Class 4

Data pre-processing in Machine Learning

----Overview of Assignment #1 Performance----

Assignment #2:

-Originally due today

-Now due two weeks from now: March 1

-Updated Assignment #2 will be available Monday

-Topic: applied model selection

-Next week’s class: applied model selection

Assignment #3: Dropped/Removed

Papers: Now due March 18 by Midnight

Second half: follow original workload.

Class 4: Data pre-processing

The elimination of noise instances
is one of the most difficult problems in
inductive ML

```
{
  "metadata": {
    "kernelspec": {
      "display_name": "Python 3",
      "language": "python",
      "name": "python3"
    },
    "language_info": {
      "file_extension": ".py",
"cells": [ {
      "metadata": {
,
      "source": [
        "# Data Preprocessing\n",
        "# Importing the libraries\n",
        "import numpy as np\n",
        "import matplotlib.pyplot as plt\n",
        "import pandas as pd"
      ]
],
  "nbformat": 4,
  "nbformat_minor": 1
```


Kaggle

(Intro to / Elements of Statistical Learning)

R for researchers:

- Importing data
- Object types and structures
- Missing data
- Changing types and creating new variables

Table 1
Data preprocessing techniques used by ML methods for SCE.

Source	Reference	Methods	FS[a]/CS[b]	Scaling[c]	MDT[d]
IEEE TSE	Pendharkar et al. [50]	ANN, BBN, CART	FS		
	Auer et al. [51]	CBR		[0,1]	
	Keung et al. [23]	CBR	FS, CS	[0,1]	LD
	Kocaguneli et al. [11]	CBR	FS, CS		
	Kocaguneli et al. [42]	CART, SVM	FS	[0,1]	MI
	Kocaguneli et al. [65]	CBR, CART	FS	[0,1]	MI
	Menzies et al. [52]	CBR	FS		
	Mittas and Angelis [66]	ANN, CART, CBR	FS		LD
JSS	Chiu and Huang [14]	ANN, CART, CBR		[0,1]	LD
	de Barcelos Tronto et al. [8]	ANN		[0,1]	
	Vinay Kumar et al. [24]	ANN	FS	[0,1]	
	Li et al. [25]	ANN, CART, CBR, SVM	CS	[0,1]	LD
	Azzeh et al. [12]	CBR		[0,1]	LD
	Bou et al. [4]	ANN	FS, CS		
IST	Huang and Chiu [15]	ANN, CART, CBR	FS	[0,1]	LD
	Mittas et al. [45]	ANN		[0,1]	
	Oliveira et al. [17]	SVR, GP	FS	[0,1]	
	Minku and Yao [62]	ANN, CART, CBR	FS, CS	[0,1]	MI
ESE	Li et al. [35]	CBR	FS	[0,1]	LD
	Li and Ruhe [16]	CBR	FS		LD
	Azzeh et al. [33]	CBR, ANN	FS		LD
	Li et al. [55]	ANN	FS	[0,1]	
	Mittas and Angelis [34]	CBR	FS, CS	[0,1]	LD
	Lopez-Martin et al. [67]	ANN	CS		
	Corazza et al. [37]	SVR	FS		
	Azzeh [63]	CBR	FS, CS	[0,1]	LD
	Corazza et al. [60]	CBR, SVR	FS		LD
	Kocaguneli et al. [13]	CBR	FS	[0,1]	MI
	Seo and Bae [30]	CBR	FS, CS	[−1,1]	LD
SQJ	Liu et al. [58]	CART, CBR	FS, CS		LD
	Bakır et al. [68]	SVM, CBR	FS	[0,1]	LD
	Hsu and Huang [64]	ANN, CART, CBR	FS, CS	[0,1]	LD
	Bakır et al. [69]	CART, CBR	FS	[0,1]	LD
	Khatibi et al. [59]	CBR	FS		LD
ICSE	Ramasubbu and Balan [70]	CBR	CS		
ESEM	Li et al. [46]	CBR			MI
	Mendes [47]	BBN, CBR, CART	FS		
	Keung [48]	CBR	FS, CS		
	Corazza et al. [44]	SVR, CBR, BBN	CS		

1. Data Mining

2. Feature Selection

3. Data Cleaning

Short List of Failed Credit Metrics

	As Submitted	Post BI&M
General Considerations		
Clear Ownership - State	Medium Low	Medium High
Clear Ownership - Private	Very Low	Medium
Line Items		
Appendix B: Scoring Model		
EBITDA	Low	Medium
Interest Expenses (Gross)	Medium Low	Medium
Total Debt	Low	Medium Low
Total Shareholder Equity	Low	Medium Low
Total Assets	Medium Low	Medium
Operating Lease Debt Equivalent	Medium Low	Medium Low
Current Ratio	Medium Low	Medium Low
Liquidity POV	Medium Low	Medium Low
Appendix G: Risk Ratings Criteria		
Debt to Capital Ratio	Low	Medium Low
Cash Flow Estimates	Low	Medium Low
Appendix X:		
Clear recourse to collateral (Art. 5)	Never	Never
Appendix A,C,D,E,F,L	N/A	N/A

Existing data +
Business goals
.... =

Feature equivalence
Feature acquisition

(Realistic/Minimize)

Changes to Credit Applications

	Off-Takers
Credit Policy & Overall Strength	
Number of years in business	✔
Concentration (counterparty > 25% of sales)	✔
Credit policy & controls (number of words)	✔
Record keeping quality (number of words)	✔
Number of accounts	✔
Stand alone risk manager/credit officer	✔
Number of risk or credit analysts	✔
Number of years for oldest account	✔
Transaction Terms	
For resale in local markets or export	✔
Purchase agreement $ rate > or = to Market Value	✔
Purchase agreement subject to price controls	✔
Purchase agreement force majeure clause strength	✔

Changes to Credit Metrics

	Off-Takers	Owners	Buyers
Monitoring & Early Warnings			
Country, industry, and counterparty risk	✔	✔	✔
Fraud, FCPA, and mismanagement	✔	✔	✔
Enriched Internal Data			
Application form	✔		
Application file	✔		
Proximity Analysis			
Closeness to other Noble counterparties	✔	✔	✔
Counterparty bank domestic/international	✔	✔	✔
Auditor domestic/international	✔	✔	✔
Law firm domestic/international	✔	✔	✔
Counterparty Risk			
Overall perceived strength	✔	✔	✔
MoU for GOE arrears & repayment	✔		
Breach of covenants or legal actions	✔	✔	✔
New Classification & Ratings			
Legal entity type & type rating	✔	✔	✔
Sub-sector of operations & rating	✔	✔	✔
Sub-sector subject to price controls	✔	✔	✔

1. **Data Mining:**

----How did you extract the data?
----How did you reconstruct / combine files
----Time Series?

2. **Feature Selection:**

----How did you make sense of rows/columns
----How did you deduct/identify relationships

3. **Data Cleaning:**

----Manual or Algorithmic?

Data Understanding (1)

Indent	Expenditure Category	Rel Importance	Index.....
0	All items	100.000	242.839
1	Food	13.384	248.242
2	Food at home	7.382	237.365
3	Cereals and bakery products	0.964	272.922
3	Meats, poultry, fish, and eggs	1.635	242.596
3	Dairy and related products	0.744	219.804
3	Fruits and vegetables	1.302	291.679
3	Nonalcoholic beverages and beverage materials	0.873	167.074
3	Other food at home	1.864	208.804
2	Food away from home(1)	6.002	266.079
1	Energy	7.513	199.608
2	Energy commodities	4.094	211.110
3	Fuel oil	0.109	243.347
3	Motor fuel	3.908	207.280
4	Gasoline (all types)	3.823	206.360
2	Energy services(2)	3.419	197.767
3	Electricity(2)	2.628	205.230
3	Utility (piped) gas service(2)	0.791	172.319

---What is the relationship btw levels?
---Which ones are the Predictors and Observations?
---How many features are optimal?

Data Understanding (2)

Unadjusted indexes			Unadjusted percent change		Seasonally adjusted percent change		
Jan. 2017	Dec. 2017	Jan. 2018	Jan. 2017-Jan. 2018	Dec. 2017-Jan. 2018	Oct. 2017-Nov. 2017	Nov. 2017-Dec. 2017	Dec. 2017-Jan. 2018

----- Where is the time series?

How did you match expected/assigned ML methods with data features/dataset construction?

Instance selection approaches are distinguished between filter and wrapper.

Filter evaluation only considers data reduction but does not take into account activities.

On contrary, wrapper approaches explicitly emphasize the ML aspect and evaluate results by using the specific ML algorithm to trigger instance selection.

Filter Selection:

Examples of “illegal features”:
---min/max outside of range
---variance higher than 3fold

Misspellings / Does not match

Duplication

Missing Features Values

(i) a value is missing because it was forgotten or lost;

(ii) a certain feature is not applicable for a given instance, e.g., it does not exist for a given instance;

(iii) for a given observation, the designer of a training set does not care about the value of a certain feature (so-called don’t-care value).

Missing Data Methods (1)

Method of Ignoring Instances with Unknown Feature Values: This method is the simplest: just ignore the instances, which have at least one unknown feature value.

Most Common Feature Value: The value of the feature that occurs most often is selected to be the value for all the unknown values of the feature.

Concept Most Common Feature Value: This time the value of the feature, which occurs the most common within the same class is selected to be the value for all the unknown values of the feature.

Missing Data Methods (2)

Mean substitution: Substitute a feature's mean value computed from available cases to fill in missing data values on the remaining cases.

Regression or classification methods: Develop a regression or classification model based on complete case data for a given feature, treating it as the outcome and using all other relevant features as predictors.

Hot deck imputation: Identify the most similar case to the case with a missing value and substitute the most similar case's Y value for the missing case's Y value.

Method of Treating Missing Feature Values as Special Values: treating "unknown" itself as a new value for the features that contain missing values.

DATA DISCRETIZATION

Discretization should significantly reduce the number of possible values of the continuous feature

The simplest discretization method is an unsupervised direct method named equal size discretization. It calculates the maximum and the minimum for the feature that is being discretized and partitions the range observed into k equal sized intervals.

DATA NORMALIZATION

Normalization is a "scaling down" transformation of the features. Within a feature there is often a large difference between the maximum and minimum values, e.g. 0.01 and 1000.

When normalization is performed the value magnitudes and scaled to appreciably low values.

FEATURE SELECTION (1)

Feature subset selection is the process of identifying and removing as much irrelevant and redundant features as possible

This reduces the dimensionality of the
data and enables learning algorithms to operate faster and more effectively

Features Selection (2)

Relevant: These are features have an influence on the output and their role can not be assumed by the rest

Irrelevant: Irrelevant features are defined as those features not having any influence on the output, and whose values are generated at random for each example.

Redundant: A redundancy exists whenever a feature can take the role of another (perhaps the simplest way to model redundancy).

FEATURE CONSTRUCTION (1)

The problem of feature interaction can be also addressed by constructing new features from the basic feature set. This technique is called feature construction/transformation.

The new generated features may lead to the creation of more concise and accurate classifiers.

The discovery of meaningful features contributes to comprehensibility of produced classifier + better understanding of the learned concept

Feature Construction (2) – EX:

The GALA algorithm [19] performs feature construction throughout the course of building a decision tree classifier.

Feature transformation process can also extract a set of new features from the original features through some functional mapping

Group Paper Discussion

3 weeks of class + Spring Break = 4 Weeks

Schedule:

Week of Feb 19: Finish data acquisition, understanding, and pre-processing

Week of Feb 26: Feature selection, data preparation, and model selection

Week of March 5: Model Implementation

Week of March 12: Draft of paper

Spring Break: Buffer time