Quiz 1 - MINERIA DE DADES - November, 2023

Name:	Mariona Farré Tapias

Answer the following questions in the spaces reserved for this use.

1. *(2pt)* Write whether the following problems can be solved using supervised data mining algorithms for classi fication or not. Explain very briefly why or why not.

   (a) Given a dataset with sensor data from a self-driving car, including radar, lidar, and camera inputs, predict the car's current location.

   Aquest problema no seria una bona proposta per un algorisme de classificació, ja que per trobar la localització a temps real es necessitaria una constant sortida de dades i fer càlculs d'aquests, i no una classificació de dades analitzant atributs del cotxe.

   Només es podria fer si hi hagués una categorització les zones per la localització, i es podria calcular si el cotxe està o no en aquella zona, per exemple amb un atribut binari per cada zona, 0 si està 1 si es troba a dins.

   (b) Given a dataset of customers for an e-commerce website, including purchase history and browsing behavior, predict the average age of the customers.

   No es podria resoldre amb un algorisme de classificació, ja que l'edat és un atribut independent a les accions de l'usuari en una web i no pot fer una predicció de la mitjana d'edat amb els resultats siguin valors numèrics, perquè per poder ser solucionat per un classificador, s'hauria de crear una etiqueta per cada edat possible i no es podria trobar la mitjana d'edat.

   (c) Given a dataset of text reviews from an e-commerce platform, the task is to categorize the reviews of new texts into sentiment classes, such as positive, neutral, or negative.

   Es podria resoldre amb un algorisme de classificació, ja que el sentiment de classes, sigui positiu, neutral o negatiu, es poden convertir en etiquetes que es poden utilitzar per classificar les reviews, fent que sigui una tasca per un algorisme de classificació analitzant el text i veient en quina etiqueta aquest text pertany.

   (d) Given a dataset of likes for instagram images for a specific city, the task is to predict whether a person has visited a particular tourist attraction in the city.

   Es podria resoldre amb un algorisme de classificació, si el dataset esta correctament creat amb dades de moltes persones si han anat a una atracció turística o no, amb el seu historial de m'agrades d'instagram, podem entrenar a un model que pugui classificar de forma binària una predicció de si una persona ha anat a una atracció turística amb el valor de 1, o si no ha anat amb el valor 0.

   (e) Given a dataset containing information about customers (such as age, income, and purchase history), predict whether a customer is likely to buy or not a given product.

   Es podria resoldre amb un algorisme de classificació, és un clàssic problema de data mining, ja que es pot crear un model per predir una etiqueta de si una persona comprarà o no un producte en específic, amb les dades d'altres consumidors que hauran comprat o no aquell producte i analitzant les seves dades.

2. *(1.5pt)* You are working on a classification problem. For validation purposes, you've randomly sampled the training data set into train and validation. You are confident that your model will work incredibly well on unseen data since your validation accuracy is high. However, you get shocked after getting poor test accuracy. Mention some possible reasons why this could happen.

Si un model creat funciona molt bé en dades de validació, retronant una precisió molt alta, però després en realitzar en dades no vistes, la precisió pot ser baixa per vàries raons:

- Overfitting: el sobre ajustament de les dades pot fer que el model hagi après amb certs errors de les dades del training, donant una precisió molt més baixa amb les dades noves.
- Que el validation i training set no sigui representatiu del conjunt total de dades del problema, fent que la precisió doni molt més alta en aquestes, però que baixi amb les dades noves, perquè aquest és un data set més representatiu.
- El model pot donar diferents resultats depenèn de la divisió de dades que s'han fet en el dataset pel training i la validació, es recomana dividir vàries vegades per anar aconseguint la mitjana dels resultats (K-fold cross-validation)
- Errors aleatoris, sobretot en datasets més petits on el set de validació està molt limitat de dades, pot fer que a l'hora de realitzar-ho en dades no vistes els resultats canviin.

3. *(1.25pt)* We want to learn to detect people with prostatic cancer from a training dataset containing 500 positive cases and 100.000 negative cases. What is the best approach to follow when we want to find the optimal parameters of an algorithm? Why?

(a) Run the learning algorithm and choose the parameters that return better accuracy.

(b) Run the learning algorithm and choose the parameters that returns better recall on the positive case.

(c) Run the learning algorithm and choose the parameters that returns higher precision on the positive case.

(d) Run the learning algorithm and choose the parameters that returns higher precision on the negative case

(e) Run the learning algorithm and choose the parameters that returns higher F-measure on the negative case.

(f) None of the above.

En un escenari de diagnòstics, millor fixar-se en el recall (retorn de dades) dels casos positius, ja que és més crític identificar el màxim de casos positius de la malaltia possible, igualment que això inclogui els possibles falsos positius que puguin haver-hi.

4. *(1.5pt)* Mark the true sentences. marcat en blau

(a) Cross-validation is a technique used to assess a model's performance by splitting the dataset into mul tiple subsets, training on some, and testing on others, to obtain a more robust estimate of the model's generalization performance. - Veritat

(b) When performing k-fold cross-validation, the choice of k (the number of folds) does not affect the model's performance evaluation; any value of k will yield similar results.

(c) The training and testing datasets in cross-validation are often chosen randomly, ensuring that the model is exposed to a representative sample of the data in each fold. - Veritat

(d) Validation techniques like k-fold cross-validation can be computationally expensive, especially when working with large datasets or complex models, due to the repeated training and testing steps.- Veritat

(e) ROC AUC (Receiver Operating Characteristic Area Under the Curve) is a commonly used metric for evaluating the performance of classification models, providing insight into their ability to distinguish between positive and negative classes.- Veritat

5. *(1pt)* Mark the true sentences: marcat en blau

(a) When training a classifier using the k-NN algorithm, a smaller value of k is generally more sensitive to noise in the data, which can lead to overfitting. - Veritat

(b) In k-NN classification, increasing the value of k will always result in a more complex decision boundary that can fit the training data better.

(c) k-NN is a non-parametric algorithm, meaning it doesn't make any assumptions about the underlying data distribution. - Veritat

(d) k-NN can be computationally expensive when dealing with large datasets because it requires calculating distances between the query point and all data points in the dataset.- Veritat

6. *(1pt)* Answer if each of the following sentences about the *Na¨ıve Bayes* algorithm *is true or not*.
Escrit al final de la frase en blau

(a) In general, when using *Na¨ıve Bayes* algorithm, the larger the number of features on the dataset, the better the performance Mentida, no garanteix tenir una millor performance, aquesta depèn de la rellevància i la qualitat de les etiquetes.

(b) Naive Bayes is sensitive to the scale of features, so standardization or normalization of data is typically required before using this algorithm. Mentida, és un algorisme relativament poc sensible a la normalització.

(c) The smoothing technique is used to reduce the impact of the assumption of independence of features in the dataset. Veritat, per exemple *Laplace smoothing* tracta problemes de zero probabilitat en l'algorisme

(d) When computing the conditional probability of a numerical feature with respect to the class, we always use the normal distribution. Mentida, la distribució triada dependrà del tipus de dades

7. *(0.75pt)* To reduce overfitting of a Decision Tree, mark which of the following method can be used:

marcat en blau

(a) Increase minimum number of examples allowed in leafs - Veritat

(b) Increase depth of trees

(c) Set a threshold on the minimum information gain to split a node - Veritat

8. *(1pt)* Which of the following are disadvantages of Decision Trees? marcat en blau

(a) Decision Trees will overfit the data easily if it perfectly describes the training dataset

Els arbres de decisió poden tenir un overfitting de dades fàcilment en el training set, fent que puguin donar errors per les dades errònies.

(b) Pruning a Decision Tree is a technique used to reduce its complexity and prevent overfitting by removing branches that do not contribute significantly to the model's predictive power.

(c) Decision Trees tend to be highly interpretable models, making them suitable for scenarios where under standing the decision-making process is crucial.

(d) Decision Trees are sensitive to small variations in the training data, and changing a single data point can lead to significant changes in the tree structure.

Els arbres de decisió en ser molt sensibles a variacions de dades, és difícil tenir un model que sigui precís amb tots els data tests i dades no vistes.

(e) The concept of entropy is used in the context of Decision Trees to measure the impurity of a node, with lower entropy indicating a more homogenous set of data points.

L'entropia és la mesura de bits per comprimir la informació de la distribució considerant que pot ser un resultat aleatori. Si aquest és més baix que els altres, vol dir que és un node on no es guanya molta informació en relació a els altres nodes dins del model.