlle	2
Mme	1
Mr	757
Mrs	197
Ms	2
Rev	8
Sir	1
the Countess	1

2 Rare categories can be merged under a single label – 'Dona', 'Lady', 'the Countess', 'Capt', 'Col', etc. .
Duplicate/redundant categories can be merged – the titles 'Mlle', and 'Ms' and can be merged under 'Miss'. 'Mme' can be merged with 'Mrs'.



Example: (Titanic dataset)

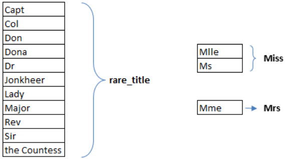
```
df = pd.read_csv('assets/train.csv')
df.head()
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.28	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.925	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
[illegible]	[illegible]	[illegible]	[illegible]	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

1 Have a (text) feature representing the passenger's name.

Title	Count
Capt	1
Col	4
Don	1
Dona	1
Dr	8
Jonkheer	1
Lady	1
Major	2
Master	61
Miss	260
Mlle	2
Mme	1
Mr	757
Mrs	197
Ms	2
Rev	8
Sir	1
the Countess	1

2 Rare categories can be merged under a single label – 'Dona', 'Lady', 'the Countess', 'Capt', 'Col', etc. .
Duplicate/redundant categories can be merged – the titles 'Mlle', and 'Ms' and can be merged under 'Miss'. 'Mme' can be merged with 'Mrs'.



3 New categorical feature with 5 unique values

Master	Miss	Mr	Mrs	rare_title
61	264	757	198	29

Importance of Feature Engineering

- Performance of a predictive model is heavily dependent on the quality of the features in the dataset used to train that model.
- Feature engineering can allow simpler models (linear vs non-linear) to provide required performance.

The Kaggle competition, Bike Sharing Demand Prediction, dealt with forecasting the rental demand based on historical usage patterns in relation with weather, time and other data. Smart feature engineering was instrumental† in securing a place in the top 5 percentile of the leaderboard, using:

Hour Bins Categorise the hour feature with number of bins determined using a decision tree.

Temp Bins: Similarly, a binned feature for the temperature variable for both registered and casual users.

Year Bins: 8 quarterly bins were created for a period of 2 years, i.e., quarterly bins.

Day Type: Days were categorised as ‘weekday’, ‘weekend’ or ‘holiday’.

† Kaggle Bike Sharing Demand Prediction — How I got in top 5 percentile of participants?

Outline

Numeric Data I

Numeric data typically represents data in the form of scalar values depicting

- observations, recordings or measurements
- frequencies or counts

It is very rich data but the pattern of interest may often be lost within the noise. Common feature engineering techniques are

- **Binarisation** — Generate a binary feature from a numeric feature.
- **Rounding** — reduce the precision in the data to simplify decision tree models, or as a step towards representing as categorical.
- **Binning** —- Divide numerical scale into bins (think histogram) to then convert to a categorical feature.
  - Bins can be fixed (equal) width or adaptive.
  - Bin width / number of bins can be determined using decision trees.

Example —- Binarisation

Consider building a recommendation system for song recommendations:

- Want to know if a person is interested or has listened to a particular song.
- If given a dataset with a feature `listen_count`, storing the number of time each individual has listened to each song, then we need to binarise our `listen_count` field as follows:

```
tmp = np.array(df["listen_count"])
tmp[tmp >= 1] = 1
df["listened"] = tmp
```

or (using more modern pandas)

```
df["listened"] = df["listen_count"].apply( lambda x: int(x>1))
```

or (using sklearn)

```
from sklearn.preprocessing import Binarizer
bn = Binarizer(threshold=0.9)
tmp = bn.transform([df["listen_count"]])[0]
df["listened"] = tmp
```

Decomposing Categorical Attributes

Dealing with special categories

Imagine you have a categorical attribute, like `Item_Color` that can be 'Red', 'Blue' or 'Unknown'. All three categories are treated equally. But there are situations in which the 'unknown' should be treated separately — for example, when it arises from missing values in the original data and the instances of missing values is linked to the target.
In this case we need to create an additional binary feature to indicate the special category.

Merging Rare (infrequent) categories

See titles in Titanic dataset.

Decomposing a Date/Time Attribute

- Decomposing date into day, month, year, day of week
- Decomposing time in hour, minute, second.
- Subtracting dates to get age.
- Extracting day of week
- Categorising into workday/weekend — see Kaggle, Bike hire dataset
- Duration between two times — eg harmonic transformation

```
def make_harmonic_features(value, period=24):
    value *= 2 * np.pi / period
    return np.cos(value), np.sin(value)
```

This transformation preserves the distance between points, which is important for algorithms that estimate distance (kNN, SVM, k-means ...)