APRENDIZAJE AUTOMÁTICO: POSIBLES PREGUNTAS EXAMEN

EDA

1. ¿Cuáles son las técnicas comunes de preprocesamiento de datos que se utilizan en el aprendizaje automático?

2. ¿Cómo se pueden aplicar estas técnicas para mejorar la calidad y la eficacia de los modelos de aprendizaje automático?

3. ¿Qué es la normalización de datos y por qué es importante en el preprocesamiento de datos? ¿Qué técnicas de normalización de datos se utilizan comúnmente en el aprendizaje automático?

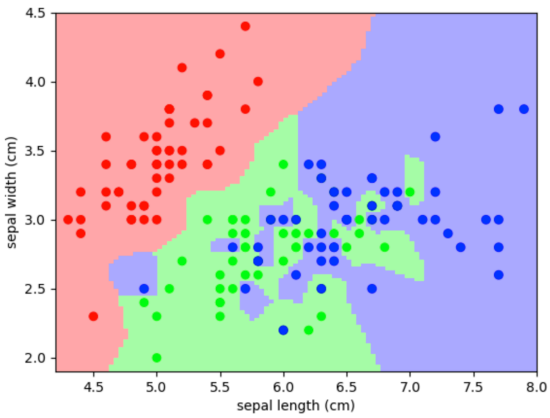
4. ¿Cómo se pueden manejar los valores atípicos (outliers) en el preprocesamiento de datos? ¿Cuáles son algunas técnicas comunes para detectar y tratar los valores atípicos en los datos?

5. ¿Qué son las técnicas de selección de características y cómo se utilizan en el preprocesamiento de datos y por qué? ¿Cuáles son algunos métodos comunes para seleccionar características relevantes en los datos?

KNN Algorithm

1. Si definimos k como el número de vecinos dentro del algoritmo de KNNClassifier, qué ocurre si incrementamos k en relación a:
- Decision boundary: más o menos complejo?
- Overfitting: aumenta o disminuye?
- Coste computacional: aumenta o disminuye?
- Error del training: aumenta o disminuye?
- Es mejor o peor si tenemos outliers, por qué?

2. Imagina que tienes el siguiente problema:



a) Si tuviéramos que discriminar las regiones verde y moradas. ¿Crees que sería mejor utilizar un k bajo o alto?
b) Si tuviéramos que discriminar las regiones roja y el resto ¿Crees que sería mejor utilizar un k bajo o alto?

3. ¿Cuáles son las ventajas y los inconvenientes de utilizar el clasificador KNN?

4. Considera un conjunto de datos con dos características: edad e ingresos. La edad se mide en años y los ingresos en miles de dólares al año.
¿Por qué es necesario normalizar los datos si utilizamos KNN? O ¿por qué no?
¿Qué ocurre al calcular la matriz de distancias?

5. Supongamos que estás trabajando en un problema de clasificación con un conjunto de datos de alta dimensión, en el que cada punto de datos tiene cientos de características. Has aplicado KNN a este conjunto de datos y has observado que el rendimiento de la clasificación es muy sensible al valor de k. Cuando utilizas un valor pequeño de k, el modelo alcanza una gran precisión en los datos de entrenamiento, pero el rendimiento en los datos de validación es bajo. Cuando se utiliza un valor grande de k, el modelo no alcanza una alta precisión ni en los datos de validación, ni en los datos de entrenamiento.

¿Cómo interpretarías este comportamiento del algoritmo KNN? ¿Qué conclusiones puede sacar de esta observación sobre la estructura de los datos y las compensaciones entre sesgo y varianza en el modelo? ¿Cómo utilizaría esta información para seleccionar un valor apropiado de k para este problema?

Logistic Regression Algorithm

1. Supongamos que tenemos un conjunto de datos de calificaciones de estudiantes e información demográfica, y queremos predecir si un estudiante aprobará o no un curso basándonos en las calificaciones de sus exámenes y otros factores. Podríamos utilizar la regresión logística para construir un modelo parecido a éste
$\text{logit}(p) = \beta0 + \beta1X1 + \beta2X2 + \beta3X3 + ... + \betakXk$
donde p es la probabilidad de aprobar el curso, X1 a Xk son las variables independientes (por ejemplo, notas del examen, edad, sexo, etc.), y β0 a βk son los coeficientes que representan el impacto de cada variable independiente en la probabilidad de aprobar.
Supongamos que nuestro modelo tiene los siguientes coeficientes
$\beta0 = -3.5$
$\beta1 = 0.6$
$\beta2 = -0.1$
$\beta3 = 1.2$
Podemos interpretar estos coeficientes de la siguiente manera:
β0 coeficiente asociado a cuando todas las variables independientes son cero.
β1 coeficiente asociado a la nota del examen.
β2 coeficiente asociado a la edad.
β3 coeficiente asociado al sexo

Supongamos que tenemos un conjunto de datos de prueba con 1.000 registros de alumnos y que utilizamos nuestro modelo de regresión logística para hacer predicciones sobre este conjunto de datos.

a) ¿Qué variable influye más en la probabilidad de aprobar el curso y cómo lo sabemos?

b) Imagina que los resultados de la validación cruzada nos dan AUC en entrenamiento: 0,9 y AUC en validación: 0,7.

Dentro del algoritmo de LogisticRegression de python:
¿Cuál es la finalidad del parámetro "C" y cómo afecta al rendimiento y la interpretabilidad del modelo? Si queremos reducir el overfitting en nuestro modelo, ¿qué valor debemos utilizar para "C" y por qué?

c) ¿Por qué la regresión logística es sensible a los outliers?

d) En la clase LogisticRegression de scikit-learn en Python, ¿cuál es la diferencia entre la regularización L1 y L2, y cómo afectan a los coeficientes del modelo?

e) ¿Por qué es importante eliminar las características correlacionadas en la regresión logística?

f) ¿Qué parámetros pondrías dentro de un param_grid de validación cruzada?

LDA Algorithm

1. ¿Qué es el LDA y en qué se diferencia de otros algoritmos de clasificación?

2. En un problema binario, ¿cuál es el número de componentes elegidos por el algoritmo? ¿Por qué? ¿Cuál es la fórmula dentro del algoritmo dentro de scikit-learn?

Naive Bayes Algorithm

1. ¿Cuál es la diferencia entre los clasificadores Gaussian Naive Bayes y Multinomial Naive Bayes en scikit-learn?

2. ¿Cuál es el propósito del Laplace smoothing en Multinomial Naive Bayes?

3. ¿Cuáles son las asunciones para utilizar el algoritmo?

Decision Trees Algorithm

1. ¿Qué técnicas de preprocesamiento de datos son útiles antes de aplicar el algoritmo DecisionTreesClassifier?

2. ¿Cómo se realiza la selección de variables para el árbol de decisiones?

3. ¿Qué es lo que tiene qué ocurrir, en términos de Gini, para que una variable sea la raíz del árbol?

4. ¿Cuál es el mayor problema de un árbol de decisión?

5. ¿Qué podemos hacer para que un árbol de decisión no sufra overfitting?

6. ¿Qué quiere decir que un nodo es puro?

Random Forest Algorithm

1. ¿Con qué propósito fueron creados los Random Forest?

2. ¿Qué es la técnica de Bagging?

3. ¿Cuáles son tres de los hiperparámetros más importantes para ajustar en un modelo de Random Forest Classifier en Python? Explicalos

4. ¿un valor alto de n_estimators puede provocar overfitting?
¿un valor alto de max_depth puede provocar una overfitting?
¿un valor bajo de min_samples_split puede provocar overfitting?
¿un valor alto de max_features puede provocar overfitting?

5. ¿Cómo se puede evaluar la importancia de variables a través de los árboles que componen el Random Forest?

Boosting Algorithm

1. Cuál es la diferencia entre Bagging y Boosting?
2. Nombra algunos algoritmos de Boosting.

Results and interpretability

1. ¿Qué es la precisión (accuracy) y cómo se calcula? ¿Cuál es un ejemplo de su aplicación?

2. ¿Qué es la sensibilidad (recall) y cómo se calcula? ¿Cuál es un ejemplo de su aplicación?

3. ¿Qué es la especificidad (specificity) y cómo se calcula? ¿Cuál es un ejemplo de su aplicación?

4. ¿Qué es el overfitting y el underfitting? ¿Puedes poner un ejemplo de AUC en train/test de cada uno?
5. Supongamos que estamos evaluando un modelo de clasificación binaria que predice si un cliente de un banco solicitará un préstamo o no. El conjunto de datos de prueba consta de 1000 clientes, de los cuales 200 solicitaron un préstamo.
Después de ejecutar el modelo en el conjunto de datos de prueba, obtenemos las siguientes predicciones:
El modelo predijo que 180 clientes solicitarán un préstamo, de los cuales 150 son verdaderos positivos (TP) y 30 son falsos positivos (FP).
El modelo predijo que 820 clientes no solicitarán un préstamo, de los cuales 750 son verdaderos negativos (TN) y 70 son falsos negativos (FN).
Calcula las métricas de evaluación de un modelo