Machine Learning

Feature Selection & Engineering

Part One

Dr. Sherif Saad

Learning Objectives

Introduce the students to feature engineering

Understand the importance of feature engineering.

Understand the problems feature engineering solve.

Hand-on experience with feature engineering.

Outlines

- Introduction Feature Engineering
- Feature Engineering Techniques
- Feature Engineering Examples

Feature Engineering: The What

A feature is a data item that is likely useful for prediction

The quality and the quantity of the feature affects the performance of the machine learning model.

Given a machine learning problem and a set of raw data for this problem, could we create new features that increase the reliability and the quality of the learning process.

The process of constructing relevant features from the raw data or other features to increase the predictive power of the learning algorithm.

Feature Engineering: The What

The process of using domain knowledge of the data to create features that enhance the accuracy of the machine learning algorithms.

Most of the time feature engineering is a manual process that require domain-specific knowledge.

Feature engineering is an essential part of applying machine learning to solve specific problem.

"Coming up with features is difficult, time-consuming, requires expert knowledge. ***Applied machine learning*** *is basically feature engineering."* [*Andrew Ng*]

Feature Engineering: The What

"Feature engineering is 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." *[Tomasz Malisiewicz]*

Feature Engineering affects:

- What you can do? The problem type (classification, regression, clustering)
- How we evaluate or measure the mode performance?
- The selection of the model. Should we use NN, SVM or DT
- What data you need to collect and the granularity of the data?

Feature Engineering: TiTanic Dataset

Data Dictionary

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

Variable Notes

pclass: A proxy for socio-economic status (SES)
1st = Upper
2nd = Middle

Feature Engineering, Selection and Representation

What is feature Engineering?

feature engineering focuses on adding more information or signal into the feature space.

What is feature Selection?

Selecting features from the feature space that are most relevant to the machine learning problem you are working on

What is Feature Representation?

focuses on finding the best encoding method to store and represent feature in format that fit the machine learning algorithm 1

Feature Engineering, Selection and Representation

We can think of feature selection and representation as subtasks of the feature engineering task.

Feature engineering is not a formally defined process, just a commonly agreed set of tasks related to designing feature sets for ML applications

In general, the process of feature engineering is divided into several subtasks. The description and number of these subtasks sometimes is problem specific.

Feature Engineering: The Why

"some machine learning projects succeed and some fail. What makes the difference? Easily the most important factor is the features used" [Pedro Domingos]

The algorithms we used are very standard for Kagglers. We spent most of our efforts in feature engineering. [Xavier Conort, on winning the Flight Quest challenge on Kaggle]

Feature Engineering: The Why

I usually use other people's code [...] it is usually not "efficient" (from time budget perspective) to write my own algorithm [...] I can find open source code for what I want to do, and my time is much better spent doing research and feature engineering -- Owen Zhang

http://blog.kaggle.com/2015/06/22/profiling-top-kagglers-owen-zhang-currently-1-in-the-world/

kaggle

Feature Engineering: The Why

The features we use directly affect the quality of the learning process and the construction of the predictive model.

Selecting and engineering features is sometimes more important than selecting the optimal model and parameters for the model.

Good feature set allow us to use less complex and inexpensive models.

Better feature design usually results in simpler models.

Feature Engineering: The How

Feature engineering is an iterative process

1. Brainstorm features:

   collect as much data (raw data) as you can. Study features of other similar problems and learn how features were engineered for these problem.

2. Devise | Craft features:

   Apply manual feature construction, or use automatic feature learning (e.g. vector quantization, principle component analysis, deep learning)

Feature Engineering: The How

3. Select Features:

   Use different feature scoring and feature selection methods to select the most relevant feature to the problem you are trying to solve. The output of this stage could be multiple subset of features.

4. Evaluate Models:

   Use the set or subsets of feature from the previous step and measure the model accuracy suing unseen (new) data.

Feature Engineering: Examples

Example 01:

Botnets Detection Through Network Behavior Analysis

Example 02:

Soccer Game Prediction

Example 01: P2P Botnets Detection Through Network Behavior Analysis

Source "Detecting P2P botnets through network behavior analysis and machine learning"

http://ieeexplore.ieee.org/document/5971980/

Example 01: P2P Botnets Detection Through Network Behavior Analysis

Botnets have become one of the major threats on the Internet for serving as a vector for carrying attacks against organizations and committing cybercrimes.

They are used to generate spam, carry out DDOS attacks and click-fraud, and steal sensitive information.

P2P bots, are the newest and most challenging types of botnets currently available.

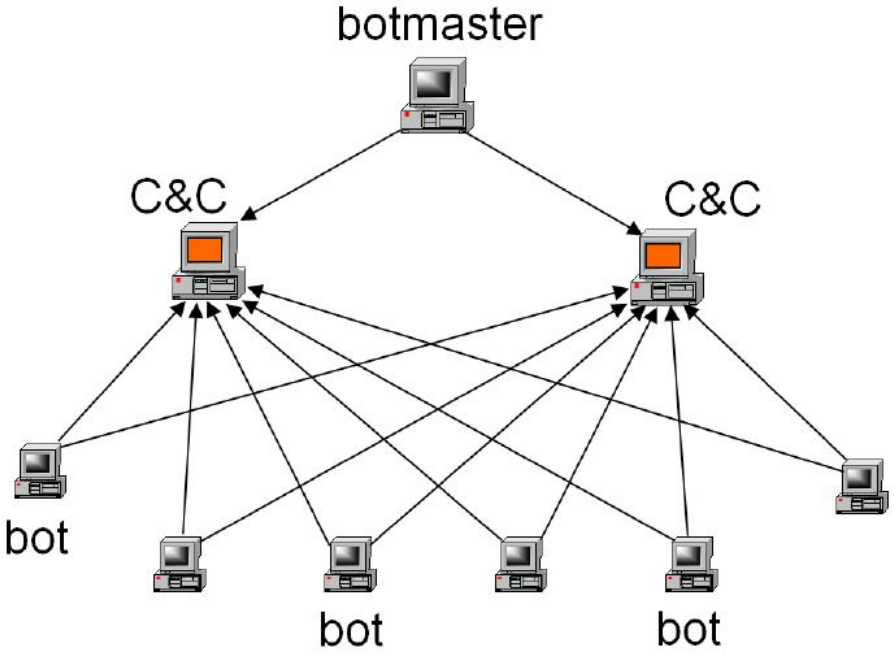
Example 01: P2P Botnets Detection Through Network Behavior Analysis

A botnet is a collection of computers connected to the Internet which have been compromised and are being controlled remotely by an intruder via malicious software called bots.

Bots connect over TCP and UDP networks and some uses IRC or HTTP

We can divided the botnet lifecycle into four phases, namely, formation, C&C, attack, and post-attack phases

Example 01: P2P Botnets Detection Through Network Behavior Analysis



Not a P2P Botnet

Example 01: P2P Botnets Detection Through Network Behavior Analysis

Brainstorm Features & Crafting Features

Works with your group

Example 02: Soccer Game Prediction

A training data set containing the results of over 200,000 soccer matches. Challenge participants should use this data set to construct a model that predicts the outcome of future soccer matches contained in the prediction data set.

Example 02: Soccer Game Prediction

In particular, there are two main issues to address. First, how should one derive predictive features from the data? There is no obvious way of doing this – many approaches are conceivable.

Second, the data consists of matches from different soccer leagues around the world. How can these data be combined to maximize the size of the training data that can be used to construct a model? Again, there may be many ways of doing this.

Example 02: Soccer Game Prediction

	Sea	Lge	Date	HT	AT	HS	AS	GD	WDL
1	00-01	GER1	11/08/2000	Dortmund	Hansa Rostock	1	0	1	W
2	00-01	GER1	12/08/2000	Bayern Munich	Hertha Berlin	4	1	3	W
3	00-01	GER1	12/08/2000	Freiburg	VfB Stuttgart	4	0	4	W
4	00-01	GER1	12/08/2000	Hamburger SV	Munich 1860	2	2	0	D
5	00-01	GER1	12/08/2000	Kaiserslautern	Bochum	0	1	-1	L
6	00-01	GER1	12/08/2000	Leverkusen	Wolfsburg	2	0	2	W
...									
205177	16-17	FIN1	23/10/2016	IFK Mariehamn	Ilves Tampere	2	1	1	W
205178	16-17	FIN1	23/10/2016	Vaasan PS	FC Lahti	0	1	-1	L
205179	16-17	FIN1	23/10/2016	Kuopion PS	Kemi Kings	1	0	1	W
205180	16-17	FIN1	23/10/2016	Rovaniemi PS	Helsingfors IFK	0	0	0	D
205181	16-17	FIN1	23/10/2016	PK35 Vantaa	Inter Turku	2	1	1	W
205182	16-17	FIN1	23/10/2016	HJK Helsinki	Seinajoki	0	0	0	D

Example 02: Soccer Game Prediction

Brainstorm Features & Crafting Features

Works with your group

Questions