PAC 4: optimització

Presentació

En aquesta prova s'aplicaran diferents algorismes d'optimització per tal d'ajustar els paràmetres d'un altre algorisme, en aquest cas de classificació. Aquesta és una aplicació molt habitual dels algorismes d'optimització.

Competències

En aquest enunciat es treballen en un determinat grau les següents competències general de màster:

- 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 d'investigació, desenvolupament i innovació en tots els àmbits relacionats amb l'enginyeria en informàtica
- Capacitat per a l'aplicació dels coneixments adquirits i de solucionar problemes en entorns nous o poc coneguts dins de contextes més amplis i multidisciplinars, essent capaços d'integrar aquests coneixements.
- Posseir habilitats per a l'aprenentatge continuat, autodirigit i autònom.
- Capacitat per a modelar, dissenyar, definir l'arquitectura, implantar, gestionar, operar, administrar y 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 és treballen 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

En aquesta PAC es practicaran els conceptes del temari relacionats amb optimització, en una vertent pràctica amb un cas concret d'ajust de paràmetres.

Descripció de la PAC/pràctica a realitzar

Dades

El conjunt de dades consta de 2000 dígits del 0 al 9 escrits a mà per diferents persones. Les imatges obtingudes han estat netejades de soroll, centrades i escalades (15x16 píxels), de manera que totes tenen la mateixa mida.

El nombre de dígits de cada classe 0..9 és el mateix (200), i a tots els fitxers de dades apareixen ordenats (els primers 200 són '0', els següents 200 són '1', etc.).

En aquest conjunt de dades es proporcionen diferents mesures útils pel processament d'imatges, de les quals en aquest cas es farà servir el següent:

- Les correlacions dels perfils (216 valors, fitxer *mfeat-fac.txt*).

Aquest conjunt de dades es diu “Multiple Features” i s'ha obtingut del Machine Learning Repository de la Universitat de California Irvine:

https://archive.ics.uci.edu/ml/datasets/Multiple+Features

Es vol entrenar un classificador amb màquines de suport vectorial i nucli gaussià i es volen determinar els valors òptims de dos paràmetres per aquest classificador:

- **C**: penalització per mostres mal classificades. Un valor baix tendeix a crear corves de decisió més suaus (no és important si una mostra està mal classificada), i un valor més alt tendeix a crear corves de decisió molt ajustades a les dades (es penalitza molt que hi hagi una mostra mal classificada).
- **gamma**: inversa del radi d'influència de cada mostra al nucli gaussià. Valors baixos signifiquen que la influència de cada mostra té un radi molt gran, mentre que valors alts signifiquen que la influència de cada mostra té un radi petit.

A la pàgina següent podeu veure una discussió de l'ús, significat i efecte d'aquests paràmetres en el funcionament de les SVM amb sklearn:

http://scikit-learn.org/stable/auto_examples/svm/plot_rbf_parameters.html

Activitat 1

Efectuar, si cal, el tractament previ de les dades (tractament de valors absents o erronis, normalització, centrat, etc.). Justifiqueu la necessitat de cada operació segons les característiques del problema.

Per cada experiment que es faci caldrà dividir el conjunt de dades en tres:

- Entrenament: seran les dades amb les quals s'entrenarà el classificador.
- Prova: seran les dades amb les quals l'algorisme d'optimització valorarà una solució concreta. És a dir, a cada iteració de l'algorisme d'optimització es calcularà
- Validació: seran les dades amb les quals es valorarà el resultat final de cada experiment.

Disposeu les dades de manera que es divideixin en aquests tres conjunts complint les propietats que considereu necessàries.

Nota: la solución de esta PEC se encuentra en el fichero **solucion.py**.

Según las especificaciones de los ficheros de entrada, no hay valores ausentes o erróneos. Se aplica un escalado estándar a los datos para evitar escalas diferentes en los diferentes atributos, aunque no es necesario por lo que indican los autores de los datos.

Para llevar a cabo estos experimentos se dividirán los datos iniciales en tres conjuntos:

-Entrenamiento (60% de las muestras). Con estos datos se entrena el clasificador cada vez que se quiera probar una configuración.

-Prueba (20% de las muestras). Con estos datos se calcula el grado de acierto del clasificador en cada configuración probada.

-Validación (20% de las muestras). Estas muestras servirán para comprobar, al acabar la el proceso de optimización, que los parámetros obtenidos no sólo sirven para los datos de prueba sino también para otros datos, es decir el resultado es generalizable.

La asignación de las muestras a cada subconjunto se hace de manera aleatoria. Dado el tamaño del conjunto de datos no parece crítico comprobar que los conjuntos estén equilibrados (se supone que lo estarán al haber muchas muestras), aunque conviene comprobarlo.

Esta división se puede hacer varias veces y ejecutar los experimentos con diferentes selecciones para analizar la varianza entre experimentos.

Activitat 2

Trieu una mesura de la qualitat del classificador. Convé que sigui un únic valor per tal de fer-lo servir com a funció objectiu dels algorismes d'optimització.

Como en esta PEC se usará el clasificador de máquinas de vectores de soporte de la biblioteca **sklearn**, se usará la función **score** que incluye y que da el grado de acierto de la clasificación en el rango [0,1].

De hecho como lo que espera el optimizador es una función para **minimizar**, lo que se hará será definir como función objetivo **1 – score**, es decir minimizar el error de clasificación.

Activitat 3

Apliqueu l'algorisme d'optimització **recocció simulada** per trobar els valors òptims per **C** i **gamma**. Compareu els resultats sobre les dades de prova i les de validació. Hi ha gaires diferències? Per què?

Opcional: si voleu experimentar una mica, podeu estudiar la diferència entre els resultats amb les dades de prova i validació en funció dels paràmetres **C** i **gamma**. Hi ha alguna correlació? Per què?

Se ha adaptado el código de la recocción simulada de los materiales de la asignatura a este problema. Principalmente se ha tenido que definir una función para generar vecinos, que en este caso se ha tomado como vecino una configuración con alguno de los dos parámetros (C o gamma) multiplicado o dividido por 1.1. Todo esto con diferentes selecciones aleatorias.

Mientras **C** es independientes de la escala de los datos, **gamma** sí que depende, y al haber escalado los datos hay que ajustar la escala de **gamma**. En este caso, se ha estudiado el intervalo [10^-4, 1].

Los resultados obtenidos con este método con 50 iteraciones son:

Datos	C	Gamma	Score
Test	1.331	0.513	0.922
Validación	1.331	0.513	0.920

Lógicamente los resultados con los datos de validación son un poco inferiores a los de los datos de test, porque el optimizador ha buscado la configuración que mejor se adaptara a los datos de test. De todos modos la diferencia no es significativa, lo que significa que el clasificador es suficientemente general como para aplicarse a nuevos datos.

Activitat 4

Apliqueu **algorismes genètics** per trobar els valors òptims per **C** i **gamma**. Compareu amb els resultats anteriors.

Quin mètode considereu més adient per aquest problema?

La adaptación del código de algoritmos genéticos a este problema también es bastante directa. En este caso, el cruzamiento de individuos es muy sencillo: como sólo hay dos atributos por individuo, se toma el primero de un progenitor y el segundo del otro. Las mutaciones también son bastante sencillas y consisten, como en la actividad anterior, en multiplicar o dividir alguno de los atributos por un valor.

Los resultados obtenidos, con una población de 10 individuos y 50 iteraciones, son los siguientes:

Datos	C	Gamma	Score
Test	4856	0.121	0.568
Validación	4856	0.121	0.555

Los resultados son bastante peores, seguramente porque en recocción simulada se ha partido de un punto relativamente bueno mientras que genéticos está explorando aleatoriamente todo el espacio de soluciones.

Nótese que el número de ejecuciones de la SVM es mucho mayor (individuos * iteraciones) que en la recocción simulada, y esto hace que en este problema se prefiera aquel método. Esto, en definitiva, es a causa de:

-Coste relativamente alto de calcular la función objetivo (entrenamiento + predicción SVM), por lo que conviene ahorrar evaluaciones.

-Pocos atributos (dimensionalidad), que es donde los algoritmos genéticos pueden suponer una ayuda mayor.

Por todos estos motivos se recomienda, en este problema, la recocción simulada.

Recursos

Aquest PAC requereix dels següents recursos:

Bàsics: Fitxer de dades adjunts a l'enunciat (*dades_pac4.zip*).

Complementaris:

Manual de teoria de l'assignatura. En especial els fitxers de codi del tema 5.

Criteris de valoració

Els exercicis tindran la següent valoració associada:

Activitat 1: 3 punts

Activitat 2: 2.5 punts

Activitat 3: 2.5 punts

Activitat 4: 2 punts

Raoneu la resposta en tots els exercicis. Les respostes sense justificació no rebran puntuació.

Format i data de lliurament

La PAC s'ha de lliurar abans del proper 5 de juny (a les 24h).

La solució a entregar consisteix en un informe en format PDF fent servir la plantilla penjada al tauler de l'assignatura més els fitxers de codi (*.py) que heu fet servir per resoldre la prova. Aquests fitxers s'han de comprimir en un fitxer ZIP.

Adjunteu el fitxer a un missatge a l'apartat de **Lliurament i Registre d'AC (RAC).** El nom del fitxer ha de ser *CognomsNom*_IAA_PAC4 amb l'extensió *.zip.*

Per a dubtes i aclariments sobre l'enunciat, adreceu-vos al consultor responsable de la vostra 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.