DATA MINING: Quiz 2

Student's name: Mariona Farré

1. Assume you have 5 *independent* classifiers, each of them with an accuracy of 0.7. Compute which is the accuracy for the *Majority Vote* algorithm for those 5 classifiers.

Per calcular el Majority Vote es necessita calcular la distribució binomial, la probabilitat que la majoria dels classificadors(3,4 o 5 dels 5 classificadors) siguin certs:

$$P(k; n, p) = C(n, k) \cdot p^k \cdot (1-p)^{n-k}$$

On:

- P(k;n,p) és la probabilitat de k sigui certa en n torns si hi ha una probabilitat de p
- C(n, k) és el coeficient binominal, el número triat serà cert k vegades dels n torns

L'ccuracy es queda com:

Pmv= P(3; 5, 0.7) + P(4; 5, 0.7) + P(5; 5, 0.7)

Els valors queden com:

Pmv = C(5, 3) · 0.7^3 · 0.3^2 + C(5, 4) · 0.7^4 · 0.3^1 + C(5, 5) · 0.7^5 · 0.3^0

= 10 · 0,343 · 0,027+ 5· 0,2401 · 0,3 + 1· 0,16807 · 1

= 0.09261 + 0.36015 + 0.16807

= 0.62083

El resultat de majority voting dels 5 classifiers independents amb una accuracy de 0,7 és de: **0,62**

2. Briefly explain if each of the following claims is true or not and why:
   a. The larger the number of iterations in the bagging method, the lower the variance of results and the larger the accuracy obtained

És veritat, ja que bagging és una tècnica que redueix la variança fent una mitjana de tots els valors trobats, si s'incrementen les iteracions, la variança de les prediccions baixarà, no vol dir que la accuracy pujarà per cada iteració, si que millorarà, però fins a un cert punt on s'estabilitzarà.

   b. Boosting cannot be applied to support vector machines because the linear combination of hyperplanes is another hyperplane.

És fals, el boosting es pot aplicar a qualsevol algorisme d'aprenentatge, incloent-hi les svms. Igualment que l'enunciat de combinació lineal dels hiperplans és correcte, el boosting entrena els models de manera secuencial i no necessita cap propietat específica.

   c. When the "a" parameter in random forests is set to the number of features, random forest is equivalent to bagging with decision trees.

És cert, si en random forest el paràmetre "a" és el nombre de característiques a considerar a cada node en l'arbre de decisió, serà el mateix que si es fa servir bagging per després fer els arbres de decisió.

d. Diversity of classifiers is the source of success in meta-method. In order to ensure this diversity, we always train classifiers with different training datasets.

És cert, la diversitat és molt important en les meta-mètodes, entrenar cada classificador amb un subconjunt diferent podrem garantir que cada classificador sigui diferent i que cometin errors diferents, permet conèixer millor a les dades no vistes, ja que si un error comet un error, probablement aquest serà corregit per un altre.

3. When implementing the main loop of the *Adaboost* procedure, what should we do when the error produced by the classifier on the training set (feed with a set of examples according to the current iteration weights) is equal to 0? Briefly explain why you think so.

    a. Stop the boosting iterations and return the weighted ensemble of classifiers built until that moment.
    b. Return that last classifier as the final classifier.
    c. Remove that classifier and continue the boosting loop until the limit number of iterations is achieved.
    d. Reduce the confidence on that classifier (with respect to its theoretical confidence) and continue the boosting loop until the limit number of iterations is achieved.
    e. Boosting cannot be applied in that case.

Si en l'execució de l'Adaboost si un classificador produeix zero errors en el training set, vol dir que classifica perfectament tots els exemples d'aquell set, per solucionar-ho s'hauria de fer l'apartat "a" : "Stop the boosting iterations and return the weighted ensemble of classifiers built until that moment. "

L'Adaboost treballa sequencialment agregant classificadors, fixant-se en els exemples que els altres classificadors s'han equivocat, si un classificador retorna zero errors, vol dir que ha classificat tots correctament i sobretot els exemples anteriors que abans és donaven per erronis, per això no parar i agregar més classificadors és poc probable que millori el model, ja que no tindrà exemples erronis per fitxar-se en arreglar.

Les altres opcions no seran les correctes, l'apartat "b" l'últim classificador no vol dir que sigui el millor ja que s'ignoren les millores que classificadors anteriors haguin pogut aconseguir. En l'apartat "c" si treiem el classificador que ha retornat zero errors perdrem el classificador que potencialment seria la millor opció pel mètode i arribar el màxim d'iteracions no vol dir que s'aconsegueixin els millors resultats. En l'apartat "d" reduir la confiança no és necessari, si el classificador ja treu zero errors, no hi ha cap raó per reduir la seva influència en el classificador final i en l'apartat "e" és erroni, ja que boosting es pot aplicar igualment que el classificador retorni zero errors en el training set.

4. After building a support vector machine with a linear kernel with a given C, we found the number of support vectors is very large. If we want to decrease the number of support vector, what should we do? Explain why.

    a. Decrease the C value
    b. Increase the C value
    c. Change to the RBF kernel
    d. Try a Polynomic kernel
    e. None of the above

En les SVM el paràmetre “C” és el que controla l'equilibri entre aconseguir un error baix en les dades d’entrenament i que el model no tingui una complexitat massa elevada, si volem decréixer el nombre de vectors de suport hem de triar la “b”: “Increase the C value ”
Si C és gran s’intentarà minimitzar el nombre d’errors en el training set, igualment que això faci que sigui més complex, portant a que es facin servir menys vectors de suport, perquè el límit de decisió pot ser més flexible i ajustar millor les dades.
Les altres opcions no son correctes, ja que en l'apartat “a” quan C és petit SVM té un marge més gran en un major nombre d’errors en el training set fent que poguessin augmentar els vectors de suport. Canviar en un kernel RBF o polinòmic com en l’aparat “c” i “d” no afecta directament al número de vectors de suport, podria canviar els límits de decisió però no el nombre de vectors de suport. I en l'apartat “e”, seria incorrecte perquè en l’apartat “a” ens dona una opció òptima per ajudar a reduir el nombre de vectors de suport.

5. In the last few years, Artificial Intelligence has advanced a lot. Believe it or not, in the attached file "ChatGPT answers about SVMs.pdf" you will find a dialog I had about Support Vector Machines with ChatGPT. ChatGPT is an amazing chat bot developed by OpenAI that has been trained on a lot of textual data of different types (but without internet access). Its answers have really surprised me for their clarity, expressiveness and knowledge of the topic. However, ChatGPT answers are known to be not always correct (even when it gives convincing explanations... which turn out to be wrong). Your goal is to detect the answers that are wrong (if any) in the SVMs dialog. You have to write here the number of the questions (in red in the pdf) you think are wrong together with the correct answers.
(More space for answers if you need it)

Les respostes que ChatGPT ha respost sobre les SVM són:

- Q1: Respon correctament amb una descripció general de les SVM

Dona resum que són les SVM, descriu què són els límits de decisió, els hiperplans i el marge, com es fan les prediccions una vegada entrenats, la seva efectivitat i la seva implementació.

- Q2: Respon correctament a que és el marge en les SVM

Dona la definició de què és un marge, les seves implicacions si és més gran o més petit, el seu objectiu amb les SVM i la importància en la optimització del límit de decisió.

- Q3: Respon correctament a que són els suports en les SVM

Dona la definició de que és una SVM i que són els suports i la importància d'aquests, els marges i les mesures de l'execució del rendiment i l'objectiu d'aquest classificador.

- Q4: Respon correctament a que és un kernel en les SVM

Explica la funció del kernel, com trobar el límit de decisió, la transformació a una dimensió més en l'espai dels hiperplans on es poden separar millor les classes i com trobar els productes de punts de les SVM.

- Q5: Respon incorrectament que si hi ha una relació entre el nombre de suports en una SVM i la generalització de l'error.

Explica el nombre de suports i la relació amb la generalització de l'error, però aquesta relació no és directe, igualment que sigui cert, que hi ha un augment en el nombre de suports no vol dir exactament que l'error baixi. El nombre de suports pot ajudar a decidir la complexitat però si hi han masses pot causar un overfitting a les dades. També diu que un nombre menor de suports indica que hi hagi overfitting, però normalment no és el cas, ja que generalment causaria un underfitting de les dades.

- Q6: Respon correctament a quins son els valors d'alpha en les SVM

Explica els possibles valors de alpha, com es troba els límits de decisions, i els algorismes i problemes d'optimització per solucionar correctament les SVM.

- Q7: Respon correctament a que és un kernel en les SVM

Dona informació dels valors d'alpha i les Lagrange multiplicadors, les limitacions d'alpha entre zero i el paràmetre C, el rang de valors vàlids i altres limitacions que té i el seu paper de determinar el pes de cada training set en les SVM.

- Q8: Respon incorrectament a que és el valor de alpha en un non-support vector

Dona informació de el valor d'alpha i el seu rol en les SVM, explicant també els vectors de suport correctament, però explica que els non-support vectors tenen un valor d'alpha més petit o igual a zero, dada incorrecte ja que segons les SVM serien dades incorrectes perquè el valor de alpha estrictament no pot ser negatiu.

- Q9: Respon correctament a que és una variable slack en les SVM

Explica que son les variables slack en les SVM, per així aconseguir una classificació amb un marge "soft" permetent alguns errors de classificació per aconseguir un límit de decisió més robust, com aquest pot ser ajustat i tolerar errors per aconseguir un millor rendiment en total.

- Q10: Respon correctament a si els gats existeixen a l'espai definit d'un rbf kernel

Dona informació sobre els rbf kernels i que transforma les dades en un espai dimensional més alt a través d'un hiperpla per poder classificar millor les dades, en aquest espai no hi té res a veure un animal molt menys un gat, només tindrà sentit aquest és un nom d'un classificador.

- Q11: Respon correctament quin grup polític dona suport les SVM

Deixa clar que les SVM son algorismes matemàtics i no subjectes de cap grup polític ni de l'oposició, dos temes completament diferents, i si les SVM analitzen les dades de grups polítics són totalment objectius i crearien classificadors imparcials.

Així que les respostes errònies són: Q5 i Q8