PAC2: Extracció de Característiques i Classificació

Presentació

En aquesta prova d'avaluació estudiarem com aplicar tècniques d'extracció de característiques a dades d'un estudi mèdic sobre malalties del cor.

Competències

En aquest enunciat es treballaran en un determinat grau les competències generals de màster següents:

- Capacitat per a projectar, calcular i dissenyar productes, processos i instal·lacions en tots els àmbits de l'enginyeria en informàtica
- Capacitat per al modelat matemàtic, càlcul i simulació en centres tecnològics i d'enginyeria d'empresa, particularment en tasques de recerca, desenvolupament i innovació en tots els àmbits relacionats amb l'enginyeria en informàtica
- Capacitat per a l'aplicació dels coneixements adquirits i per solucionar problemes en entorns nous o poc coneguts dins de contextos més amplis i multidisciplinars, sent capaços d'integrar aquest coneixements
- Posseir habilitats per a l'aprenentatge continuat, autodirigit i autònom
- Capacitat per modelar, dissenyar, definir l'arquitectura, implementar, gestionar, operar, administrar i mantenir aplicacions, xarxes, sistemes, serveis i continguts informàtics
- Capacitat per assegurar, gestionar, auditar i certificar la qualitat dels desenvolupaments, processos, sistemes, serveis, aplicacions i productes informàtics

Les competències específiques d'aquesta assignatura que es treballaran són:

- Entendre que és l'aprenentatge automàtic en el context de la Intel·ligència Artificial
- Distingir entre els diferents tipus i mètodes d'aprenentatge
- Aplicar les tècniques estudiades a un cas concret

Objectius

L'objectiu d'aquesta prova d'avaluació és l'aplicació de tècniques d'extracció de característiques i de classificació sobre dades de persones classificades segons el seu estat del cor, en un total de 5 categories. Per la qual cosa estarem resolent un problema multi-classe (versus binari o dos classes).

Descripció de la PAC

En aquesta prova ens familiaritzarem amb els mètodes d'extracció i selecció de característiques (mòdul 3) i mètodes d'aprenentatge supervisat (classificadors) d'acord amb el que s'explica al mòdul 4.

L'arxiu de dades conté informació sobre 303 persones. Cada persona està etiquetada amb una classe que pot prendre fins a 5 valors:

- 0: no hi ha cap malaltia del cor
- 1: malaltia simple en els vasos sanguini
- 2: malaltia doble en els vasos sanguinis
- 3: malaltia triple en els vasos sanguinis
- 4: malaltia en l'arteria principal coronaria

La URL per descarregar les dades és
http://archive.ics.uci.edu/ml/datasets/heart+disease

De fet, trobareu fins a 4 conjunts de dades, en funció de l'hospital que les ha recollit: Cleveland, Hungary, Switzerland, i VA Long Beach. A més a més, hi ha dos versions de les dades, de 75 variables i de 13 variables. En el darrer cas, el procés de reducció de les variables ha estat realitzat per experts en el domini. En aquesta PAC, per simplicitat, treballarem amb la versió processada de les dades Cleveland: 13 variables. L'arxiu corresponent és "**processedCleveland.csv**".

En la URL trobareu més detalls de les explicacions de les variables, així com diferents articles que les descriuen. Una de les referències citades on expliquen millor les dades és:

Liping Wei and Russ B. Altman. An Automated System for Generating Comparative Disease Profiles and Making Diagnoses. Section on Medical Informatics Stanford University School of Medicine, MSOB X215.

Requeriments:

- Matplotlib

https://matplotlib.org/users/installing.html

```
python –mpip install –U pip
python –mpip install –U matplotlib
```

- Scikit-Learn

http://scikit-learn.org/stable/install.html

```
pip install -U scikit-learn
```

- Scipy

https://scipy.org/install.html

```
pip install -U scipy
```

- Pandas

https://pandas.pydata.org/pandas-docs/version/0.20/install.html

```
pip install pandas
```

Draft code:

Es proporciona un codi inicial on teniu el mètode LoadAndPreprocess(...) per llegir l'arxiu, i fer un petit pre-processament, amb la finalitat d'eliminar els valors desconeguts i escalar les dades. Retorna dos matrius de dades: els exemples, i les classes a la qual pertany cada exemple.

El codi també està pensat perquè desenvolupeu el que se us demana en cadascun dels exercicis descrits aquí sota.

Exercici 1

Apliqueu una anàlisi PCA a les dades de la PAC:

a) Quantes components principals són necessàries per representar un 95% de la variància de les dades originals? Raoneu la resposta tenint en compte els diferents arxius disponibles a la URL.
b) Reconstruïu el conjunt de dades a partir de les 2, 4 i 8 components principals (mitjançant el mètode *inverse_transform*), i calculeu la pèrdua d'informació respecte al conjunt original per a cada cas. Per a fer-ho, podeu calcular la mitjana de les diferències elevades al quadrat entre cada element del conjunt reconstruït i l'original. Quina relació tenen aquests valors respecte a les variàncies acumulades calculades a l'apartat anterior?
c) Visualitzar les dades original (2 primers variables) i les dades transformades segons els 2 components principals. Comenteu el que veieu.

Exercici 2

En aquest exercici compararem els resultats d'aplicar un classificador simple amb les dades originals i les dades d'acord amb les característiques obtingudes pel PCA.

a) Aplicar KNeighborsClassifier, amb k=3, sobre el conjunt de dades originals
b) Aplicar KNeighborsClassifier, amb k=3, sobre el conjunt de dades transformades pel PCA amb 8 components
c) Comentar els resultats obtinguts, i la implicació respecte a la reducció de la dimensionalitat amb datasets complexos.

Exercici 3

En aquest exercici treballarem amb els següents classificadors de scikit-learn, amb els paràmetres indicats a continuació, per tal d'obtenir l'*score*, el *training time*, i el *prediction time* quan l'apliquem a les dades d'aquesta PAC.

- k Nearest Neighbors (mòdul KneighborsClassifier de sklearn.neighbors): amb 3, 4 i 5 veïns (primer paràmetre).

- Linear SVM (mòdul SVC de sklearn.svm): *kernel="linear"*, *C=0.025*, la resta de paràmetres, valor per defecte.

- Decision Tree (mòdul DecisionTreeClassifier de sklearn.tree): criterion='entropy', *max_depth=5*, la resta de paràmetres, valor per defecte.

- AdaBoost (mòdul AdaBoostClassifier de sklearn.ensemble): paràmetres per defecte.

- Gaussian Naive Bayes (mòdul GaussianNB de sklearn.naive_bayes): paràmetres per defecte.

Exercici 4

El data set proporcionat està des balancejat. Amb el mètode que hàgiu obtingut millor resultat en l'apartat anterior:

a) Compareu el resultat de la mesura per defecte i la precision_score amb el paràmetre average="macro". Com es veuen afectat els resultats.

b) Ara compareu els resultats depenent de si useu kFold o StratifiedKfold. Quina és l'opció vàlida?

Recordeu consultar la documentació del scikit-learn a la seva web (www.scikit-learn.org) per tal de descobrir els mètodes i funcionalitats ja implementats que us poden facilitar el desenvolupament de les vostres solucions.

Recursos

Aquesta pràctica requereix els recursos següents:

Bàsics:

Per a realitzar aquesta PAC disposeu d'uns fitxers adjunts:

- processedClevaland.csv

Complementaris: manual de teoria de l'assignatura, vídeos de l'assignatura, web de scikit-learn.

Criteris de valoració

Els exercicis tindran la valoració següent associada:

Exercici 1: 2 punts

Exercici 2: 2 punts

Exercici 3: 3 punts

Exercici 4: 3 punts

S'ha d'incloure el codi font de les vostres solucions a l'entrega.

S'han de raonar les respostes de tots els exercicis. Les respostes sense justificació no rebran puntuació.

Format i data de lliurament

La pràctica s'ha de lliurar abans del **proper 1 de maig** (inclòs).

La solució ha de consistir en un arxiu zip que contingui un informe en format pdf i els arxius en format python (*.py) que corresponguin a la solució adoptada.

Adjunteu l'arxiu a un missatge en el apartat de **Lliurament i Registre de AC (RAC)**. El nom de l'arxiu ha de ser CognomsNom_IAA_PAC2 amb extensió zip.

Per a dubtes i aclariments sobre l'enunciat, dirigiu-vos al consultor responsable de l'aula.

Nota: Propietat intel·lectual

Sovint és inevitable, en produir una obra multimèdia, fer ús de recursos creats per terceres persones. És per tant comprensible fer-ho en el marc d'una pràctica dels estudis del Grau Multimèdia, sempre i això es documenti clarament i no suposi plagi en la pràctica.

Per tant, en presentar una pràctica que faci ús de recursos aliens, s'ha de presentar juntament amb ella un document en què es detallin tots ells, especificant el nom de cada recurs, el seu autor, el lloc on es va obtenir i el seu estatus legal: si l'obra està protegida pel copyright o s'acull a alguna altra llicència d'ús (Creative Commons, llicència GNU, GPL ...). L'estudiant haurà d'assegurar-se que la llicència que sigui no impedeix específicament seu ús en el marc de la pràctica. En cas de no trobar la informació corresponent haurà d'assumir que l'obra està protegida pel copyright.

Hauran, a més, adjuntar els fitxers originals quan les obres utilitzades siguin digitals, i el seu codi font si correspon.

Un altre punt a considerar és que qualsevol pràctica que faci ús de recursos protegits pel copyright no podrà en cap cas publicar-se en Mosaic, la revista del Graduat en Multimèdia a la UOC, a no ser que els propietaris dels drets intel·lectuals donin la seva autorització explícita.