Data Mining

Optimització, Preprocés i *IBL* aplicats amb scikit-learn

http://www.salle.url.edu

Enginyeria La Salle – Universitat Ramon Llull

Pràctica 2 (Week 8): Aprendre a categoritzar imatges de dígits per IBL

Objectiu

Els principals objectius d'aquesta pràctica son:

- Aprendre a dominar el paquet de Python scikit-learn
- Agafar fluïdesa amb la selecció, normalització, preprocés i cerca d'atributs
- Conèixer les diferents implementacions de l'algorisme d'aprenentatge basat amb exemples IBL
- Treballar amb el concepte d'optimització d'un *learner*.

Requirements

La pràctica es pot fer en qualsevol sistema operatiu: Windows / Mac OS X / Linux però es necessita d'un intèrpret de Python. Si no esteu acostumats a l'entorn Python adreceu-vos a la guia penjada a l'e-study:
https://estudy.salle.url.edu/mod/resource/view.php?id=463912

Els recursos per fer la pràctica es troben a:
https://estudy.salle.url.edu/mod/resource/view.php?id=463947

El pou es troba a:
https://estudy.salle.url.edu/mod/assign/view.php?id=463948

Si teniu algun problema, comenceu una discussió a :
https://estudy.salle.url.edu/mod/forum/view.php?id=463938

També recomano l'ús d'un editor avançat de programació tals com:

Sublime Text Editor : `http://www.sublimetext.com/`

Notepad++ : `http://notepad-plus-plus.org/`

Temps estimat: 2 hores. Màxim 3 hores

Deadline: 14 de Desembre a les 23:59:59 CET

Descripció

La pràctica proporciona els següents recursos:

- **Assignment2.pdf:** Enunciat de la pràctica
- **Recordeu que a través de pip o easy-install s'instalen els paquests numpy, scipy i scikit-learn**

Entrega:

- **Es demana un document en format PDF donant resposta a les qüestions plantejades en aquesta pràctica, acompanyat del codi Python emprat per la mateixa**

Funcionament:

1. **Analitzar el conjunt de dades de dígits (Digits Data Set)**
   - *El primer pas és importar les llibreries de python requerides per la pràctica. Es recomanen carregar les següents llibreries:*

```
import numpy
import sklearn
import sklearn.datasets
import sklearn.model_selection
import sklearn.decomposition
import sklearn.neighbors
import sklearn.metrics
```

   - *Ara ja podeu carregar el dataset de digits a les varibles X i Y:*

```
digits = sklearn.datasets.load_digits()
X= digits.data
Y= digits.target
```

```
print X.shape, Y.shape
```

- *La matriu X té 1797 files i 64 columnes (que es corresponen a matrius de 8x8) mentre que la variable Y té 1797 files i una columna. Podeu executar* `print digits.DESCR` *per obtenir més informació.*

➢ Implementa aquestes funcions en l'script python

➢ Fes una breu descripció al teu informe sobre quines són les estadístiques bàsiques d'aquestes dades. Mitjanes, desviacions típiques, nombre de elements d'entrenament per cada classe…

➢ (OPCIONAL) prova d'importar la llibreria matplotlib i fer un imshow plot de 8x8 d'algun dígit. Et pots ajudar de la informació de:
http://matplotlib.org/users/image_tutorial.html

2. Divisió amb train i test i normalització de les dades

- Mira i estudia la funció: `sklearn.model_selection.train_test_split`

http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html

➢ Divideix les dades entre train (70%) i test (30%) i posa-ho a les variables X_train, X_test, Y_train i Y_test

➢ **Mira i estudia:**
http://scikit-learn.org/stable/modules/preprocessing.html

➢ Normalitza les dades X (train i test) per tal que estiguin centrades a 0 amb desviació típica 1 (normalització z-score) tenint en compte les estadístiques del train.

3. Projecció en diferents components principals

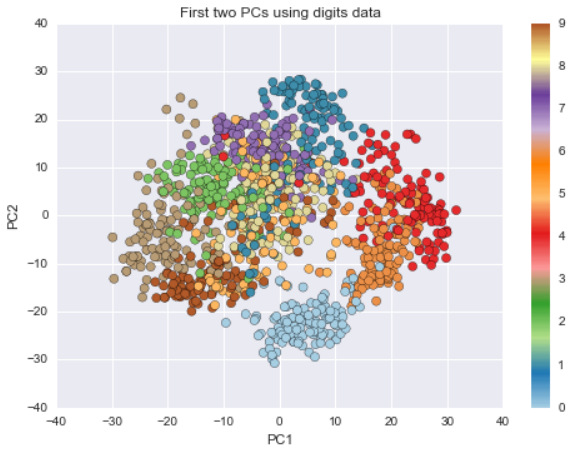
- Mira la descomposició de les dades en components principals mitjançant:
    - Anàlisi de components principals:

http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

    - Descomposició en valors singulars:

http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.TruncatedSVD.html

- Descomposa les dades en 2 components principals segons els dos mètodes presentats
- (OPCIONAL) Prova de fer un scatter Plot de les dades per classes amb les 2 tècniques que s'han presentat. T'hauria de sortir quelcom així:



- Se t'acudeix algun altre mètode de descomposició en projeccions per aconseguir millor separació de les dades? Prova de trobar la seva implementació a la web scikit-learn… Pots aplicar-lo a les dades?

4. **Fes servir validació creuada per estimar el nombre òptim de veïns K**

- Fent servir la funció `sklearn.model_selection.KFold`, farem una divisió en *10-fold cross validation* per estimar primer el nombre òptim de veïns, el nombre òptim de dimensions.

➢ **Definició de la funció de test**

➢ Defineix una funció de test que avalui la idoneitat, ajuda't de sklearn.metrics.

- http://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics

```
def compute_test(x_test,y_test,clf,cv):
    Kfolds= sklearn.model_selection.Kfold(…)
```

http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html

```
    scores=[]
    for i,j in Kfolds:
        …
        scores.append(…)
    return scores
```

➢ **Implementació de la cerca dels K més propers**

- Per fer una parametrització respecte els K elements més propers fem servir (http://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html):
  - `parameters = {'n_neighbors':k,…}`
  - `knearest=sklearn.neighbors.KneighborsClassifier()`
  - `model=fit(train_X,train_y)`
  - `predicted_y=predict(test_X)`

➢ Combina la funció 4.1 amb la funció 4.2 per obtenir els valors cross-validats segons certs paràmetres.

➢ **Com es pot fer un procés de cerca de paràmetres?**

- Ara que ja tens el mètode d'aprenentatge, ja pots fer una cerca sobre:
  - Mètode de transformació de l'entrada
  - Nombre de dimensions d'entrada del classificador

- Nombre de veïns a emprar
- Esquema de ponderació de l'algorisme
- Et pots ajudar de:

http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

- Fes una cerca de paràmetres per tenir el millor classificador possible, printa o visualitza gràficament la evolució de la cerca i mostra els resultats i com hi has arribat.

5. Conclusions

- Reporta les conclusions de la pràctica discutint, almenys, els següents apartats amb exemples i judicis

a) Explicació l'efecte de la dimensionalitat en KNN

b) Com comprens i entens els resultats

Felicitats! No només has après IBL amb Sckikit-Learn sinó que també has après a reconèixer dígits, transformar l'espai d'entrada i treballar en processos de cerca!