Práctica 2

Siguiendo las principales etapas de un proyecto analítico, las diferentes tareas a realizar (y justificar) son las siguientes:

1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?

El conjunto de datos objeto de análisis se ha obtenido a partir del enlace en Kaggle, el cual contiene la longitud y la anchura de los pétalos y sépalos y la especie de 150 flores iris. De manera que es un conjunto de datos multivariante comprendido por 5 características (columnas) de 150 flores iris (filas o registros).

El famoso estadístico Sir Ronald. A. Fisher usó este conjunto de datos en su artículo «The Use of Multiple Measurements in Taxonomic Problems» (Annals of Eugenics 7 (1936), pp. 179–188). A veces se llama el conjunto de datos Iris de Anderson porque Edgar Anderson recopiló los datos para cuantificar la variación morfológica de las flores de Iris de tres especies relacionadas. Dos de las tres especies fueron recogidas en la Península de Gaspé "todas del mismo pasto, y recogidas el mismo día y medidas al mismo tiempo por la misma persona con el mismo aparato". El conjunto de datos consta de 50 muestras de cada una de las tres especies de Iris (Iris setosa, Iris virginica e Iris versicolor). Se midieron cuatro características de cada muestra: la longitud y la anchura de los sépalos y pétalos, en centímetros. Basándose en la combinación de estas cuatro características, Fisher desarrolló un modelo discriminatorio lineal para distinguir la especie entre sí.

La idea es realizar con este conjunto de datos un análisis exploratorio o descriptivo que permita resumir, representar y explicar los datos concretos a disposición. Igualmente se pretenden plantear un modelo estadístico que logre predecir o clasificar las tres especies a partir de los 4 atributos enunciados anteriormente, lo cual se convierta en un caso de prueba y aprendizaje para las técnicas de clasificación estadística en el aprendizaje automático.

2. Integración y selección de los datos de interés a analizar.

No se realizaron procesos de integración o fusión de datos tales como añadir nuevos atributos o registros a la base original, pues no se considera necesario, por ahora, al logro de los objetivos planteados.

En cuanto a la selección de los datos se consideran todos los atributos a excepción del primer campo Id, dado que no es un atributo que mida algún tipo de característica relevante que aporte al ejercicio analítico.

Ahora bien al revisar el tipo de atributo o variable del dataset importado se observa que todos los atributos son numéricos a excepción del atributo Species, el cual se ha importado como un vector de palabras: lo indica el chr, de character, en la fila correspondiente del resultado de str. Esta variable Species es de tipo categórico y tiene asociada una descripción, una cadena de caracteres y, al mismo tiempo cuenta con un limitado número de valores posibles. Almacenar estos datos directamente como cadenas de caracteres implica un uso de memoria innecesario, ya que cada una de las apariciones en la base de datos puede asociarse con un índice numérico sobre el conjunto total de valores posibles, obteniendo una representación mucho más compacta. Para tal fin, el

atributo Species se crea como factor, de tal manera que este -el factor- se almacena internamente como un número y las etiquetas asociadas a cada valor se denominan niveles, que en este caso serán tres.

3. Limpieza de los datos.

3.1. ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?

No se identifican valores perdidos o ausentes (NA) en el dataset. En la gestión de datos ausentes se puede optar por eliminarlos o sustituirlos por el valor promedio de la columna o el valor más frecuente e incluso pueden ser reemplazados a partir de un modelo de regresión que predice dicho valor vacío; el camino a seguir dependerá de los datos a disposición y de los objetivos del análisis a realizar.

3.2. Identificación y tratamiento de valores extremos.

Con el fin de identificar valores extremos se presentan diagramas de caja por cada una de las cuatro características y según la especie de flor. Pero antes se presentan estadísticos descriptivos de cada una de las cuatro características y según la especie de flor con el fin de notar diferencias entre las especies.

Los estadísticos de tendencia central de la longitud y del ancho del sépalo entre las especies presentan diferencias marcadas; por ejemplo, la media y mediana de la longitud del sépalo de la especie virginica es mayor que las otras dos especies. En contraste, el ancho del sépalo -su media y mediana- es superior en la especie setosa. En cuanto a la longitud y ancho del sépalo, la especie virginica es mayor frente a las otras dos especies.

En los diagramas de caja de los cuatro atributos se observan diferencias marcadas en su mediana tanto en las longitudes como en los anchos del sépalo y pétalo de las flores iris. También se identifican algunos valores extremos en el atributo Ancho del sépalo.

Igualmente al realizar los diagramas de caja de los atributos de acuerdo con cada especie de iris se observan diferencias relevantes en las medianas de la longitud y ancho del pétalo, así como la longitud del sépalo; pero en el ancho del sépalo, si bien se presentan diferencias en sus medianas estas son menos marcadas. Por otra parte, también se identifican valores extremos en las características de longitud y ancho del pétalo de la especie setosa y de la especie virginica en las características de longitud y ancho del sépalo. Por ahora, se mantendrán todos los valores extremos identificados en el ejercicio analítico de este dataset.

4. Análisis de los datos.

4.1. Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).

Se divide el dataset iris en varios datasets, los cuales contienen cada uno las muestras pertenecientes a una especie de flor que sería interesante analizar y/o comparar; sin embargo, no todos se utilizarían en la realización de pruebas estadísticas posteriores.

4.2. Comprobación de la normalidad y homogeneidad de la varianza.

Con el fin de comprobar la normalidad de cada uno de los atributos y según su especie de flor se presenta el histograma y curva de densidad, gráfico de cuantiles teóricos (Q-Q plot), así como los test de normalidad Anderson-Darling cuyo nivel de significación se fija en 0.05.

En el caso de los atributos longitud y ancho del sépalo, los gráficos indican que su distribución se aleja de la distribución normal, lo cual se verifica en el test aplicado, pues el valor P (0.02251 y 0.01455 respectivamente) es menor que el nivel de significancia (0.05), por tanto, existe evidencia para rechazar hipótesis nula, es decir, que los datos no provienen de una población con distribución normal.

En el caso de los atributos longitud y ancho del pétalo, los gráficos indican que su distribución se aleja de la distribución normal, lo cual se verifica en el test aplicado, pues el valor P (2.2e-16 y 1.427e-12 respectivamente) es menor que el nivel de significancia (0.05), por tanto, existe evidencia para rechazar hipótesis nula, es decir, que los datos no provienen de una población con distribución normal.

Estos resultados -en particular la distribución de los atributos- también sugieren la presencia de diferentes muestras, es decir, se evidencia la influencia de las tres especies de flores.

Ahora bien, al analizar los atributos de longitud y ancho del sépalo y pétalo según cada una de las tres especies, a partir del histograma y curva de densidad, así como el grafico Q-Q y el test de normalidad Anderson-Darling se observa lo siguiente:

En el caso de la especie setosa el ancho del pétalo tiene una distribución sesgada a la derecha mientras que las otras variables tienen distribuciones aproximadamente normales. De acuerdo con el gráfico Q-Q se presentan algunas desviaciones de la línea recta y esto indica posibles desviaciones de una distribución normal, particularmente en el ancho del pétalo. Y según el test de normalidad, las variables longitud y ancho del pétalo no provienen de poblaciones normales, lo cual confirma lo señalado en los gráficos descritos anteriormente.

En cuanto a la especie versicolor el ancho del pétalo tiene una distribución sesgada a la derecha mientras que las otras variables tienen distribuciones aproximadamente normales. De acuerdo con el gráfico Q-Q se presentan algunas desviaciones de la línea recta y esto indica posibles desviaciones de una distribución normal, particularmente en el ancho del pétalo. Y según el test de normalidad, la variable ancho del pétalo no proviene de poblaciones normales, lo cual confirma lo señalado en los gráficos descritos anteriormente.

Y finalmente, en la especie virginica, todas las variables tienen distribuciones aproximadamente normales. Si bien se presentan ciertas desviaciones de la línea recta en el grafico Q-Q no son pronunciadas en las variables en estudio. Y el test de normalidad indica que sin excepción todas las variables provienes de poblaciones normales.

Las gráficas de histograma de frecuencias, curvas de densidad y dispersión -según las especies de flor iris- visualizadas en conjunto permiten identificar la posible separación de las especies y la superposición de valores de cada especie para un atributo en específico.

Con relación a la homogeneidad de la varianza (la varianza es constante (no varía) en los diferentes niveles de un factor) se utiliza el test de Levene. Este test de Levene se caracteriza, porque en primer lugar se puede comparar 2 o más poblaciones (en este caso son tres muestras) y, en segundo lugar permite elegir entre diferentes estadísticos de centralidad: mediana (por defecto), media, media truncada, lo cual es importante a la hora de contrastar la homocedasticidad, dependiendo de si los grupos se distribuyen de forma normal o no, lo cual como se anotó anteriormente algunas variables no siguen una distribución normal.

De acuerdo con el test aplicado se encuentran diferencias entre los tres grupos de especies de iris en todos los atributos a excepción de la característica ancho del sépalo, de lo cual ya se tenia cierto indicio desde el punto de vista gráfico con los diagramas de caja desglasado por especie, pues se indicaba que si bien se presentaban diferencias en las medianas, estas no eran muy marcadas en comparación a los otros atributos según especie de iris.

4.3. Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.

Se realiza un Anova con el fin comparar las medias de cada uno de los atributos entre los grupos o especies de flores iris. Al establecer el valor alfa en 0.05 y al ver en la tabla que el valor de p es menor a alfa, se rechaza la hipótesis nula de que las medias son iguales, y se concluye que la media de la longitud y ancho del sépalo y pétalo es distinta entre las tres especies en todos los casos.

Al verificar la correlación entre los atributos de longitud y ancho del sépalo y pétalo se observa que entre la longitud del sépalo y la longitud y ancho del pétalo guardan una correlación positiva superior al 80%. Mientras que el ancho del sépalo guarda una correlación negativa con estas mismas variables, pero mucho menor (entre el 35% y 42%). Finalmente entre la longitud y ancho del pétalo su correlación es positiva y es del 96% (muy alta) y, entre la longitud y ancho del sépalo su correlación es negativa y muy baja (10%).

A continuación se presentan las correlaciones y diagramas de dispersión con línea de regresión, por parejas de atributos, pero desglosadas por cada especie de iris.

En cuanto a la correlación para cada uno de los grupos o especies de flores iris se observa que en el caso de la especie setosa, los atributos con correlaciones relevantes son la longitud y ancho del sépalo (76%). En cuanto a la especie versicolor, las correlaciones superiores al 60% ocurren entre las longitudes del sépalo y pétalo (75%) y entre los anchos del sépalo y pétalo (66%). Y, en la especie virginica, se presenta una única correlación alta entre las longitudes del sépalo y pétalo (86%).

Finalmente, el diagrama de dispersión -según las especies de flor iris- visualizadas en conjunto permiten identificar la posible separación de las especies. Se observa que las variables longitud y ancho del pétalo son las dos variables con más potencial para poder separar entre especies. Sin embargo, como se indicó en párrafos anteriores están altamente correlacionadas, por lo que la información que aportan es en gran medida redundante.

Ahora bien con el fin de clasificar las tres especies de iris a partir de sus atributos de longitud y ancho del sépalo y pétalo se plantea desarrollar un Análisis Discriminante Lineal o Linear Discrimiant Analysis (LDA). La LDA es un método de clasificación de variables cualitativas en el que dos o más grupos son conocidos a priori y nuevas observaciones se clasifican en uno de ellos en función de sus características. Haciendo uso del teorema de Bayes, LDA estima la probabilidad de que una observación, dado un determinado valor de los predictores, pertenezca a cada una de las clases de la variable cualitativa, $P(Y=k|X=x)$. Finalmente se asigna la observación a la clase k para la que la probabilidad predicha es mayor.

Se requieren las siguientes dos condiciones para que el LDA se considera valido:

La primera es que cada predictor que forma parte del modelo se distribuye de forma normal en cada una de las clases de la variable respuesta. En un apartado anterior se presentaron los resultados y en general se puede decir que la mayoría de los predictores en cada una de las clases sigu en la distribución normal, a excepción de la variable ancho del pétalo, la cual no se distribuye de forma normal en las especies setosa y versicolor.

En el caso de múltiples predictores, las observaciones siguen una distribución normal multivariante en todas las clases. Con el fin de verificar el cumplimiento de esta condición se aplica el test de normalidad multivariante royston. El test muestra evidencias significativas de falta de normalidad multivariante. El LDA tiene cierta robustez frente a la falta de normalidad multivariante, pero es importante tenerlo en cuenta en la conclusión del análisis.

Y la segunda condición a cumplir es que la varianza del predictor es igual en todas las clases de la variable respuesta. En el caso de múltiples predictores, la matriz de covarianza es igual en todas las clases. Si esto no se cumple se recurre a Análisis Discriminante Cuadrático (QDA).

Con el fin de verificar el cumplimiento de esta condición se aplica el test Box M, el cual se utiliza en el caso multivariante y permite contrastar la igualdad de matrices entre grupos.

El test Box's M muestra evidencias de que la matriz de covarianza no es constante en todos los grupos, $p(2.2e\text{-}16)<\alpha(0.05)$, lo que a priori descartaría el método LDA en favor del QDA. Sin embargo, como el test Box's M es muy sensible a la falta de normalidad multivariante, con frecuencia resulta significativo no porque la matriz de covarianza no sea constante sino por la falta de normalidad, cosa que ocurre para los datos en estudio. Por esta razón se va a asumir que la matriz de covarianza sí es constante y que LDA puede alcanzar una buena precisión en la clasificación. En la evaluación del modelo se verá como de buena es esta aproximación.

5. Representación de los resultados a partir de tablas y gráficas.

Se procede al cálculo de la función discriminante

Se procede a realizar la predicción con el mismo dataset.

Una vez que las normas de clasificación se han establecido, se tiene que evaluar como de buena es la clasificación resultante. En otras palabras, evaluar el porcentaje de aciertos en las clasificaciones. Para tal fin se presenta la matriz de confusión, la cual presenta el número de verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos.

Solo 3 de las 150 predicciones que ha realizado el modelo han sido erróneas.

Ahora bien, para evaluar el error de clasificación se emplean las mismas observaciones con las que se ha creado el modelo, obteniendo así lo que se denomina el training error. Si bien esta es una forma sencilla de estimar la precisión en la clasificación, tiende a ser excesivamente optimista. Es más adecuado evaluar el modelo empleando observaciones nuevas que el modelo no ha visto, obteniendo así el test error.

El trainig error es muy bajo (2%), lo que apunta a que el modelo es bueno. Sin embargo, para validarlo es necesario un nuevo set de datos con el que calcular el test error o recurrir a validación cruzada.

Se presenta a continuación una visualización que representa los límites de clasificación de un modelo discriminante lineal para cada par de predictores. Cada color representa una región de clasificación acorde al modelo, se muestra el centroide de cada región y el valor real de las observaciones.

Creación de un sets de entrenamiento y prueba

Se crea un set de entrenamiento para generar un modelo predictivo, y un set de prueba, para comprobar la eficacia de este modelo para hacer predicciones correctas. Se obtiene un subconjunto del dataset original, que consiste en 70% del total de ellos. Y se obtiene el subconjunto de datos complementario al de entrenamiento para el set de prueba, esto es, el 30% restante.

De acuerdo con el modelo ninguna de las 45 predicciones realizadas ha sido incorrecta; el test de error es del 0%.

6. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?

El modelo de clasificación LDA desarrollado presentan muy buenos resultados en la tarea de predecir o clasificar las especies de flores iris a partir de los atributos de longitud y ancho del sépalo y pétalo, tanto en los datos originales como con datos de prueba.

Contribuciones	Firma
Investigación previa	HHM
Redacción de las respuestas	HHM
Desarrollo código	HHM