AAI.EX.20240902

Ejercicios elaborados con fines educativos, inspirados en los contenidos evaluados en el exámen del 02/09/2024 (convocatoria Sep-2024) 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.

Test

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

AAI.EX.20240902.T.1

Enunciado AAI.EX.20240902.T.1
Se entrena una SVM con kernel RBF y se observa subajuste incluso en el conjunto de entrenamiento. ¿Qué ajuste suele ser el más adecuado para aumentar la capacidad del modelo?

a) Aumentar γ y/o C.

b) Disminuir γ y/o C.

c) Aumentar γ y/o disminuir C.

d) Disminuir γ y/o aumentar C.

Solución AAI.EX.20240902.T.1
Respuesta correcta: **a)**

Justificación: aumentar γ permite fronteras más complejas y aumentar C reduce la regularización, ambos incrementan la capacidad del modelo y ayudan a combatir el subajuste.

AAI.EX.20240902.T.2

Enunciado AAI.EX.20240902.T.2
Algunos clasificadores pueden asignar una instancia a varias clases simultáneamente. ¿En qué escenario ocurre esto?

a) En situaciones de sobreajuste.

b) En clasificadores binarios.

c) Cuando la exactitud por clase es del 50%.

d) En clasificadores multi-etiqueta.

Solución AAI.EX.20240902.T.2
Respuesta correcta: **d)**

Justificación: en **clasificación multi-etiqueta** una misma instancia puede pertenecer a varias clases a la vez.

AAI.EX.20240902.T.3

Enunciado AAI.EX.20240902.T.3
En una SVM, ¿qué hiperparámetro controla la penalización por errores de clasificación?

a) γ.

b) α.

c) β.

d) C.

Solución AAI.EX.20240902.T.3
Respuesta correcta: **d)**

Justificación: C controla el compromiso entre maximizar el margen y penalizar los errores de clasificación.

AAI.EX.20240902.T.4

Enunciado AAI.EX.20240902.T.4
¿Cuál de las siguientes afirmaciones sobre SVM es incorrecta?

a) SVM clasifica más de dos clases de forma nativa.

b) SVM admite clasificación y regresión, tanto lineal como no lineal.

c) SVM suele funcionar bien en problemas complejos de tamaño pequeño o mediano.

d) SVM es sensible a la escala de las variables.

Solución AAI.EX.20240902.T.4
Respuesta correcta: **a)**

Justificación: las SVM son **intrínsecamente binarias**; la multiclase se logra mediante estrategias como one-vs-rest u one-vs-one.

AAI.EX.20240902.T.5

Enunciado AAI.EX.20240902.T.5
¿Cuál de las siguientes aplicaciones no corresponde al uso de algoritmos de clustering?

a) Segmentación de clientes.

b) Organización de resultados en motores de búsqueda.

c) Detección de anomalías y novedades.

d) Clasificación de datos con etiquetas conocidas.

Solución AAI.EX.20240902.T.5
Respuesta correcta: **d)**

Justificación: clasificar datos etiquetados es un problema **supervisado**, no de clustering.

AAI.EX.20240902.T.6

Enunciado AAI.EX.20240902.T.6
¿Cuál de los siguientes métodos pertenece al aprendizaje semi-supervisado?

a) Clustering jerárquico aglomerativo.

b) BIRCH.

c) Propagación de etiquetas.

d) Mean-Shift.

Solución AAI.EX.20240902.T.6
Respuesta correcta: **c)**

Justificación: la **propagación de etiquetas** utiliza pocos datos etiquetados y muchos no etiquetados.

AAI.EX.20240902.T.7

Enunciado AAI.EX.20240902.T.7
¿Cuál de las siguientes funciones se emplea habitualmente como activación en neuronas artificiales?

a) Función lineal.

b) Función gaussiana.

c) Función tangente hiperbólica.

d) Ninguna de las anteriores.

Solución AAI.EX.20240902.T.7
Respuesta correcta: **c)**

Justificación: la **tangente hiperbólica** es una función de activación clásica, especialmente en capas ocultas.

AAI.EX.20240902.T.8

Enunciado AAI.EX.20240902.T.8
¿Cuál de las siguientes **no** es una característica de las SVM?

a) Pueden manejar múltiples variables predictoras.

b) Solo permiten fronteras de decisión lineales.

c) Pueden manejar datos no perfectamente separables.

d) Pueden abordar problemas multiclase mediante estrategias estándar.

Solución AAI.EX.20240902.T.8
Respuesta correcta: **b)**

Justificación: gracias a los **kernels**, las SVM pueden aprender fronteras **no lineales**.

AAI.EX.20240902.T.9

Enunciado AAI.EX.20240902.T.9
¿Qué tipo de forma y variabilidad pueden presentar los clusters en un Modelo de Mezcla Gaussiana (GMM)?

a) Circular y de tamaños similares.

b) No se asume ninguna estructura.

c) Elipsoidal con tamaños y densidades similares.

d) Elipsoidal con tamaños y densidades potencialmente diferentes.

Solución AAI.EX.20240902.T.9
Respuesta correcta: **d)**

Justificación: los GMM modelan cada cluster como una gaussiana con su propia media y covarianza, permitiendo **elipses** de distintos tamaños y densidades.

AAI.EX.20240902.T.10

Enunciado AAI.EX.20240902.T.10
En un MLP con 10 entradas, una capa oculta de 50 neuronas y una capa de salida de 3 neuronas, ¿qué dimensiones tiene la matriz de salida Y?

a) 3×50.

b) $M \times 3$, siendo M el tamaño del lote.

c) $3 \times M$, siendo M el tamaño del lote.

d) 3×10.

Solución AAI.EX.20240902.T.10
Respuesta correcta: **b)**

Justificación: cada fila de Y corresponde a una instancia del lote y cada columna a una neurona de salida, por lo que $Y \in \mathbb{R}^{M \times 3}$.

AAI.EX.20240902.D

AAI.EX.20240902.D.1

- Puntuación máxima: 5 puntos
- Extensión máxima orientativa: 2 caras

Enunciado AAI.EX.20240902.D.1
Durante las últimas semanas se ha detectado un incremento de incidentes delictivos en horario nocturno en determinadas estaciones de la red de metro de una

gran ciudad. Los hechos presentan un patrón común y se concentran en un grupo reducido de estaciones situadas en la zona centro.

El equipo de análisis de una unidad policial dispone de información diaria correspondiente a los últimos 30 días. Para cada día y para cada una de las 10 estaciones consideradas se registran:

- el número de viajeros que acceden a la estación,
- el número de viajeros que abandonan la estación,
- el número de trenes que circulan por la estación,
- un indicador binario que señala si se produjo o no algún delito ese día en la estación.

Además, se cuenta con una estimación actual de las variables de afluencia y circulación de trenes para el día de hoy en esas mismas estaciones.

El objetivo es diseñar un sistema basado en aprendizaje automático que permita estimar la probabilidad de que ocurra un delito en cada estación en un día determinado, utilizando la información histórica disponible.

Responda de forma razonada a las siguientes cuestiones:

- ¿Qué tipos de algoritmos de aprendizaje automático serían adecuados para abordar este problema? Proponga uno y explique su funcionamiento.
- Describa un proceso completo de aprendizaje automático, identificando las principales etapas desde la preparación de los datos hasta la obtención de las predicciones.
- ¿Es necesario ajustar parámetros o hiperparámetros en el modelo propuesto? Indique cuáles y explique su influencia en el rendimiento del sistema.

Solución AAI.EX.20240902.D.1

Planteamiento de la solución:

Se disponen de las siguientes variables para cada combinación de los 30 días y las 10 estaciones:

- Número de entradas (`Num-E`), var. explicativa
- Número de salidas (`Num-S`), , var. explicativa
- Número de trenes (`Num-T`), , var. explicativa.
- Si ese día ocurrió o no al menos un delito en la estación (`D`), variable binaria D.

El objetivo es, a partir de estos datos históricos y de los valores estimados para el día actual (`Num-E`, `Num-S` y `Num-T`), predecir la **probabilidad de que ocurra un delito** en cada una de las 10 estaciones.

Este es un problema de clasificación binaria con dependencia temporal implícita y con un conjunto de datos relativamente pequeño (30 días $\times$ 10 estaciones = 300 observaciones).

Modelo elegido:

Se propone utilizar un **árbol de decisión** para clasificación. Este modelo es adecuado porque:

- maneja relaciones no lineales entre las variables
- no requiere escalado
- funciona bien con pocos datos
- permite obtener probabilidades de clase
- es eficiente en entrenamiento y predicción

Funcionamiento del modelo:

El árbol aprende reglas del tipo:

- si Num-E $>$ umbral y Num-T $<$ umbral
- si Num-S $<$ umbral

Cada hoja del árbol contiene observaciones con comportamiento similar. La probabilidad de delito se estima como:

$$P(D = 1 \mid x) = \frac{\text{nº de observaciones con delito en la hoja}}{\text{nº total de observaciones en la hoja}}$$

Proceso de aprendizaje automático:

1) Definición del dataset supervisado
   Cada observación corresponde a una pareja estación–día.
   Entradas: Num-E, Num-S, Num-T.
   Salida: D.

2) Preprocesamiento

   - revisión de valores faltantes o inconsistentes
   - no es necesario escalado
   - codificación de la estación si se incluye como variable adicional

3) División entrenamiento–validación
   Se respeta el orden temporal:

   - entrenamiento con los primeros días
   - validación con los últimos días

4) Entrenamiento
   Se ajusta el árbol minimizando la impureza (índice Gini o entropía).

5) Evaluación
   Se utilizan métricas de clasificación como:
   - accuracy
   - precision, recall y F1
   - AUC-ROC si se evalúan probabilidades
6) Predicción operativa
   Con los valores del día actual, el modelo produce para cada estación una probabilidad $P(D = 1 \mid x)$, que puede utilizarse para priorizar vigilancia.

Ajuste de hiperparámetros:

Para evitar sobreajuste se controlan:

- profundidad máxima del árbol
- número mínimo de muestras por hoja
- número mínimo de muestras para dividir un nodo

Conclusión:

Un árbol de decisión permite estimar de forma eficiente la probabilidad de delito por estación y día, utilizando únicamente técnicas del temario y adaptándose al tamaño reducido del conjunto de datos.

Solución alternativa: Regresión logística:

Como solución alternativa, se podía haber propuesto la regresión la **regresión logística** . No obstante, no entra dentro del temario de la asignatura por lo que no parece la respuesta que el profesor esperara.

No obstante esta solución es también adecuada ya que:

- Produce directamente probabilidades.
- Es interpretable, algo importante en un contexto policial.
- Funciona bien con pocos datos si se regulariza adecuadamente.
- Permite entender cómo influyen Num-E, Num-S y Num-T en el riesgo de delito.

Funcionamiento de la regresión logística:

La regresión logística modela la probabilidad de que ocurra un delito como:

$$P(D = 1 \mid x) = \sigma(z) = \frac{1}{1 + e^{-z}}$$

donde:

$$z = \beta_0 + \beta_1 \,\text{Num-E} + \beta_2 \,\text{Num-S} + \beta_3 \,\text{Num-T}$$

Los coeficientes β se aprenden maximizando la verosimilitud (o, equivalentemente, minimizando la log-loss). El resultado es una probabilidad entre 0 y 1 que puede interpretarse como riesgo estimado de delito en la estación ese día.

Proceso de aprendizaje automático propuesto:

1) **Definición del dataset supervisado**

   Cada observación corresponde a una pareja estación–día. Las variables de entrada son Num-E, Num-S y Num-T, y la variable objetivo es D.

   Opcionalmente, se pueden añadir:

   - Identificador de estación (como variable categórica codificada).
   - Variables temporales como día de la semana, si se considera que el patrón nocturno varía según el día.

2) **Preprocesamiento de datos**:

   - Revisión de valores faltantes o inconsistentes.
   - Escalado de las variables Num-E, Num-S y Num-T, ya que pueden tener órdenes de magnitud distintos.
   - Codificación de la estación si se incluye como variable categórica.

3) **División entrenamiento–validación**:

   Dado el carácter temporal de los datos, la división debe respetar el orden cronológico:

   - Entrenar con los primeros días (por ejemplo, días 1 a 24).
   - Validar con los últimos días (por ejemplo, días 25 a 30).

   Esto evita utilizar información futura para predecir el pasado.

4) **Entrenamiento del modelo**

   Se ajusta la regresión logística usando los datos de entrenamiento. El modelo aprende los coeficientes que mejor explican la ocurrencia de delitos en función de las variables de entrada.

5) **Evaluación del modelo**

   Se evalúa el rendimiento en el conjunto de validación usando métricas adecuadas para clasificación:

   - Log-loss (adecuada para probabilidades).
   - AUC-ROC.
   - Precisión, recall o F1 si interesa priorizar la detección de estaciones peligrosas.

6) **Predicción operativa**

   Con los valores estimados de hoy (Num-E, Num-S, Num-T), el modelo genera para cada estación una probabilidad de que ocurra un delito. Es-

tas probabilidades pueden usarse para priorizar vigilancia o despliegue policial.

Afinado de hiperparámetros:

Sí, es necesario afinar algunos hiperparámetros, especialmente debido al tamaño limitado del dataset.

1) **Fuerza de regularización**

   La regresión logística suele incluir regularización para evitar sobreajuste. El hiperparámetro controla cuánto se penalizan los coeficientes grandes.

   - Regularización L2 (ridge): penaliza el cuadrado de los coeficientes y tiende a repartir el peso entre variables.
   - Regularización L1 (lasso): puede llevar algunos coeficientes a cero, actuando como selección de variables.

   El parámetro asociado (por ejemplo, C en muchas implementaciones) controla la intensidad de la regularización: valores pequeños implican regularización fuerte.

2) **Tipo de regularización**

   Elegir entre L1, L2 o una combinación (elastic net) afecta tanto a la estabilidad del modelo como a su interpretabilidad. Con pocos datos, L2 suele ser una opción segura.

3) **Umbral de decisión**

   Aunque el modelo produce probabilidades, la decisión final de "riesgo alto" o "riesgo bajo" depende de un umbral. Ajustar este umbral permite priorizar:

   - Recall alto (detectar la mayoría de estaciones con delito).
   - Precisión alta (reducir falsas alarmas).

   Este ajuste no cambia el modelo, pero sí su uso operativo.

4) **Inclusión o no de variables adicionales**

   No es un hiperparámetro formal, pero decidir si se incluye la estación como variable categórica o si se añaden variables temporales afecta de forma directa al rendimiento y debe validarse empíricamente.

   En conjunto, un modelo de clasificación supervisada como la regresión logística, bien regularizado y validado temporalmente, permite estimar probabilidades de delito por estación de forma interpretable y adecuada al tamaño y naturaleza del conjunto de datos disponible.