Facultad de Ciencias, UNAM
Reconocimiento de patrones y aprendizaje automatizado
Tarea 2

Rubí Rojas Tania Michelle

11 de febrero de 2021

1. Cada una de las líneas en el documento `tarea2_docs.txt` lo vamos a considerar como un documento. Realiza la limpieza y preprocesamiento necesarios para contestar las siguientes preguntas:

   - ¿En qué tópicos se pueden dividir?
   - ¿Qué documentos habla de México y su cultura?

2. Descarga el dataset de datos de sonar de la siguiente liga: `https://archive.ics.uci.edu/ml/machine-learning-databases/undocumented/connectionist-bench/sonar/sonar.all-data`

   Mismo que deberás explorar y entender lo que hacen sus atributos.

   Lleva a cabo una clasificación utilizándo la regresión logística de los datos. Luego, efectúa una reducción de dimensión y haz tu clasificación otra vez. La comparación entre las clasificaciones la harás con las métricas que tú elijas.

   ¿Qué observas y por qué sucede esto?

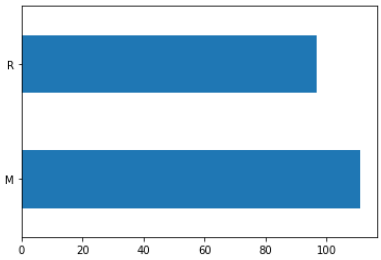
   Solución: **sonar.all-data** es un conjunto de datos que describe el sonido (o más bien, el chirrido) del sonar al rebotar sobre diferentes superficies. Las 59 variables de entrada son la fuerza de los retornos en ángulos diferentes y se encuentran en un rango de 0 a 1. La variable 60 de salida es un string **M** para Mina y **R** para Roca.

   El problema es de clasificación binaria, la cual requiere de un modelo para diferenciar rocas de cilindros metálicos.

   Descargamos el conjunto de datos y lo visualizamos.

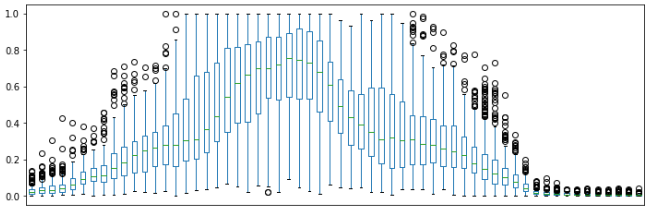
0	1	2	3	4	5	6	7	8	9	...	51	52	53	54	55	56	57	58	59	60
0.0200	0.0371	0.0428	0.0207	0.0954	0.0986	0.1539	0.1601	0.3109	0.2111	...	0.0027	0.0065	0.0159	0.0072	0.0167	0.0180	0.0084	0.0090	0.0032	R
0.0453	0.0523	0.0843	0.0689	0.1183	0.2583	0.2156	0.3481	0.3337	0.2872	...	0.0084	0.0089	0.0048	0.0094	0.0191	0.0140	0.0049	0.0052	0.0044	R
0.0262	0.0582	0.1099	0.1083	0.0974	0.2280	0.2431	0.3771	0.5598	0.6194	...	0.0232	0.0166	0.0095	0.0180	0.0244	0.0316	0.0164	0.0095	0.0078	R
0.0100	0.0171	0.0623	0.0205	0.0205	0.0368	0.1098	0.1276	0.0598	0.1264	...	0.0121	0.0036	0.0150	0.0085	0.0073	0.0050	0.0044	0.0040	0.0117	R
0.0762	0.0666	0.0481	0.0394	0.0590	0.0649	0.1209	0.2467	0.3564	0.4459	...	0.0031	0.0054	0.0105	0.0110	0.0015	0.0072	0.0048	0.0107	0.0094	R
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
0.0187	0.0346	0.0168	0.0177	0.0393	0.1630	0.2028	0.1694	0.2328	0.2684	...	0.0116	0.0098	0.0199	0.0033	0.0101	0.0065	0.0115	0.0193	0.0157	M
0.0323	0.0101	0.0298	0.0564	0.0760	0.0958	0.0990	0.1018	0.1030	0.2154	...	0.0061	0.0093	0.0135	0.0063	0.0063	0.0034	0.0032	0.0062	0.0067	M
0.0522	0.0437	0.0180	0.0292	0.0351	0.1171	0.1257	0.1178	0.1258	0.2529	...	0.0160	0.0029	0.0051	0.0062	0.0089	0.0140	0.0138	0.0077	0.0031	M
0.0303	0.0353	0.0490	0.0608	0.0167	0.1354	0.1465	0.1123	0.1945	0.2354	...	0.0086	0.0046	0.0126	0.0036	0.0035	0.0034	0.0079	0.0036	0.0048	M
0.0260	0.0363	0.0136	0.0272	0.0214	0.0338	0.0655	0.1400	0.1843	0.2354	...	0.0146	0.0129	0.0047	0.0039	0.0061	0.0040	0.0036	0.0061	0.0115	M

   Veamos si la proporción de los datos en cada clase es equilibrada.

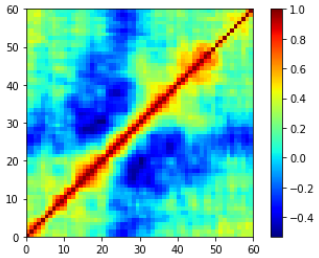


donde existen 111 datos para la clase Mina y 97 datos para la clase Roca. Como no hay mucha diferencia entre ambas clases, entonces no es necesario rebalancear de alguna manera el conjunto de datos.

Aunque los metadatos dicen que todos los predictores van de 0 a 1, todavía encontramos que hay varios de ellos que tienen valores relativamente menores en comparación con la mayoría. También encontramos que no hay ningún valor alto más grande que 1.



La matriz de correlación nos indica que hay alguna estructura en el órden de los atributos: el color rojo alrededor de la diagonal sugiere que los atributos que están uno al lado del otro están más correlacionados entre sí (es decir, cada variable tiene una gran correlación positiva con sus vecinos), las manchas azules sugieren alguna correlación negativa moderada entre los atributos que están más alejados unos de otros. Esto tiene sentido si el órden de los atributos se refiere al ángulo de los sensores para el chirrido del sonar.



Ahora bien, para el procesamiento de datos separamos al conjunto de datos en dos:

- X: el cual está conformado por las primeras 59 columnas del dataset original.
- y: el cual está conformado por la columna 60 del dataset original.

La gráfica de cajas nos sugiere que debemos estandarizar los datos de entrada, por lo que usamos la clase **StandardScaler** para lograr esto (recordándo que ésta elimina la media y escala los datos de forma que su varianza sea igual a 1).

La variable de salida (la columna 60) es un string **M** o **R**. Debemos convertirlos en valores enteros entre 0 y 1, y para lograr esto usamos la clase **LabelEncoder** (la cual modelará la codificación requerida usando todo el conjunto de datos a traves de la función `fit()` y luego aplicará la codificación para crear una nueva variable de salida usando la función `transform()`).

Para crear el modelo de regresión lineal, dividimos nuestro conjunto de datos ya modificado en entrenamiento (80 %) y prueba (20 %). Luego, simplemente creamos una instancia de `LogisticRegression` y lo entrenamos.

Los resultados obtenidos fueron los siguientes:

```
print('Accuracy of LR classifier on training set: {:.2f}'
     .format(logistic_regression.score(X_train, y_train)))
print('Accuracy of LR classifier on test set: {:.2f}'
     .format(logistic_regression.score(X_test, y_test)))
```

```
Accuracy of LR classifier on training set: 0.90
Accuracy of LR classifier on test set: 0.88
```

```
print("Mean Absolute Error: %.2f" % mean_absolute_error(y_test, y_pred))
# Error Cuadrado Medio
print("Mean squared error: %.2f" % mean_squared_error(y_test, y_pred))
print("Root Mean Squared Error: %.2f" % np.sqrt(mean_squared_error(y_test, y_pred)))
# Puntaje de Varianza. El mejor puntaje es un 1.0
print('Variance score: %.2f' % r2_score(y_test, y_pred))
```

```
Mean Absolute Error: 0.12
Mean squared error: 0.12
Root Mean Squared Error: 0.35
Variance score: 0.52
```

```
print(classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.86      0.90      0.88        21
           1       0.90      0.86      0.88        21

    accuracy                           0.88        42
   macro avg       0.88      0.88      0.88        42
weighted avg       0.88      0.88      0.88        42
```

donde en particular podemos notar que

- Obtenemos un 90 % de precisión en el conjunto de entrenamiento y un 88 % de precisión en el conjunto de prueba.
- Obtenemos MSE de 0.12 (con un puntaje de varianza de 0.52).
- Obtenemos 86 % de precisión para los datos correspondientes a metales y 90 % de precisión para los datos correspondientes a las rocas.

Por otro lado, usaremos PCA para obtener la lista de atributos que tienen mayor varianza (mayor poder explicativo). Éstos serán los componentes principales. Realizamos un par de gráficas para poder determinar el número de estos componentes.

Donde obtenemos que con 43 componentes tenemos un 99 % de la varianza explicada (queremos una v.e. entre 95 % y 99 %, pero me ha gustado más al resultado con este valor). Así, creamos una instancia de PCA con un número de 43 componentes principales y la entrenamos, para después pasarle este nuevo conjunto X_pca al modelo de `LogisticRegression` y entrenarlo.

Los resultados obtenidos fueron los siguientes:

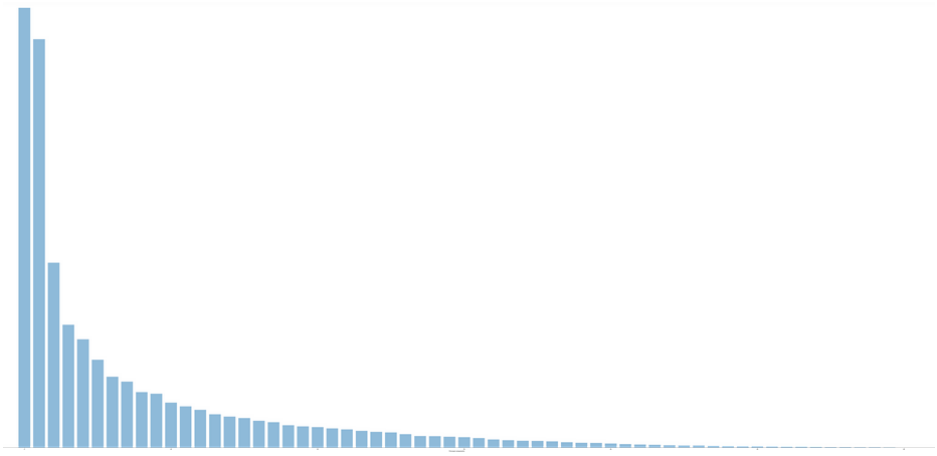


Figura 1: Variance Ratio vs Principal componentes

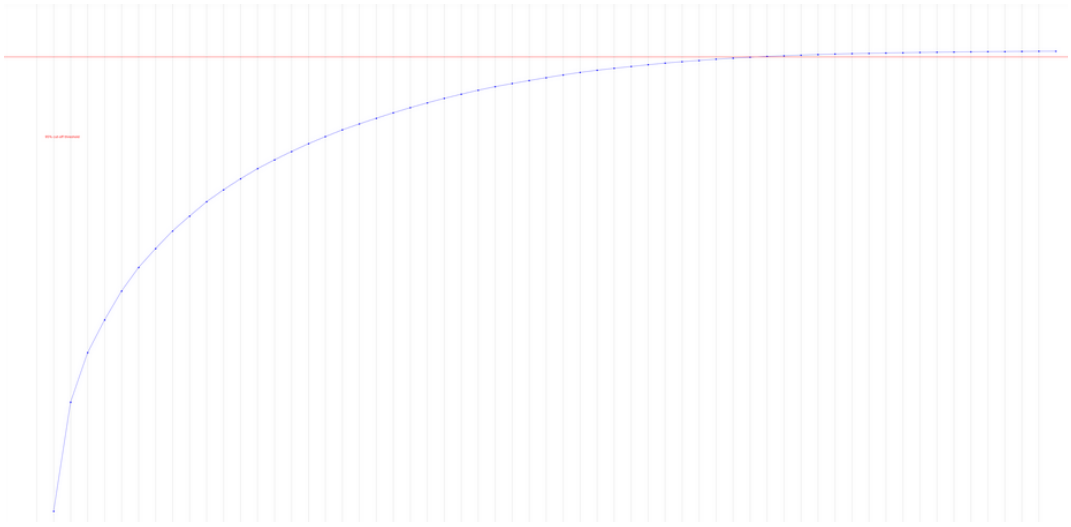


Figura 2: Cumulative variance vs Number of Components

```
print('Accuracy of LR classifier on training set: {:.2f}'
     .format(logistic_regression.score(X_train, y_train)))
print('Accuracy of LR classifier on test set: {:.2f}'
     .format(logistic_regression.score(X_test, y_test)))
```

```
Accuracy of LR classifier on training set: 0.90
Accuracy of LR classifier on test set: 0.88
```

```
print("Mean Absolute Error: %.2f" % mean_absolute_error(y_test, y_pred))
# Error Cuadrado Medio
print("Mean squared error: %.2f" % mean_squared_error(y_test, y_pred))
print("Root Mean Squared Error: %.2f" % np.sqrt(mean_squared_error(y_test, y_pred)))
# Puntaje de Varianza. El mejor puntaje es un 1.0
print('Variance score: %.2f' % r2_score(y_test, y_pred))
```

```
Mean Absolute Error: 0.12
Mean squared error: 0.12
Root Mean Squared Error: 0.35
Variance score: 0.52
```

	precision	recall	f1-score	support
0	0.86	0.90	0.88	21
1	0.90	0.86	0.88	21
accuracy			0.88	42
macro avg	0.88	0.88	0.88	42
weighted avg	0.88	0.88	0.88	42

donde en particular podemos notar que

- Obtenemos un 90 % de precisión en el conjunto de entrenamiento y un 88 % de precisión en el conjunto de prueba.
- Obtenemos MSE de 0.12 (con un puntaje de varianza de 0.52).
- Obtenemos 86 % de precisión para los datos correspondientes a metales y 90 % de precisión para los datos correspondientes a las rocas.

Así, podemos concluir que obtenemos la misma precisión que en el modelo anterior usando 43 componentes principales (16 columnas menos). Esto puede ser causado por el nivel de correlación que notamos al analizar los datos.

3. Diseña un experimento para determinar para qué umbral es posible tener un conjunto de datos no balanceado.

4. Reproduce la gráfica de la información mutua del seno contra él mismo considerándo un corrimiento de 50 posiciones.

5. El dataset de `seatbelt.csv` representa una serie de datos temporales de diversos atributos que hacen referencia a la mortalidad asociada a los accidentes de tránsito en Gran Bretaña en el periodo de 1969 y 1984. La legislación para utilizar el cinturón de seguridad de manera obligatoria fue introducida el 31 de enero de 1983. Ajusta un modelo lineal generalizado para determinar la probabilidad de morir en un accidente de tránsito y responde las siguientes preguntas:

   - ¿Volver obligatorio el cinturón de seguridad disminuyó la probabilidad de morir en algún accidente de tránsito?
   - ¿Qué otra conclusión puedes generar a partir del modelo que ajustaste?

   Se espera que lleves a cabo la normalización o estandarización del dataset, la transformación de las variables que consideres pertinente y que propongas qué variables tienen más o menos importancia en la probabilidad de morir en un accidente de tránsito.