AAI.EX.20250902

Ejercicios elaborados con fines educativos, inspirados en los contenidos evaluados en el exámen del 02/09/2025 (convocatoria Sep-2025) de Aprendizaje Automático 1 del MUICD de la UNED.

Este documento no es una copia ni una transcripción del examen oficial, sino una redacción propia de ejercicios conceptualmente equivalentes.

AAI.EX.20250902.T

Pregunta correcta: +0.5 puntos
Pregunta incorrecta: -0.1 puntos

AAI.EX.20250902.T.1

Enunciado AAI.EX.20250902.T.1 Durante el entrenamiento de un clasificador Naive Bayes, ¿qué cantidades se estiman a partir de los datos?

a) Las probabilidades de cada clase y las probabilidades de cada clase condicionadas por cada característica.

b) Las probabilidades de cada clase y las probabilidades de cada característica condicionadas por cada clase.

c) Solo las probabilidades de cada clase.

d) Las probabilidades de cada clase condicionadas por cada característica.

Solución AAI.EX.20250902.T.1 Respuesta correcta: **b)**

Justificación: Naive Bayes aprende los *priors* $P(y)$ y las *verosimilitudes* $P(x_j \mid y)$, asumiendo independencia condicional entre características dado la clase.

AAI.EX.20250902.T.2

Enunciado AAI.EX.20250902.T.2 ¿Cuál de las siguientes afirmaciones sobre DBSCAN no es correcta?

a) Tiene noción explícita de ruido y puede detectar outliers.

b) Requiere únicamente dos parámetros, pero es sensible al orden de los puntos en el dataset.

c) No necesita especificar el número de clusters por adelantado.

d) Puede encontrar clusters con formas arbitrarias.

Solución AAI.EX.20250902.T.2 Respuesta correcta: **b)**

Justificación: DBSCAN depende de dos parámetros principales (ε y `min_samples`), pero **no es sensible al orden** de los puntos; el resultado no cambia por el orden del dataset.

AAI.EX.20250902.T.3

Enunciado AAI.EX.20250902.T.3 En una clasificación multiclase con muchas variables numéricas y categóricas (ya preprocesadas), si el objetivo principal es que el modelo sea interpretable por humanos, ¿qué tipo de algoritmo es más adecuado?

a) SVM.

b) Árboles de decisión.

c) Naive Bayes.

d) Redes neuronales.

Solución AAI.EX.20250902.T.3 Respuesta correcta: **b)**

Justificación: los árboles de decisión producen reglas claras tipo "si–entonces", lo que los hace especialmente interpretables frente a modelos de caja negra.

AAI.EX.20250902.T.4

Enunciado AAI.EX.20250902.T.4 ¿Por qué es importante realizar selección de características antes de entrenar una SVM?

a) Para aumentar el tamaño del dataset.

b) Para incrementar la complejidad del modelo.

c) Para reducir dimensionalidad y mejorar eficiencia y generalización.

d) Para asegurar que todas las variables se utilicen.

Solución AAI.EX.20250902.T.4 Respuesta correcta: **c)**

Justificación: reducir características irrelevantes o redundantes disminuye el coste computacional y puede mejorar la generalización de la SVM.

AAI.EX.20250902.T.5

Enunciado AAI.EX.20250902.T.5 En un MLP usado para regresión, cuando se desean salidas reales sin restricciones, ¿qué activación se emplea en la capa de salida?

a) ReLU.

b) Sigmoide.

c) Softmax.

d) Ninguna función de activación.

Solución AAI.EX.20250902.T.5 Respuesta correcta: **d)**

Justificación: una salida lineal (sin activación) permite predecir cualquier valor real sin acotaciones artificiales.

AAI.EX.20250902.T.6

Enunciado AAI.EX.20250902.T.6 ¿Cuál es una ventaja conocida de las SVM en aplicaciones como la clasificación de imágenes médicas?

a) Funcionan sin ningún preprocesamiento.

b) Son fáciles de interpretar visualmente.

c) Son robustas al sobreajuste con muchos parámetros.

d) Funcionan bien incluso con pocos datos de entrenamiento.

Solución AAI.EX.20250902.T.6 Respuesta correcta: **d)**

Justificación: las SVM suelen generalizar bien en espacios de alta dimensión incluso cuando el número de muestras es limitado.

AAI.EX.20250902.T.7

Enunciado AAI.EX.20250902.T.7 En Aprendizaje Automático, ¿cómo se interpreta la entropía cuando se usa como medida de impureza?

a) Es cero cuando el conjunto contiene menos información que otros.

b) Es cero cuando hay el mismo número de instancias de cada clase.

c) Es cero cuando todas las instancias pertenecen a una sola clase.

d) Es cero cuando los conjuntos son idénticos.

Solución AAI.EX.20250902.T.7 Respuesta correcta: **c)**

Justificación: la entropía mide desorden; alcanza su valor mínimo (0) cuando el conjunto es completamente puro (una sola clase).

AAI.EX.20250902.T.8

Enunciado AAI.EX.20250902.T.8 ¿Cuál es el objetivo de la detección de anomalías?

a) Detectar anomalías solo en el conjunto de entrenamiento.

b) Detectar anomalías solo en nuevas instancias.

c) Detectar anomalías tanto en el entrenamiento como en nuevas instancias.

d) Detectar anomalías solo en el conjunto de validación.

Solución AAI.EX.20250902.T.8 Respuesta correcta: **c)**

Justificación: en detección de anomalías se asume que pueden existir valores atípicos tanto en los datos históricos como en los datos futuros.

AAI.EX.20250902.T.9

Enunciado AAI.EX.20250902.T.9 En una neurona de una capa oculta se obtiene una salida cercana a 0.1 para cualquier entrada. ¿Qué función de activación podría estar usándose?

a) ReLU.

b) tanh.

c) Sigmoide.

d) Ninguna de ellas.

Solución AAI.EX.20250902.T.9 Respuesta correcta: **b)**

Justificación: la función **tanh** está centrada en 0 y produce valores pequeños positivos o negativos alrededor de ese punto; la sigmoide suele producir valores cercanos a 0.5 para entradas pequeñas.

AAI.EX.20250902.T.10

Enunciado AAI.EX.20250902.T.10 ¿Cuáles son motivaciones habituales para aplicar reducción de dimensionalidad?

a) Acelerar y mejorar algoritmos posteriores.

b) Incrementar la complejidad computacional.

c) Facilitar interpretación y visualización de los datos transformados.

d) a) y c).

Solución AAI.EX.20250902.T.10 Respuesta correcta: **d)**

Justificación: reducir dimensionalidad puede mejorar eficiencia, reducir ruido y facilitar interpretación, pero no busca aumentar complejidad.

AAI.EX.20250902.D

AAI.EX.20250902.D.1

Puntuación máxima: 5 puntos Extensión máxima orientativa: 2 caras

Enunciado AAI.EX.20250902.D.1
Formas parte del área de datos de una cadena de supermercados con presencia nacional. Se ha implantado un sistema de recogida homogénea de información diaria por tienda, de modo que para cada establecimiento se dispone de:

- un identificador de tienda,
- la ciudad donde se ubica,
- la fecha,
- el número de cámaras/refrigeradores instalados,
- la superficie del local (m^2),
- el número total de compras registradas (tickets),
- el consumo eléctrico total del día (kW).

Dado que los registros son diarios, las variables que cambian con el tiempo son la fecha, los tickets y el consumo. Además, se cuenta con datos meteorológicos por ciudad y fecha (temperatura mínima y máxima) obtenidos de un servicio oficial, y el histórico disponible cubre el año 2024 y parte del 2025. El servicio meteorológico puede proporcionar también previsiones de temperatura mínima y máxima para los próximos 7 días en cada ciudad.

El objetivo es diseñar una metodología para predecir el consumo eléctrico diario por tienda con un horizonte de hasta una semana. Al revisar trabajos previos sobre problemas similares, se observa que dos enfoques habituales son los modelos basados en árboles de decisión y las redes neuronales.

Responda razonadamente:

1. Si se opta únicamente por modelos basados en árboles de decisión, ¿cómo plantearía el sistema de predicción?, ¿qué rasgos tendría el modelo y cómo lo mejoraría mediante un proceso de ajuste/refinamiento?
2. Si se opta únicamente por redes neuronales, ¿cómo plantearía el sistema?, ¿qué tipo de arquitectura propondría y cómo sería la capa de salida en un pronóstico a 7 días?
3. Dado que el objetivo original es numérico (regresión), proponga una manera de reformularlo como un problema de clasificación y explique qué modificaciones habría que introducir en los enfoques de los puntos 1 y 2.

Solución AAI.EX.20250902.D.1

Planteamiento común del problema:

Se trata de un problema de predicción temporal (*forecasting*) del consumo eléctrico diario por tienda, con un horizonte máximo de 7 días. La unidad de predicción es tienda–día y la variable objetivo es el consumo eléctrico en kW. El entrenamiento se realiza con datos de 2024 y parte de 2025, respetando siempre el orden temporal para evitar fugas de información.

Las variables disponibles pueden dividirse en:

- Variables estáticas por tienda: ciudad, metros cuadrados, número de refrigeradores.
- Variables dinámicas diarias: fecha, número de tickets, consumo eléctrico.
- Variables meteorológicas: temperatura mínima y máxima por ciudad y día, con disponibilidad de pronóstico para los próximos 7 días.

Para evaluar los modelos es necesario usar validación temporal, por ejemplo entrenando con un bloque inicial y validando con un bloque posterior. Las métricas adecuadas incluyen MAE y RMSE.

1) Uso exclusivo de árboles de decisión:

Proceso de modelado:

El primer paso consiste en construir un conjunto de datos supervisado. Para cada tienda y cada día se generan características explicativas a partir de la información histórica:

- Variables de calendario: día de la semana, mes, indicadores de fin de semana y festivos.

- Variables retardadas (*lags*): consumo eléctrico en días anteriores como $t-1$, $t-7$ o $t-14$, que permiten capturar inercia y estacionalidad semanal.
- Variables agregadas: medias móviles, medianas o máximos del consumo en ventanas temporales (por ejemplo, últimos 7 o 28 días).
- Variables relacionadas con los tickets: valores retardados y agregados históricos, evitando usar valores futuros no disponibles.
- Variables meteorológicas: temperaturas históricas y, de forma clave, temperatura mínima y máxima pronosticadas para el día objetivo $t+h$.
- Variables estáticas: metros cuadrados y número de refrigeradores.
- Variable de ciudad codificada como categórica.

El objetivo puede definirse de dos formas: entrenar un modelo por cada horizonte (1 a 7 días) o entrenar un único modelo con múltiples salidas, una por cada día futuro.

Características del modelo:

Aunque se pueden usar árboles de decisión simples, estos tienden a sobreajustar. En la práctica, se emplean conjuntos de árboles:

- Random Forest, que reduce la varianza mediante el promedio de múltiples árboles.
- Gradient Boosting con árboles, que construye el modelo de forma secuencial corrigiendo errores anteriores.

Estos modelos manejan bien relaciones no lineales, interacciones entre variables y no requieren escalado de características. Además, permiten cierto grado de interpretabilidad mediante importancias de variables.

Proceso de refinamiento:

El refinamiento del modelo incluye:

- Control de la complejidad mediante hiperparámetros como profundidad máxima, número mínimo de muestras por hoja y número de árboles.
- Ajuste de la tasa de aprendizaje y uso de *early stopping* en modelos de boosting.
- Evaluación mediante validación temporal con ventanas deslizantes.
- Análisis de errores por tienda, ciudad o estación del año.
- Tratamiento de valores atípicos en el consumo eléctrico.

2) Uso exclusivo de redes neuronales:

Proceso de modelado:

Existen dos enfoques principales. El primero consiste en tratar el problema como tabular, usando una red neuronal multicapa (MLP). En este caso se utilizan las mismas características que con los árboles (lags, agregados, calendario, meteorología y variables estáticas), aplicando normalización a las variables numéricas.

Las variables categóricas, como la ciudad o el identificador de tienda, pueden codificarse mediante *one-hot encoding* o mediante embeddings.

El segundo enfoque es secuencial. Para cada tienda se construyen ventanas temporales de longitud fija (por ejemplo, 28 días) que contienen:

- Secuencias históricas de consumo y tickets.
- Información meteorológica histórica.
- Variables estáticas replicadas en el tiempo.
- Variables meteorológicas futuras conocidas para los próximos 7 días.

Sobre estas secuencias pueden entrenarse redes convolucionales temporales o redes recurrentes como LSTM o GRU.

Características de la red:

Las redes neuronales suelen incluir:

- Varias capas ocultas con activación ReLU.
- Regularización mediante dropout y penalización L2.
- Optimización con Adam.
- Función de pérdida MAE o Huber para reducir la sensibilidad a valores extremos.
- *Early stopping* basado en el error de validación.

Neuronas de salida:

Al tratarse de un problema de regresión:

- Para un único horizonte se utiliza una sola neurona de salida con activación lineal.
- Para un horizonte de 7 días se utilizan 7 neuronas de salida, una por cada día futuro, todas con activación lineal. La pérdida total se calcula agregando el error de cada horizonte.

3) Conversión del problema de regresión a clasificación:

Una forma directa de transformar el problema es discretizar el consumo eléctrico en categorías. Por ejemplo, se pueden definir clases basadas en cuantiles del consumo histórico, como consumo bajo, medio y alto. De este modo, el objetivo pasa a ser predecir la clase de consumo en lugar del valor exacto.

Cambios en los modelos basados en árboles:

Los modelos pasan de ser regresores a clasificadores. Se utilizan árboles de clasificación, Random Forest Classifier o Gradient Boosting Classifier. El criterio de división pasa a ser Gini o entropía, y las métricas de evaluación incluyen F1 macro, matriz de confusión y *balanced accuracy*. Las predicciones pueden expresarse como probabilidades por clase.

Cambios en las redes neuronales:

La capa de salida se modifica para tener tantas neuronas como clases definidas. Se utiliza activación softmax y la función de pérdida pasa a ser entropía cruzada. Para un horizonte de 7 días se pueden definir salidas independientes por horizonte. Las métricas habituales incluyen accuracy y F1 macro.

Este enfoque es especialmente útil cuando las decisiones de negocio se basan en rangos de consumo en lugar de valores exactos.