Padrón: _____ Nombre y Apellido:_____

Parcial 26-10-2023

Ejercicio 1

Explique brevemente los conceptos de **Overfitting** y **Underfitting**. Mencione cómo puede detectarse cada una de estas situaciones y qué se podría hacer para evitarlas.

*El **overfitting** es el sobreajuste del modelo a los datos de entrenamiento, el modelo aprende incluso el ruido en los datos de entrada. También hay que analizar la complejidad del modelo.*
Se puede detectar cuando un modelo performa muy bien en entrenamiento y no tiene poder de generalización, es decir, tiene baja performance sobre datos nuevos. Los modelos que incurren en overfitting suelen tener alta varianza, lo que implica que son extremadamente sensibles a pequeñas alteraciones en los datos, Para evitarlo se pueden usar técnicas de regularización en los modelos, usar validación cruzada, etc.

*El **underfitting** es cuando el modelo no logra aprender sobre los datos de entrenamiento. Un modelo con underfitting es aquel en donde los errores tanto de entrenamiento como de validación son similares y relativamente altos. Los modelos víctimas del underfitting tienen un bias alto. También hay que analizar la complejidad del modelo, ya que para evitarlo se puede usar modelos más complejos, recolectar más datos si son escasos, reducir la regularización, etc.*

Ejercicio 2

Explique brevemente el concepto de **Outliers** e indique cuál es la diferencia entre valores atípicos **univariados y multivariados**. Mencione métodos/técnicas para detectar estos valores en cada caso. Dé un ejemplo concreto de algún método utilizado en el TP1.

*Los **outliers o valores atípicos** son observaciones distantes del resto de los datos, pueden deberse a un error de medición, aleatoriedad,etc. Siempre son subjetivos al problema que estamos estudiando y deben ser cuidadosamente inspeccionados porque pueden estar alertando anomalías.*
*Los **outliers univariados** son valores atípicos que podemos encontrar en una simple variable por ejemplo: en un conjunto de datos correspondientes a niños en edad escolar, encontramos una observación donde la edad es de 45 años. Métodos para detectarlos: IQR, Z-score, Z-score modificado.*
*Los **outliers multivariados** son valores atípicos que se pueden encontrar en un espacio n-dimensional, es decir analizando más de una variable o atributo. Por ejemplo en un dataset de frutas una manzana de color naranja. Métodos para detectarlos: Mahalanobis, LOF, Isolation Forest, Clustering.*

EJEMPLO CONCRETO DE APLICACIÓN AL TP1

Padrón: _____ Nombre y Apellido:_____

Ejercicio 3

Explique brevemente las principales diferencias entre:

a. Métodos de ensamble de tipo **bagging** y los de tipo **boosting**.

Bagging

- *Los conjuntos de entrenamiento se construyen mediante bootstrap (muestras aleatorias con reemplazo) para entrenar distintos modelos, y luego combinarlos.*
- *Los modelos se entrenan en paralelo.*
- *Disminuye la varianza en el modelo final*
- *….*

Boosting

- *Se construyen nuevos modelos para las instancias mal clasificadas por los anteriores.*
- *Los modelos se entrenan secuencialmente.*
- *Las nuevas instancias se clasifican usando una votación ponderada de todos los modelos construidos.*
- *….*

b. Métodos de ensamble de tipo **stacking** y los de tipo **voting**.

*Ambos son ensambles híbridos, se entrenan N modelos que pueden ser diferentes (árboles, KNN, Regresión Logística, SVM, etc). En **voting** se aplica votación mayoritaria o ponderada para decidir y en **stacking** se utiliza un metamodelo.*

Ejercicio 4

¿Cuál es la diferencia entre **clasificación, regresión y agrupamiento**? De un ejemplo de cada caso, detallando qué algoritmos podrían utilizarse en cada tipo de problema.

*La **clasificación** es una tarea de aprendizaje automático supervisada. En un problema de **clasificación**, buscamos, para ciertos datos de entrada, un categoría c de un conjunto C de categorías posibles. Estas categorías no solo son finitas, sino que además son conocidas de antemano. Ejemplo: clasificar si una persona está sana o enferma. Algoritmos: regresión logística (clasif. binaria), Árboles de Clasificación, XGBoost, etc.*

*La **regresión** es una tarea de aprendizaje automático supervisada. En un problema de **regresión** buscamos predecir un valor en un rango continuo, para ciertos valores de entrada. Ejemplo: predecir el precio de una propiedad. Algoritmos: regresión lineal simple, regresión lineal múltiple, XGBoost Regressor, Árboles de Regresión, etc.*

*El **agrupamiento** es una tarea de aprendizaje automático no supervisada. En este tipo de problemas se trata de agrupar los datos. Agruparlos de tal forma que queden definidos N conjuntos distinguibles, aunque no necesariamente se sepa que signifiquen esos conjuntos.*
El agrupamiento siempre será por características similares. Algoritmos: K-Means, Clustering Jerárquico,DBSCAN, etc

Padrón: _____ Nombre y Apellido:_____

Ejercicio 5

Responda las siguientes preguntas en relación al tema: **Reducción de Dimensionalidad**:

a. ¿Para qué usaría un gráfico "**Scree plot**", al utilizar **PCA**?

Para seleccionar el número de componentes principales, utilizando el método "elbow", de acuerdo al porcentaje de varianza explicada.

b. ¿Qué técnica utilizaría si sospecha que los datos están distribuidos sobre una ***variedad*** del espacio total?

ISOMAP: Su objetivo es preservar la geometría intrínseca de los datos reflejada en las distancias geodésicas de la variedad. La clave es encontrar una forma eficiente de calcular la verdadera distancia geodésica entre observaciones, dada solamente su distancia Euclídea en el espacio de alta dimensión

c. ¿Qué técnica utilizaría si desea conservar los ***clusters***?

TSNE, ya que la proyección de t-SNE preserva los clusters.

Ejercicio 6
Responda las siguientes preguntas en relación al tema: **Árboles de decisión**:

a. ¿Para qué utilizaría la técnica de **poda** en un **árbol C4.5**?

Se utiliza para controlar el overfitting se hace después del entrenamiento, y consiste en "podar" o eliminar algunos nodos.

b. ¿Puede un árbol manejar valores de entrada numéricos? ¿Cómo?

Sí, dado un atributo A de rango continuo de valores, se debe establecer un umbral C que divida en dos el rango maximizando la ganancia de información. Con dicho umbral se creará un booleano Ac tal que si A < C Ac = TRUE, sino Ac = FALSE.

c. ¿Para qué se usa la impureza de **Gini**? ¿Qué significa que un nodo sea puro o impuro?

Para determinar el nodo raiz y para generar las divisiones de los nodos del árbol, seleccionando el atributo que clasifica mejor. Para ello se mide la impureza de Gini. Un nodo puro tiene el 100% de instancias de la misma clase, un nodo impuro tiene instancias de diferentes clases.

Padrón: _____ Nombre y Apellido:_____

Ejercicio 7

Mencione cuál es la diferencia entre las métricas **Precision y Recall** y cómo se calculan las mismas.

Precision: TP/ TP+FP
¿Cuántas veces lo que mi modelo dice es realmente cierto?
Precision se centra en lo que el modelo dice y luego lo compara con la realidad.

Recall: TP/TP+FN
¿Cuántas veces mi modelo es capaz de identificar la verdad?
Recall parte de la realidad, y después evalúa que tan bueno es el modelo para reconocerla.

a. ¿En qué tipo de problemas sería conveniente maximizar la métrica Precision y en cuáles Recall?. Ejemplifique.

   Precisión cuando el costo de los falsos positivos sea alto Por ejemplo, en el diagnóstico médico, diagnosticar erróneamente una enfermedad a una persona sana puede generar tratamientos y gastos innecesarios.

   Recall cuando el costo de los falsos negativos sea alto: por ejemplo, en la detección de fraudes, no detectar una transacción fraudulenta puede provocar pérdidas financieras importantes.

b. ¿Son esas métricas adecuadas para resolver un problema de regresión? En caso de responder negativamente, expliqué qué métricas usaría.

   No, no son las métricas adecuadas. Para evaluar el rendimiento de los modelos en problemas de regresión se mide el error, las métricas que se pueden utilizar son MAE, MSE, RMSE etc.