Fake Names Detection

Project: machine learning recognition of fake names

Marco Zanotti

–/04/2019

Process Analysis

- processo di acquisto FlyUvet
- valutazione casistiche di inserimento (solo upper case, nome e cognomi separati, lunghezza stringhe, ecc)
- database
- valutazione operatività modello (dove inserire il filtro ed efficiency)

Data Acquisition

- nomi reali da database pagato
- nomi falsi da database fake?
- explorative data analysis (max/min lengths, spaces, punctuations, patterns, ecc) on real and fake strings/names
- analisi provenienza real names
- valutazione sbilanciamento (over-sampling / under-sampling / mixed)
- valutazione download dati anagrafici pubblici

Synthetic Data Creation

- creazione nomi veri in base alla distribuzione dei dati e ai pattern e non
- creazione nomi falsi (generazione di stringhe casuali in base ai pattern e non)

Variables Extraction & Labelling

- manipolazione dei nomi e creazione del dataset con tutte le variabili (lunghezza, num. vocali/consonanti, proporzione vocali/consonanti, vocali/consonanti consecutive, ecc.)
- etichettatura dei dati

Modelling

- study unbalanced case (balancing data or use models that allow for unbalanced data)
- train, try different models: SVM and Advanced SVM, KNN, Logistic Regression, Bayesian, Tree, Random Forest, XGBoost, extra trees, regularized greedy forest, Decision Jungle, Neural Network, Deep Learning (keras/tensorflow)
- cross-validation interna
- cross-validation esterna
- test
- evaluation (accuracy, precision, recall, F1, AUC/ROC, other metrics)

Reporting

- Shiny web app testing names (input -> nome, output -> TRUE/FALSE + Prob)

Production