Buscando la paz interior...

twiter: @richardyantas5

TITLE:

Introduction to Data Science Competition

fuente: www.prometec.net

Versión en LaTeX:
Msc(c). Richard Valentín Yantas Alcantaraaa

Contents

Chapter 1

Introduction

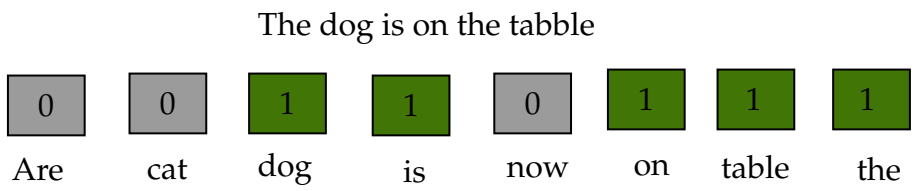
1.1 Introductionnn

1.1.1 Feature extractionnnñ from texts and image

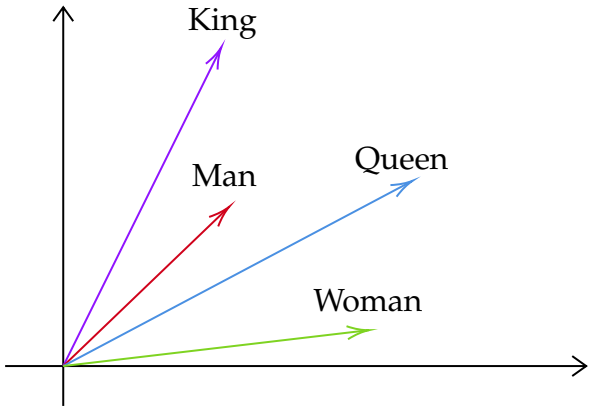
This is the titanic data sets

Text to vector

1. Bag of words:



2. Embeggins:



Text preprocessing

1. Lowercase
2. Lemmatization
3. Stemming

4. StopWords

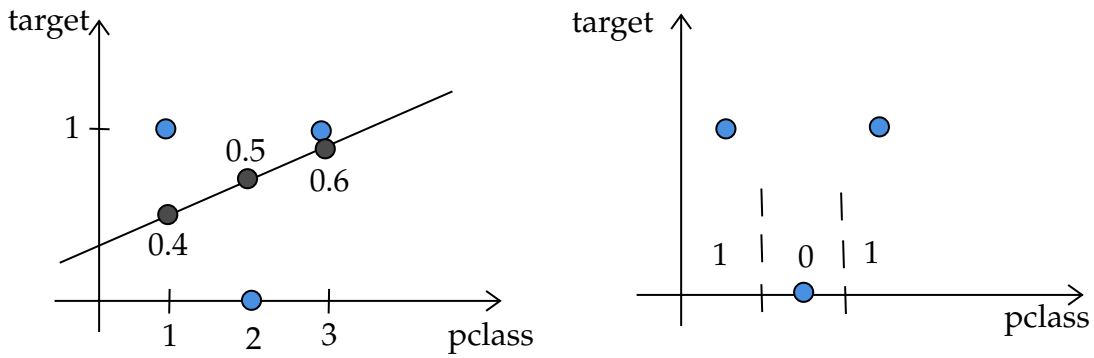
1.1.2 Categorical and ordinal features

Categorical

Ordinal Features

Label enconding

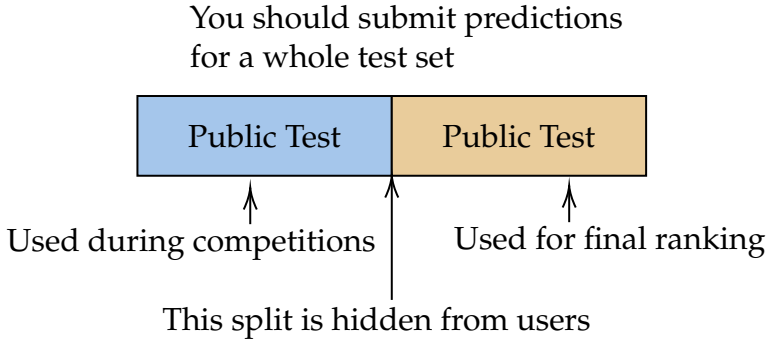
pclass	1	2	3
target	1	0	1



Frequency enconding

1.1.3 Competition Mechanics

1. Evaluation function Exists some evaluation functions like:
    - Accuracy
    - Logistic loss
    - AUC
    - RMSE
    - MAE
2. Public/Private tests



3. Sites:

    - Kaggle
    - Driven data
    - CrowdAaqnalityx
    - CodaLab
    - DataScienceChallenge.net
    - DataScience.net
    - Single-Competition sites like KDD,VizDooM

4. Conclusions:

    - Main Concepts(Data, Model, Submission, Evaluation, Leaderboard)
    - Competition platforms
    - Reasons for participating

1.1.4 Datetime and Coordinates

1.1.5 Handling missing

1.1.6 Handling missing

1.1.7 Kaglle overview

1.1.8 Numeric features

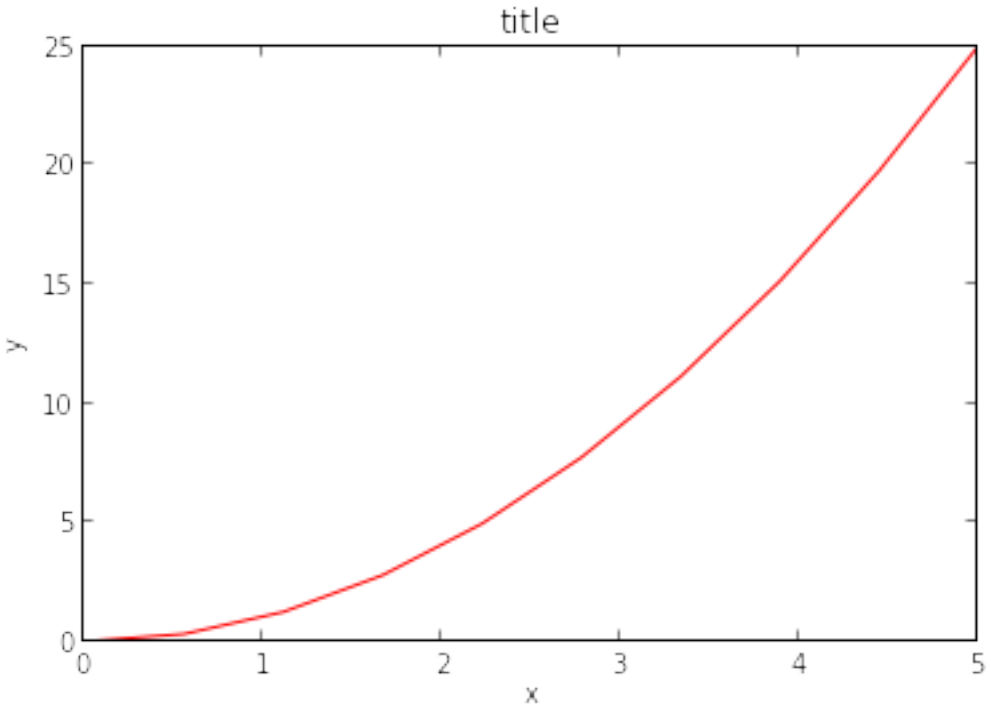
1.1.9 Recap of main algorithms

```
In [3]: %matplotlib inline
```

```
In [4]: from pylab import *
```

```
In [5]: x = linspace(0, 5, 10)
        y = x ** 2
```

```
In [6]: figure()
        plot(x, y, 'r')
        xlabel('x')
        ylabel('y')
        title('title')
        show()
```


title
25
20
15
10
5
0
y
0
1
2
3
4
5
x

Chapter 2

Exploratory Data Analysis

2.1 Exploratory Data Analysis

Nowadays, most ecological research is donde with hypothesis testing and modelling in mind. However , Exploratory Data Analysis(EDA), which uses visualization tools and computer synthetic descriptors, is still required at the beginning of the statistical analysis of multidimensional data, in order to:

- Get an overview of the data
- Transform or recode some variables
- Orient further analyses

2.1.1 Building Intuition about the data

1. Getting domain knowledge It helps to deeper understand the problem
2. CHecking if the data is intuitive To be agree with domain knowledge
3. Understanding how the data was generated A it is crucial to set up a poper Validation

2.1.2 Data cleaning

1. Constant features
2. Duplicated features
3. Duplicated rows
4. Check if datasets is shuffled

here graphics!! :D

2.1.3 Data splitting strategy

1.

2.1.4 Exploratory data Analysis

2.1.5 Exploratory anonymized data

2.1.6 Validation and overfitting

2.1.7 Validation strategies

2.1.8 Visualizations