Data-preprocessing

Advanced Machine Learning
École Polytechnique Fédérale de Lausanne, Switzerland

EPFL

March 20, 2020

Learning Outcomes

How to deal with:

- Missing values
- Categorical data
- Text datasets
- Unbalanced datasets

Missing data

Values may be missing for various reasons:

- sensor malfunction
- expensive data-gathering
- information non entered (formular)

Usually denoted by: NaN, or ?

Example the SECOM datatset.

Missing data

- Replace with the mean/most frequent value
- Approximate with regression/classification
- Interpolate in case of time-series

Missing data

Replacing with the mean/most frequent value

	Car width	Manufacturer
car 1	2.3.	?
car 2	?	Volkswagen
car 3	1.7	Volkswagen
car 4	1.8	Volkswagen
car 5	2	BMW
car 6	2.4	BMW
car 7	2.1	BMW
car 8	1.8	Volkswagen

Table: A toy dataset with missing values

Continuous data $\rightarrow$ replace with mean $\rightarrow$ width of car 2 is 2.01
Categorical data $\rightarrow$ replace with the most frequent label $\rightarrow$ Manufacturer of car 1 is Volskwagen

Missing data

Approximate missing values by performing regression/classification

$$y = f(x)$$

- y is the dimension with missing data and
- x all the other dimensions.
- x contain only samples without missing data
- perform cross-validation as usual (find hyperparameters)

Missing data

Approximate missing values by performing regression/classification

	Car width	Manufacturer
car 1	2.3.	?
car 2	?	Volkswagen
car 3	1.7	Volkswagen
car 4	1.8	Volkswagen
car 5	2	BMW
car 6	2.4	BMW
car 7	2.1	BMW
car 8	1.8	Volkswagen

utilise ces données

Table: A toy dataset with missing values

Regression for the car width
Classification for the manufacturer

Missing data – Expectation maximization[1]

Model data as a mixture models

- Gaussian for continuous
- Bernouli for discrete

Idea: Handle missing values similarly to unknown model parameters

- Joint distribution:

unknown value

$$p(X|\theta) = p\left(X_o, X_m|\theta\right) = p\left(X_o|\theta\right) p\left(X_m|X_o, \theta\right) \tag{1}$$

observed value

- Log-likelihood:

$$\begin{aligned} L(\theta|X) &= L\left(\theta|X_o, X_m\right) \\ &= L\left(\theta|X_o\right) + \log P\left(X_m|X_o, \theta\right) \end{aligned} \tag{2}$$

[1] Ghahramani, Zoubin, and Michael I. Jordan. "Supervised learning from incomplete data via an EM approach." Advances in neural information processing systems. 1994.

Categorical values

A dimension of the data may take categorical values

Example: Car manufactures at the Automobile Data Set

- Convert categorical to numeric using one hot encoding.

Create a new dimension for each of the categorical values.

Assign binary values to those dimensions according to the occurrence of the label

Categorical values

One hot encoding

- Create a new dimension for each of the categorical values.
- Assign binary values to those dimensions according to the occurrence of the label

fct fait ça en python/matlab

	Car width	Manufacturer
car 1	2.3.	BMW
car 2	1.75	Volkswagen
car 3	1.7	Volkswagen
car 4	1.8	Volkswagen
car 5	2	BMW
car 6	2.4	BMW
car 7	2.1	BMW
car 8	1.8	Volkswagen

Table: Before one hot encoding

	Car width	Volkswagen	BMW
car 1	2.3.	0	1
car 2	1.5	1	0
car 3	1.7	1	0
car 4	1.8	1	0
car 5	2	0	1
car 6	2.4	0	1
car 7	2.1	0	1
car 8	1.8	1	0

Table: After one hot encoding

Text Datasets

Each sample is text which needs to be classified

Example: The BBC dataset

Bag-of-words: Each distinct word is a dimension of the sample. The value of each dimension corresponds to word counts at each text (sample)

Example:

Sample 1 : *Camera phones are 'must-haves'*

Sample 2 : *Musical future for phones*

	camera	phones	are	must-haves	musical	future	for
sample 1	1	1	1	1	0	0	0
sample 2	0	1	0	0	1	1	1

Table: After one hot encoding

Unbalanced datasets

Unbalanced datasets: One class contains significantly less samples than others (example: the Exoplanet detection dataset)

- Report significant metrics (classification error vs confusion matrix)
- Down-sampling of dominant classes
- Create artificial samples for non-dominant classes

Unbalanced Datasets – Reporting significant metrics

Confusion matrices provide information on which classes are merged (confused) with which by a classifier

Actual Class \ Predicted Class	C_1	C_2	$\cdots$	C_c
C_1	n_{11}	n_{12}	$n_{1\ldots}$	n_{1c}
C_2	n_{21}	n_{22}	$n_{2\ldots}$	n_{2c}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
C_c	n_{c1}	n_{c2}	$\cdots$	n_{cc}

Table: Structure of a confusion matrix

$n_{ij} \rightarrow$ Number of samples that belong to class i and classified at class j

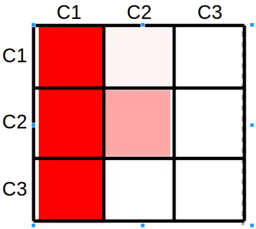
Unbalanced Datasets – Reporting significant metrics

Classification error vs Confusion matrix

Comparison of two classifiers (kNN - GMM) on an unbalanced dataset:

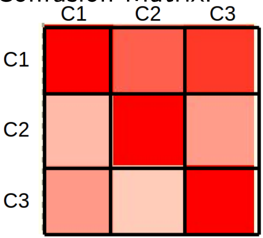
kNN:

- Classification Error: 17.4 %
- Confusion Matrix:



GMM:

- Classification Error: 55 %
- Confusion Matrix:



Unbalanced Datasets – Down-sampling

Match the samples' number of the classes by down-sampling the most dominant class/classes

1. Partition to training/testing dataset
2. At the training set down-sample the most dominant classes:
   - Select randomly samples and remove them (uniform proba)
3. Train the classifier with the down-sampled dataset.
4. Evaluate on the testing set
5. Repeat n times (n-fold validation)

Drawbacks:

- Ending up with very few samples for training
- Downsampled dataset does not capture reliably the distribution of dominant classes.

Unbalanced Datasets – Oversampling

Create artificial samples of the non-dominant classes

1. Partition to training/testing dataset
2. At the training set over-sample the non-dominant classes:
   - Approximate the distribution of a class with GMM
   - Sample the required amound of data from the GMM
3. Train the classifier with the over-sampled dataset.
4. Evaluate on the testing set
5. Repeat n times (n-fold validation)

Drawbacks:

- Computational complexity

Summary – Data preprocessing

- Missing values:
  - Replace with mean (continous) or most frequent value (categorical)
  - Approximate them with regression/classification
  - Expectation-Maximization
- Categorical values:
  - One hot encoding
- Text datasets:
  - Bag of words
- Unbalanced datashets:
  - Importance of performance metric
  - Undersampling
  - Oversampling

Supplementary material can be found at moodle