Clasificación: Regresión Logística y LDA.

José I. González Etchemaite
Aprendizaje Estadístico

17 de junio de 2020

1. Desarrollo

1.1. Ejercicio 1

En este ejercicio se intenta ajustar un clasificador de abalones (en adulto o infante) utilizando tres parámetros: longitud, peso total y cantidad de anillos. Se analiza la performance del clasificador utilizando los tres features por separados y luego juntos, dando un total de cuatro clasificadores a ser comparados.

Para ello se separa el conjunto de muestras disponibles en dos grupos: el 75 % será utilizado para entrenamiento, y el 25 % restante para probar los clasificadores obtenidos.

En la siguiente tabla se resume las métricas de los cuatro clasificadores obtenidos:

	Precision	Recall	Accuracy
Longitud	0.50	0.73	0.78
Peso total	0.62	0.71	0.80
Anillos	0.43	0.76	0.77
Conjunta	0.70	0.75	0.83

Figura 1: Resumen de las métricas obtenidas.

Se puede ver que el clasificador con mayor tasa de éxito es el que toma las tres variables, mientras que la variable mas explicativa de las tres es el peso total del abalone. Además, el clasificador que utiliza solamente la cantidad de anillos será el que mejor clasifique infantes como tal, pues tiene el mayor recall(o sensibilidad).

1.2. Ejercicio 2

En este ejercicio se hace la clasificación de abalones mediante el método de los discriminantes lineales (LDA), donde se asume que la distribución de los features(o variables predictoras) para cada clase(adulto e infante) siguen una distribución normal, con parámetros estimados a partir de los datos de entrenamiento.

Se comienza con el modelo que utiliza una sola de las variables para clasificar y se elige la variable que mejor clasifique bajo cierto criterio. Para ello se separan el 25 % de las muestras

como en el ejercicio anterior y el modelo se prueba sobre ellas. La variable ganadora es incorporada al modelo y el proceso anterior se repite, evaluando cómo varía la métrica al incorporar una de las variables restantes. Esto se repite hasta que el modelo quede completo.

El criterio utilizado para evaluar la performance de cada modelo es el accuracy, que equivale a la proporción de muestras bien clasificadas sobre el total. En la tabla a continuación se muestra el accuracy para cada modelo.

	Accuracy								
P	**1**	**2**	3	**4**	**5**	**6**	**7**	**8**	**TOTAL**
Longitud	0.7807	0.7893	0.817	0.8103	0.8103	0.8103	0.8113		x7
Diametro	0.7826	0.7902	0.8056	0.8056	0.8103	0.8103	0.8075	0.8084	x8
Altura	0.7816	0.8017	0.8142	0.8123	0.8132	0.8132			x6
Peso total	0.796	0.8008	0.817						x3
Peso Carne	0.7826	0.8046	0.8132	0.8151	0.8161				x5
Peso Viscera	0.8074								x1
Peso Caparazon	0.794	0.7969	0.8151	0.8161					x4
Cantidad Anillos	0.7739	0.8142							x2

Figura 2: Accuracy para distintas dimensiones.

Se observa que la variable que mejor explica la clase a detectar es el peso de las vísceras. Luego, se espera que la variable que mejor explique sea la menos correlacionada a esta. Tiene sentido creer que la cantidad de anillos puede tener mayor descorrelación que otras variables asociadas a la masa y geometría del abalone. El resultado comprueba esta intuición satisfactoriamente.

Por último, se puede observar que el modelo con mayor accuracy es el que incluye las tres variables, a saber: peso de las vísceras, numero de anillos y peso total. A partir de aquí, ningún agregado de variables al modelo logra aumentar la métrica.

2. Conclusiones

Se intentó resolver un problema por dos métodos distintos y se obtuvieron resultados diferentes. Para el primer caso se consiguió mayor proporción de clasificaciones correctas teniendo en cuenta las variables longitud, peso total y cantidad de anillos del abalone. Para el segundo, esto se consiguió considerando el peso de las vísceras, cantidad de anillos y peso total, resultando este clasificador con menor accuracy que el anterior.

También se evaluó para cada método aplicado, el entrenamiento de un modelo tomando las variables que eran óptimas para el otro, concluyendo de que estas terminan explicando de manera distinta la clase a predecir según el caso:

	Accuracy	
Variables predictoras (X)	Reg. Logística	LDA
(peso total, anillos, longitud)	0.83	0.8
(peso total, anillos, peso vísceras)	0.81	0.82

Figura 3: Comparación de modelos según Accuracy.

Para modelos de una sola variable se obtienen mejores resultados con LDA(utilizando peso de las vísceras como variable predictora principal) que en regresión logística(que utiliza peso total). Pero con tres variables, se observa que el mejor resultado se consigue cuando se utiliza un modelo de regresión logística. Esto puede explicarse debido a la naturaleza del problema. Si se intenta clasificar utilizando LDA, se está asumiendo que las variables explicativas siguen una distribución de densidad de probabilidad normal, con parámetros a estimar a partir de las muestras de entrenamiento. Esto puede resultar en un mejor o peor clasificador cuanto mas cerca o lejos esté la realidad de los datos(mas precisamente, los datos de prueba) de estas hipótesis. Por el contrario, la clasificación mediante regresión logística tiene mejor performance cuando la naturaleza de las clases no son como las supuestas en LDA.

Si bien no se muestran los resultados en este informe, se hizo una ligera prueba sobre algunos modelos utilizando QDA, y los resultados no fueron satisfactorios, obteniendo clasificadores de menor accuracy que los considerados anteriormente.