Damegender: using NLTK and Scikit in a real case

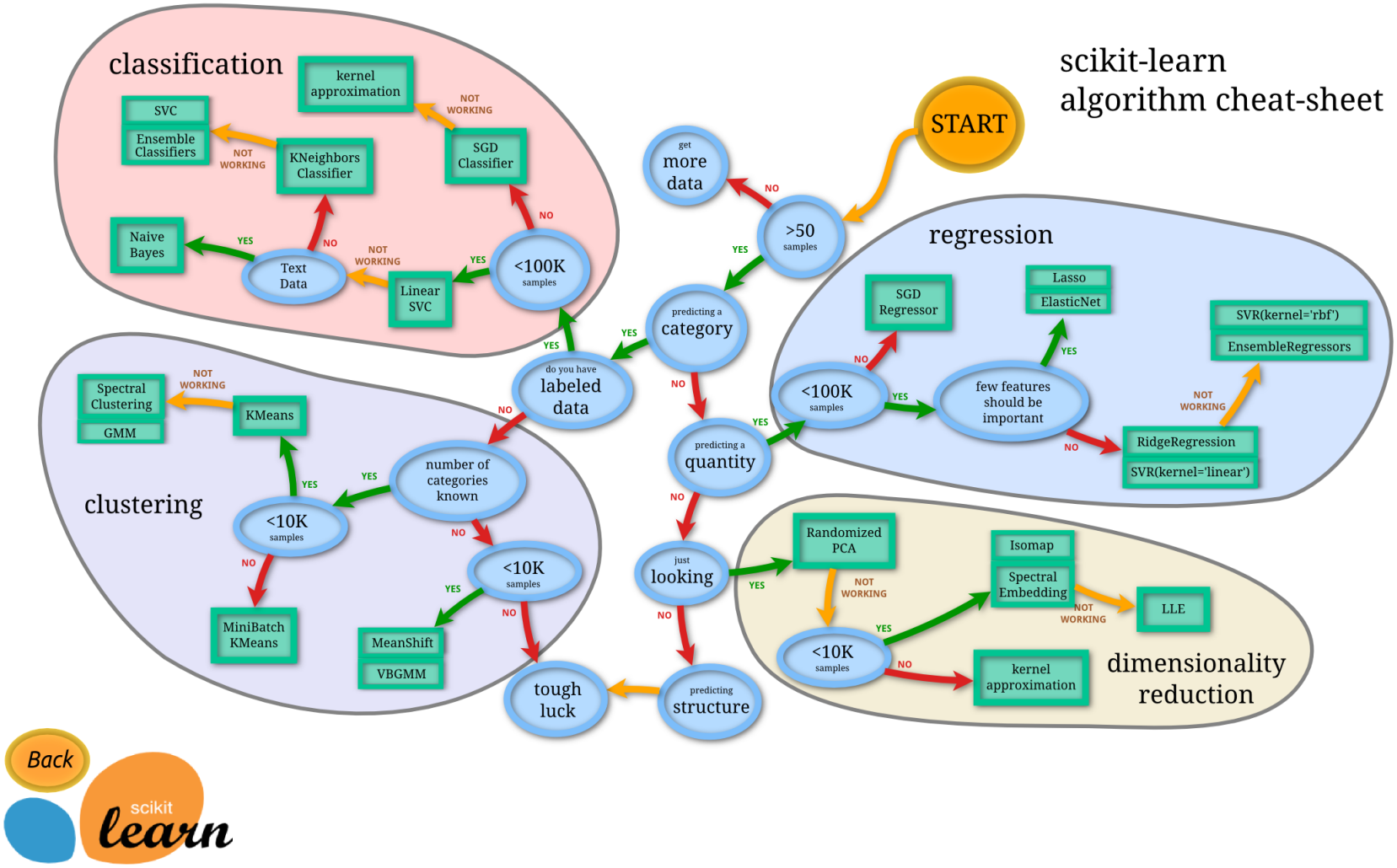
David Arroyo Menéndez

May 16, 2019

Machine Learning is for predictions

- Quantities
- Categories (with data training or not)
- What properties is predicting

Scikit Graph



In classification our data are properties and a variable for the prediction

	Predictors				Response
	Outlook	**Temperature**	**Humidity**	**Wind**	**Class** **Play=Yes** **Play=No**
Day1	Sunny	Hot	High	Weak	No
Day2	Sunny	Hot	High	Strong	No
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	No
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	No
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	No

Practical NLTK

```
$ git clone https://github.com/davidam/GAPLEN.git
$ sudo pip3 install GAPLEN
```

In NLP the machine learning is a classification task:

1. Sentiment Analysis
2. Detect Gender
3. Sentence Similarity
4. Text Summary
5. Classify Documents
6. Manage Words

singulars/plurals, dictionary entries, stopwords

1. Gramatical Trees
2. Extract Keywords
3. Disambiguation

Now Damegender

Name

damegender is a gender detection tool coded by David Arroyo MEnéndez (DAME)

Why?

- If you want determine gender gap in free software projects or mailing lists.
- If you don't know the gender about a name in spanish or english (current status).
- If you want research with statistics about why a name is related with males or females.
- If you want use the main solutions in gender detection (genderize, genderapi, namsor, nameapi and gender guesser) from a command.

DAMe Gender is for you!

Installing Software

Posible Debian/Ubuntu dependencies

```
$ sudo apt-get install python3-nose-exclude dict dict-freedict-
```

From sources

```
$ git clone https://github.com/davidam/damegender
$ cd damegender
$ pip3 install -r requirements.txt
```

With python package

```
$ python3 -m venv /tmp/d
$ cd /tmp/d
$ source bin/activate
$ pip install --upgrade pip
$ pip3 install damegender
$ cd lib/python3.5/site-packages/damegender
$ python3 main.py David
```

Obtaining an api key

Currently you can need an api key from:

- `https://store.genderize.io/documentation`
- `https://gender-api.com`
- `https://www.nameapi.org/`

You can execute:

```
$ python3 apikeyadd.py
```

To configure your api key

Configuring nltk

```
$ python3
>>> import nltk
>>> nltk.download('names')
```

Checking All tests

```
$ nosetest3 test
```

Checking Single test

```
$ nosetests3 test/test_dame_sexmachine.py:TddInPythonExample.te
```

Execute program

Detect gender from a name (INE is the dataset used by default)

$ python3 main.py David
David gender is male 363559 males for David from INE.es 0 females for David from INE.es

Detect gender from a name from multiple dataset

$ python3 main.py David –total="all"
David gender is male 375099 males and 9 females from all census (INE + Uk census + USA census)

Detect gender from a name only using machine learning (experimental way)

$ python3 main.py Mesa –ml=nltk
Mesa gender is female 0 males for Mesa from INE.es 0 females for Mesa from INE.es

Detecting gender in mailing lists and repositories

Count gender from a git repository

$ python3 git2gender.py
`https://github.com/chaoss/grimoirelab-perceval.git`
–directory="/tmp/clonedir"
The number of males sending commits is 15 The number of females sending commits is 7

Count gender from a mailing list

$ cd files/mbox
$ wget -c `http://mail-archives.apache.org/mod_mbox/httpd-announce/201706.mbox`
$ cd ..
$ python3 mail2gender.py
`http://mail-archives.apache.org/mod_mbox/httpd-announce/`

Using external tools to detect gender

Use an api to detect the gender

$ python3 api2gender.py Leticia –surname="Martin" –api=namsor
female scale: 0.99

Google popularity for a name

$ python3 gendergoogle.py Leticia
Google results of Leticia as male: 42300 Google results of Leticia as female: 63400

Give me informative features

$ python3 infofeatures.py
Females with last letter a: 0.4705246078961601
Males with last letter a: 0.04867256637168141 5
Females with last letter consonant: 0.2735841767750908
Males with last letter consonant: 0.6355328972681801
Females with last letter vocal: 0.726261299544 1552
Males with last letter vocal: 0.3640823393612928

To measure success

Damengeder

$ python3 accuracy.py –csv=files/names/min.csv
Gender list: [1, 1, 1, 1, 2, 1, 0, 0] Guess list: [1, 1, 1, 1, 0, 1, 0, 0] Dame
Gender accuracy: 0.875

Genderize

$ python3 accuracy.py –api="genderize" –csv=files/names/min.csv
Gender list: [1, 1, 1, 1, 2, 1, 0, 0] Guess list: [1, 1, 1, 1, 2, 1, 0, 0]
Genderize accuracy: 1

Confusion

Namsor

```
$ python3 confusion.py –api="namsor"
[[ 2, 0, 0]
[ 0, 5, 0]]
```

Benchmark

Table 1 **Comparison table showing relevant features for the gender inference services under study.** Note that although Gender API does provide a specific API end point for handling surnames, our results employ the version that does not make use of them.

	Gender API	gender-guesser	genderize.io	NameAPI	NamSor
Database size (January 2018)	1,877,787	45,376	216,286	510,000	1,300,000
Regular data updates	yes	no	yes	yes	yes
Handles unstructured full name strings	yes	no	no	yes	no
Handles surnames	yes	no	no	yes	yes
Handles non-Latin alphabets	partially	no	partially	yes	yes
Implicit geo-localization	no	no	no	yes	yes
Assignment type	probabilistic	binary	probabilistic	probabilistic	probabilistic
Free parameters	accuracy, samples	–	probability, count	confidence	scale
Open source	no	yes	no	no	no
API	yes	no	yes	yes	yes
Monthly free requests	500	unlimited	30,000	10,000	1,000
Monthly subscription cost (100,000 requests/month)	79 €	Free	7 €	150 €	80 €
Provider	Gender-API.com	Israel Saeta Pérez	Casper Strømgren	Optimaize GmbH	NamSor SAS

Damegender Market Study

	damegender	gender api
Database size	60000	1877787
Regular data updates	yes, developing	yes
Handles unstructured full name strings	yes	yes
Handles surnames	yes	yes
Handles non-Latin alphabets	no	partially
Implicit geo-localization	no	no
Assingment type	binary	probabilistic
Free parameters	-	accuracy, samples
Free license	yes	no
API	future	yes
Monthly free requests	free license	500

Accuracies

	Accuracy
Namsor	0.7539570378745054
Genderize	0.715375918598078
Gender Guesser	0.6902204635387225
Dame Gender	0.667750141322781 2

Machine Learning Algorithms and DameGender

These results are experimental, we are improving the choosing of features.

- Stochastic Gradient Descendent accuracy: 0.5873374788015828
- Support Vector Machines accuracy: 0.7049180327868853
- Gaussian Naive Bayes accuracy: 0.5960994912379876
- Multinomial Naive Bayes accuracy: 0.5876201243640475
- Bernoulli Naive Bayes accuracy: 0.5962408140192199
- Dame Gender (nltk bayes) accuracy: 0.6677501413227812

License