Pràctica: Anàlisi de patrons amb dades sintètiques

Presentació

A l'àmbit de la intel·ligència artificial, sovint és convenient disposar de dades simulades per tal de poder validar el funcionament d'una determinada tècnica de reconeiximent de patrons. Aquesta classe de dades s'anomena *dades sintètiques* perquè no han estat obtingudes d'un sistema real mitjançant tècniques experimentals.

L'objectiu d'aquesta pràctica és doble: d'una banda, es vol que l'alumne sigui capaç de generar dades sintètiques amb determinades propietats. D'altra, la intenció és aprofundir més en la comprensió de diferents tècniques emprades durant el curs, en particular els mètodes de reducció de dimensionalitat i les tècniques de classificació.

L'hipòtesi de partida és que en aplicar-les a un conjunt de dades sintètiques amb propietats conegudes, serà més fàcil fer un anàlisi dels resultats obtinguts i per tant comprendre'n amb més detall el funcionament de cada tècnica.

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 generar dades sintètiques i aplicar-ne diferents tècniques de reconeiximent de patrons. Per la implementació dels diferents mètodes d'anàlisi es faran servir les eines d'intel·ligència artificial incloses a la llibreria scikit-learn: http://scikit-learn.org/stable/index.html

Descripció de la pràctica a realitzar

Exercici 1: reducció de la dimensionalitat

a) Genereu un conjunt de dades sintètiques amb N=1000 observacions i 4 variables. La primera variable (*x1*) ha de seguir una **distribució normal univariada** amb mitjana 0.5 i desviació típica 0.2, és a dir x1~N(0.5, 0.1). La segona variables serà x2~N(4.5, 0.6). La tercera variable serà una combinació lineal de la primera x3 = 3*x1 + 1. La quarta, una combinació no lineal de les dues primeres x4 = $2*x1^2$ + x2. Finalment, construïu una matriu de dades *A* de tamany Nx4 amb les variables *x1, x2, x3, x4*.

http://docs.scipy.org/doc/numpy/reference/generated/numpy.concatenate.html

b) Per cada combinació de variables, representeu les dades gràficament com a núvols de punts bidimensionals (*x1* vs *x2*, etc.). Podeu fer servir la funció *scatter* de la llibreria *matplotlib*:

http://matplotlib.org/api/pyplot_api.html?highlight=scatter#matplotlib.pyplot.scatter

Comenteu breument les gràfiques obtingudes i la forma del núvol de punts tenint en compte les propietats de les dades.

c) Apliqueu PCA a les dades anteriors i comenteu els resultats obtinguts tenint en compte les propietats de les dades. La funció PCA de scikit-learn està descrita a

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

Quantes variables descorrelacionades hi han? Quants components principals són necessaris per explicar un 95% de la variabilitat de les dades?

d) Compareu els resultats obtinguts als apartats anteriors amb els que s'obtindrien en cas que les 4 variables estiguessin distribuïdes amb distribucions independents x~N(1, 0.1). Justifiqueu raonadament els resultats obtinguts.

Exercici 2

a) Genereu un conjunt de dades bidimensional *A1* amb N=2000 observacions amb una distribució normal amb mitjana [5, -4] i matriu de covariança [2, -1; -1, 2], i un altre conjunt *A2* amb mitjana [1, -3] i matriu de covariança [1, 1.5; 1.5, 3]. Feu servir la següent funció de la llibreria *numpy*:

http://docs.scipy.org/doc/numpy/reference/generated/numpy.random.multivariate_normal.html

b) Representeu les dades de tots dos conjunts en forma de núvol de punts en un únic gràfic. Feu servir simbols de color diferent per representar les dades de cada conjunt.

c) Construïu un conjunt de dades d'entrenament (*training*) amb les 1000 primeres observacions de cada conjunt de dades i un conjunt de prova (*test*) amb les altres 1000. Feu servir les següents eines de classificació disponibles a les llibreries scikit-learn per dissenyar un sistema de classificació automàtica dels conjunts A1 i A2 a partir de les dades de test.

Naïve Bayes: Funció GaussianNB del paquet sklearn.naive_bayes. http://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html

Anàlisi de Discriminants Lineal: Funció LDA del paquet sklearn.lda. http://scikit-learn.org/stable/modules/generated/sklearn.lda.LDA.html

En cada cas, determineu el percentatge d'encerts i errors de classificació sobre el conjunt de validació.

d) Repetiu l'anàlisi anterior quan la mitjana del primer conjunt de dades *A1* és [2, -4] en comptes de [5, -4]. Comenteu raonadament els resultats obtinguts i justifiqueu les diferències observades tenint en compte les característiques de les dades sintètiques utilitzades.

Recursos

Aquesta pràctica requereix els recursos següents:

Bàsics:

-materials i codi de l'assignatura

-eines i documentació de la llibreria scikit-learn:

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

Complementaris:

-llibrera de representació gràfica matplotlib:

http://matplotlib.org

Criteris de valoració

Els exercicis tindran la valoració següent associada:

Exercici 1: 5 punts (a -> 1 punt, b -> 1 punt, c -> 2 punts, d -> 1 punt)

Exercici 2: 5 punts (a -> punt, b -> 1 punt, c -> 2 punts, d -> 1 punt)

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 juny** (abans de les 24h).

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_Pràctica amb extensió xip.

Per a dubtes i aclaracions 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 Màster en Informàtica, 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.