

TÉCNICAS DE INTELIGENCIA ARTIFICIAL

TEMA 1 - INTRODUCCIÓN A LA INTELIGENCIA ARTIFICIAL

EJERCICIOS

CONCEPTUALES

EJERCICIO 1

Indica para los siguientes supuestos, si un método de aprendizaje automático flexible mostraría mejor rendimiento que uno inflexible. Justifica las respuestas

1. El número de observaciones, n , es extremadamente grande mientras que el número de predictores, p , es pequeño.
2. El número de predictores, p , es extremadamente grande mientras que el número de observaciones, n , es pequeño.
3. La relación entre respuesta y predictores es altamente no-lineal.
4. La varianza del término de error, $\sigma^2 = \text{Var}(\epsilon)$, es muy alto.

EJERCICIO 2

Explica si cada uno de los escenarios abajo detallados se corresponde con un problema de regresión o clasificación. También indica si son interesantes para inferencia o predicción. Finalmente, identifica n y p .

1. Reunimos datos de las mejores 500 empresas de Euskadi. Para cada empresa se tiene su beneficio, número de empleados, tipo de actividad y salario del gerente. Estamos interesados en comprender que factores afectan el salario del gerente.
2. Vamos a lanzar un nuevo producto y nos gustaría saber si será un *éxito* o un *fracaso*. Reunimos datos de 20 productos similares lanzados previamente. Para cada uno de esos productos, registramos si fue exitoso o fallido, precio del producto, presupuesto para marketing, y otras diez variables.
3. Estamos interesados en predecir como varía el cambio EURO/USD en relación a los cambios semanales del mercado bursátil mundial. Por lo tanto, registramos semanalmente datos de todo el año anterior: el porcentaje en el que el cambio EURO/USD ha variado, el porcentaje en el que mercado Alemán ha variado, el porcentaje en el que mercado Francés ha variado, y el porcentaje en el que mercado Estadounidense ha variado.

EJERCICIO 3

Pensar en situaciones reales donde aplicar métodos de aprendizaje automático:

1. Describe tres aplicaciones en las que la *clasificación* pueda ser útil. Describe la respuesta, Y , y los predictores, \mathbf{X} . ¿Es el objetivo de cada aplicación inferencia o predicción? Justifica las respuestas.
2. Describe tres aplicaciones en las que la *regresión* pueda ser útil. Describe la respuesta, Y , y los predictores, \mathbf{X} . ¿Es el objetivo de cada aplicación inferencia o predicción? Justifica las respuestas.
3. Describe tres aplicaciones en las que el *análisis de clusters* pueda ser útil.

EJERCICIO 4

¿Cuáles son las ventajas/desventajas del uso de un método muy flexible para regresión o clasificación? ¿Bajo que circunstancias sería preferible un método más flexible que uno menos flexible? ¿Cuándo sería preferible un método menos flexible?

EJERCICIO 5

La siguiente tabla muestra datos de entrenamiento compuestos por seis observaciones, tres predictores (X_1, X_2, X_3), y una respuesta cualitativa (Y).

Observación	X_1	X_2	X_3	Y
1	0	3	0	Rojo
2	2	0	0	Rojo
3	0	1	3	Rojo
4	0	1	2	Verde
5	-1	0	1	Verde
6	1	1	1	Rojo

Suponer que quisiéramos usar este conjunto de datos para predecir Y cuando $X_1 = X_2 = X_3 = 0$ usando KNN.

- a) Calcular la distancia Euclídea entre cada observación y el punto de test, $X_1 = X_2 = X_3 = 0$.
- b) ¿Cuál es nuestra predicción con $K = 1$? ¿Por qué?
- c) ¿Cuál es nuestra predicción con $K = 3$? ¿Por qué?
- d) Si la región de decisión de Bayes en este problema fuera altamente no-lineal, entonces ¿el *mejor* valor de K sería alto o bajo? ¿Por qué?

APLICADOS

EJERCICIO 6

Este ejercicio está relacionado con la base de datos **College** que contiene variables para un total de 777 universidades y facultades Estadounidenses. Las variables que contiene son las siguientes:

- **Private**: indicador pública/privada
 - **Apps**: número de solicitudes recibidas
 - **Accept**: número de solicitudes admitidas
 - **Enroll**: número de estudiantes matriculados
 - **Top10perc**: proporción de estudiantes que estuvieron entre el mejor 10 % de su clase en bachiller
 - **Top25perc**: proporción de estudiantes que estuvieron entre el mejor 25 % de su clase en bachiller
 - **F_Undergrad**: número de alumnos a dedicación completa
 - **P_Undergrad**: número de alumnos a dedicación parcial
 - **Outstate**: número de matrículas provenientes de fuera del Estado
 - **Room_Board**: costes de alojamiento y comida
 - **Books**: coste estimado de los libros
 - **Personal**: gasto personal estimado
 - **PhD**: porcentaje de profesorado doctor
 - **Terminal**: porcentaje de profesorado que han alcanzado el más alto rango posible en sus disciplinas
 - **S_F_Ratio**: ratio esdiantes/profesorado
 - **perc_alumni**: porcentaje de alumnado que realiza donaciones
 - **Expend**: gasto en enseñanza por estudiante
 - **Grad_Rate**: ratio de graduación
1. Calcular la media y desviación estándar de cada una de las variables cuantitativas de la base de datos.
 2. Realizar el conteo de universidades públicas y privadas.
 3. Producir los siguientes diagramas de dispersión (*scatter*):
 - **Apps** vs **Accept**

- `Enroll` vs `F_Undergrad`
 - `Room_Board` vs `Outstate`
 - `PhD` vs `Expend`
 - `S_F_Ratio` vs `Expend`
 - `perc_alumni` vs `Expend`
4. Crear una variable cualitativa de nombre `Elite` binarizando la variable `Top10perc`. Si la proporción de estudiantes que estuvieron entre el mejor 10 % de su clase en bachiller es superior al 50 % `Elite` será igual a 'Yes' de lo contrario, será igual a 'No'. Calcula el número de universidades de élite y realiza un *boxplot* de la variable `Outstate` en función de la variable `Elite`.
5. Obtener los histogramas para las siguientes variables cuantitativas y después estratificarlos en base a la variable `Elite`.
- `Top10perc`
 - `Top25perc`
 - `PhD`
 - `Terminal`
 - `Room_Board`
 - `Outstate`
-

EJERCICIO 7

Este ejercicio implica el uso de la base de datos `Auto` que usamos en la práctica de ordenador de este tema.

1. Remueve las observaciones que contengan valores atípicos.
 2. Identifica los predictores cuantitativos y los cualitativos.
 3. Calcular la media, desviación estándar y rango de cada uno de los predictores cuantitativos.
 4. Eliminar las observaciones en el rango 10 – 85. ¿Cuál es ahora el rango, media y desviación estándar de cada predictor?
 5. Usando toda la base de datos, analiza los predictores de manera gráfica haciendo uso de la función *scatter*. Crea gráficos que resalten la relación entre predictores. Resume los resultados obtenidos.
 6. Suponer que queremos predecir la autonomía del coche dada en millas por galón (`mpg`) en base a otros predictores. ¿Alguno de los gráficos obtenidos previamente sugieren que otras variables puedan ser de utilidad a la hora de predecir `mpg`?
-