AAI.EX.20220905

Ejercicios elaborados con fines educativos, inspirados en los contenidos evaluados en el exámen del 05/09/2022 (convocatoria Sep-2022) 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.20220905.T.1

Enunciado AAI.EX.20220905.T.1
Al aplicar K-means, existen técnicas prácticas para estimar un número razonable de clusters. ¿Qué enfoques se usan habitualmente?

a) Las opciones b) y d) son correctas.

b) Buscar un extremo (máximo o mínimo) en la puntuación de la silueta.

c) No existen métodos aproximados para elegir k.

d) Identificar el punto donde la inercia deja de decrecer bruscamente (método del codo).

Solución AAI.EX.20220905.T.1
Respuesta correcta: **a)**

Justificación: tanto el método del codo como la puntuación de silueta son heurísticas habituales para estimar un valor razonable de k en K-means.

AAI.EX.20220905.T.2

Enunciado AAI.EX.20220905.T.2
¿Cuál describe correctamente la regla de actualización básica del Perceptrón durante el entrenamiento?

a) Cuando una salida es incorrecta, se refuerzan los pesos asociados a entradas que habrían favorecido la predicción correcta.

b) Cuando una salida es incorrecta, se debilitan los pesos asociados a entradas que habrían favorecido la predicción correcta.

c) Ante un error, los pesos se multiplican por 1.

d) Cuando una salida es correcta, se refuerzan los pesos que habrían producido un error.

Solución AAI.EX.20220905.T.2
Respuesta correcta: **a)**

Justificación: el Perceptrón ajusta los pesos en la dirección de la clase correcta cuando comete un error, reforzando las contribuciones que empujan hacia la predicción adecuada.

AAI.EX.20220905.T.3

Enunciado AAI.EX.20220905.T.3
Un modelo obtiene un rendimiento excelente en entrenamiento pero falla al predecir nuevas instancias. ¿Cuál es la causa más probable?

a) Todas las anteriores.

b) Datos de mala calidad.

c) Sobreajuste.

d) Subajuste.

Solución AAI.EX.20220905.T.3
Respuesta correcta: **c)**

Justificación: este comportamiento es característico del sobreajuste, donde el modelo memoriza el entrenamiento pero no generaliza bien.

AAI.EX.20220905.T.4

Enunciado AAI.EX.20220905.T.4
¿Por qué la función logística fue clave para entrenar los primeros MLP mediante retropropagación?

a) Porque es derivable.

b) Porque su derivada nunca es cero.

c) Porque su derivada está acotada entre 0 y 1.

d) Porque su derivada es siempre positiva.

Solución AAI.EX.20220905.T.4
Respuesta correcta: **a)**

Justificación: la retropropagación requiere funciones derivables para poder calcular gradientes para poder calcular la pendiente y por tanto minimizar la función de pérdida. La función de activación sigmoide o logística permitió entrenar redes multicapa de forma eficiente.

AAI.EX.20220905.T.5

Enunciado AAI.EX.20220905.T.5
¿Cuál de las siguientes tareas no pertenece al aprendizaje no supervisado?

a) Descubrimiento de reglas de asociación.

b) Agrupamiento.

c) Reducción de dimensionalidad.

d) Regresión.

Solución AAI.EX.20220905.T.5
Respuesta correcta: **d)**

Justificación: la **regresión** requiere etiquetas y es una tarea de aprendizaje supervisado; las demás son no supervisadas.

AAI.EX.20220905.T.6

Enunciado AAI.EX.20220905.T.6
Respecto a las SVM, ¿qué afirmación es correcta?

a) Los vectores soporte son las instancias situadas en los bordes de la "calle" de separación y fuera de ella.

b) En el caso separable, los vectores soporte están dentro de la "calle".

c) Ninguna de las otras opciones es correcta.

d) Los vectores soporte son las instancias situadas exactamente en los bordes de la "calle" de separación.

Solución AAI.EX.20220905.T.6
Respuesta correcta: **d)**

Justificación: los **vectores soporte** son los puntos de entrenamiento que definen el margen, situados en los **bordes del margen máximo**.

AAI.EX.20220905.T.7

Enunciado AAI.EX.20220905.T.7
¿Qué par de algoritmos de clustering puede escalar razonablemente bien a grandes volúmenes de datos?

a) K-means y Mean-Shift.

b) K-means y BIRCH.

c) DBSCAN y Mean-Shift.

d) BIRCH y DBSCAN.

Solución AAI.EX.20220905.T.7
Respuesta correcta: **b)**

Justificación: **K-means** y **BIRCH** están diseñados para escalar a grandes conjuntos; Mean-Shift y DBSCAN suelen tener problemas de escalabilidad.

AAI.EX.20220905.T.8

Enunciado AAI.EX.20220905.T.8
¿Cuál de las siguientes funciones se emplea habitualmente como función de activación en neuronas artificiales?

a) Función gaussiana.

b) Función lineal.

c) Ninguna de las anteriores.

d) Función tangente hiperbólica.

Solución AAI.EX.20220905.T.8
Respuesta correcta: **d)**

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

AAI.EX.20220905.T.9

Enunciado AAI.EX.20220905.T.9
En regresión con SVM, ¿qué hiperparámetro controla el ancho del "tubo" alrededor de la función de regresión?

a) C.

b) ε.

c) β.

d) γ.

Solución AAI.EX.20220905.T.9
Respuesta correcta: **b)**

Justificación: ε define la zona de insensibilidad alrededor de la función de regresión en SVR, determinando el ancho del "tubo" dentro del cual los errores no se penalizan durante el entrenamiento. El hiperparámetro C, aunque también existe en SVR, controla únicamente la penalización de los errores que quedan fuera de dicho tubo, no su anchura.

AAI.EX.20220905.T.10

Enunciado AAI.EX.20220905.T.10
¿Puede un clasificador SVM proporcionar una medida de confianza al clasificar una instancia?

a) Sí, usando la media de las distancias a los vectores soporte.

b) Sí, usando la distancia al centroide de la clase.

c) Sí, usando la distancia a la frontera de decisión.

d) No, nunca puede proporcionar confianza.

Solución AAI.EX.20220905.T.10
Respuesta correcta: **c)**

Justificación: la **distancia a la frontera de decisión** es una medida natural de confianza en SVM (aunque no es una probabilidad).

Preguntas de desarrollo

AAI.EX.20220905.D

AAI.EX.20220905.D.1

Puntuación máxima: 5 puntos Extensión máxima orientativa: 2 caras

Enunciado AAI.EX.20220905.D.1
Una empresa del ámbito sanitario dispone de una base de datos extensa con información clínica de pacientes atendidos en un hospital especializado. El conjunto de datos incluye registros tanto de personas diagnosticadas con una determinada patología como de personas sin dicho diagnóstico. Para cada paciente se almacenan más de un centenar de variables, que recogen información demográfica, antecedentes médicos personales y familiares, así como resultados de distintas pruebas clínicas.

El equipo de análisis de datos tiene como objetivo desarrollar un sistema que permita **identificar pacientes cuyo perfil sea claramente distinto del comportamiento habitual del conjunto de pacientes**, con el fin de detectar casos atípicos que requieran una revisión más detallada.

Responda de forma razonada a las siguientes cuestiones:

- ¿Es más adecuado abordar este problema mediante técnicas de aprendizaje supervisado o no supervisado? Justifique la elección.
- ¿Qué tipos de algoritmos serían apropiados para este objetivo? Proponga uno y explique con detalle su funcionamiento.
- ¿Es necesario ajustar parámetros o hiperparámetros en el modelo propuesto? Indique cuáles y explique su papel en el comportamiento del algoritmo.

Solución AAI.EX.20220905.D.1

Planteamiento del problema:

Se dispone de un gran conjunto de datos clínicos con más de 100 características por paciente. El objetivo no es predecir directamente la enfermedad, sino **detectar pacientes con características muy diferentes al resto**, es decir, identificar **casos anómalos o atípicos** dentro de la población.

El problema se centra en encontrar patrones inusuales, no necesariamente en clasificar pacientes como enfermos o no.

¿Aprendizaje supervisado o no supervisado?

El enfoque más adecuado es el **aprendizaje no supervisado**.

Razones:

- El objetivo no es predecir una etiqueta conocida, sino **detectar desviaciones respecto al comportamiento general**.
- Aunque se conozca si un paciente tiene la enfermedad, esa información no define qué es "anómalo", ya que:
  - puede haber pacientes enfermos con perfiles comunes,
  - o pacientes no enfermos con perfiles muy inusuales.
- En problemas de detección de anomalías, las clases suelen estar mal definidas o muy desbalanceadas.

Por tanto, el aprendizaje no supervisado permite identificar pacientes raros sin imponer una definición previa de normalidad basada en etiquetas.

Tipo de algoritmo propuesto y funcionamiento:

Un algoritmo adecuado es **Isolation Forest**.

Funcionamiento:

- Isolation Forest se basa en la idea de que las anomalías son más fáciles de aislar que los puntos normales.
- Construye múltiples árboles de decisión de forma aleatoria:
  - en cada nodo se selecciona aleatoriamente una característica,
  - y un punto de corte aleatorio dentro de su rango.
- Los datos se dividen recursivamente hasta aislar cada observación.

Idea clave:

- Los pacientes "normales" suelen necesitar muchos cortes para aislarse.
- Los pacientes anómalos se aíslan con pocos cortes, porque están en regiones poco densas del espacio de características.

El modelo asigna a cada paciente una **puntuación de anomalía**, que permite:

- detectar individuos muy distintos al resto,
- o establecer un ranking de rareza clínica.

Ventajas en este contexto:

- Escala bien con muchas variables.
- No requiere normalidad ni distribución concreta.
- Funciona bien en espacios de alta dimensión.
- No necesita etiquetas.

Hiperparámetros a afinar:

Sí, es importante ajustar algunos hiperparámetros:

1) **Número de árboles (n_estimators)**:
   - Define cuántos árboles componen el bosque.
   - Más árboles → estimaciones más estables, pero mayor coste computacional.
   - Valores típicos: 100–300.

2) **Tamaño de la muestra por árbol (max_samples)**:

- Número de observaciones usadas para entrenar cada árbol.
- Valores pequeños favorecen la detección de anomalías locales.
- Puede fijarse como número absoluto o proporción del total.

3) **Número de características consideradas (max_features)**

- Número de variables candidatas en cada división.
- Con muchas variables clínicas, limitar este valor puede reducir ruido y correlación.

4) **Proporción esperada de anomalías (contamination)**

- Estima qué fracción de pacientes se espera que sea anómala.
- Determina el umbral para clasificar un paciente como normal o anómalo.
- Es un parámetro crítico y suele fijarse con conocimiento clínico o exploración previa.

5) **Preprocesamiento previo**

- Aunque no es un hiperparámetro del modelo, es clave:
- normalizar o estandarizar variables,
- tratar valores faltantes,
- eliminar variables redundantes o altamente correlacionadas.

En conjunto, un enfoque no supervisado basado en Isolation Forest permite detectar pacientes con perfiles clínicos inusuales de forma robusta, interpretable a nivel de puntuación de riesgo y adecuada para datos clínicos complejos y de alta dimensión.