Practica 1 de Reconocimiento de Formas

Santiago Valverde García

1. Tasas de acierto y matrices de confusión de Iris, Wine y Cancer

	Tasas de Acierto	
	Euclídeo	Bayesiano
Iris	92.66	98.00
Wine	72.47	99.43
Cancer	89.10	97.53

Matrices de confusión del clasificador Euclídeo:

Iris:

	Setosa	Versicolor	Virginica
Setosa	50	0	0
Versicolor	0	46	4
Virginica	0	7	43

Wine:

	1	2	3
1	50	0	9
2	3	49	19
3	1	17	30

Cancer:

	B	M
B	353	4
M	58	154

Matrices de confusión del clasificador Bayesiano:

Iris:

	Setosa	Versicolor	Virginica
Setosa	50	0	0
Versicolor	0	48	2
Virginica	0	1	49

Wine:

	1	2	3
1	59	0	0
2	1	70	0
3	0	0	48

Cancer:

	B	M
B	352	5
M	9	203

Iris: Dado que las tasas de acierto del clasificador Euclídeo y el bayesiano son muy similares, podemos deducir que las clases están muy separadas y se solapan muy poco. En concreto iris-setosa tiene un 100% de acierto para ambos clasificadores, como se puede observar en la matriz de confusión. Por lo tanto esta clase tiene una distancia entre ocurrencias muy pequeña en comparación con la distancia entre su centroide y los centroides de otras clases, es decir, está muy poco dispersa.

Wine: En este caso las tasas de acierto de los clasificadores varían enormemente. De este hecho deducimos que, aunque las clases tienen dispersiones muy diferentes, no se solapan prácticamente nada. De esta forma el clasificador bayesiano obtiene casi un 100% de acierto. En la matriz de confusión podemos ver que el único error se corresponde con una ocurrencia de la clase 2 clasificada en la clase 1 erroneamente.

Cancer: En este dataset la mejoría del clasificador Euclídeo al bayesiano es moderada. Si echamos un vistazo a las matrices de confusión vemos que la diferencia es significativa únicamente en el número de errores que se cometen al clasificar ocurrencias de M. Pasamos de 58 errores a 9, mientras que en la otra clase pasamos de 4 a 5. Esto significa 2 cosas: que la clase B está muy poco dispersa y la M mucho. Finalmente, dado que la tasa de acierto del bayesiano es muy alta también podemos deducir que las clases no se solapan casi nada.

2. Clasificación mediante la distancia Euclídea.

3. El clasificador estadístico bayesiano.

La tasa de acierto para el clasificador por distancia Euclídea es: 86.655%

La tasa de acierto para el clasificador estadístico bayesiano es: 89.40%

Como las tasas de acierto del clasificador de distancia Euclídea y el clasificador estadístico bayesiano son muy parecidas y bastante altas y bastante altas podemos deducir que las clases tienen una dispersión muy parecida en general y que se solapan muy poco. Solo se clasifica mal cerca de 1 de cada 10 datos con el clasificador bayesiano.

```
[[ 1.0121  0.0988 -0.0573 -0.0472  0.0073 -0.2082  0.7162  0.0414  0.2172]
 [ 0.0988  4.9483 -0.4938 -0.2152  2.9893  0.0278  0.5449 -0.4269 -0.5212]
 [-0.0573 -0.4938  0.3455 -0.0792 -0.5649  0.0328 -0.2135  0.1007  0.1536]
 [-0.0472 -0.2152 -0.0792  1.8296  0.094  -0.1344  0.4064 -0.1899 -0.2739]
 [ 0.0073  2.9893 -0.5649  0.094   3.6712  0.2957  1.2914 -0.0708 -0.9493]
 [-0.2082  0.0278  0.0328 -0.1344  0.2957  0.8492  0.1582 -0.098  -0.3752]
 [ 0.7162  0.5449 -0.2135  0.4064  1.2914  0.1582  2.2604  0.4648 -0.6686]
 [ 0.0414 -0.4269  0.1007 -0.1899 -0.0708 -0.098   0.4648  0.8003 -0.0425]
 [ 0.2172 -0.5212  0.1536 -0.2739 -0.9493 -0.3752 -0.6686 -0.0425  1.0195]]
```

Matriz de covarianza del digito 1

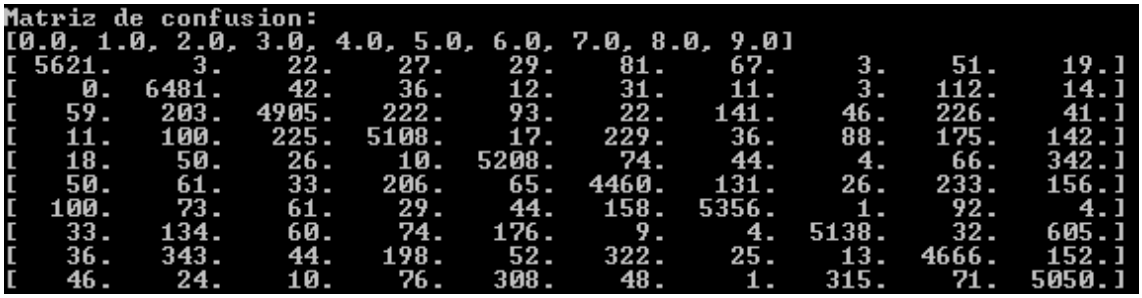
En la imagen anterior tenemos un ejemplo de matriz de covarianza. En este caso es la matriz del digito 1. Sabemos que una matriz de covarianzas con la desviación estándar en las esquinas y el resto de valores ceros, es circular (en dos dimensiones). En este caso tenemos valores muy cercanos a 1 en las esquinas lo cual quiere decir que las desviaciones estándar son ligeramente superiores a uno y no muy diferentes de los valores que tenemos en la matriz de covarianzas. Además, el resto de valores no superan el 0, salvo excepciones como el segundo parámetro que tiene una varianza grande (4.9483) y algún otro.

Esto explica que los porcentajes del clasificador mediante distancia Euclídea y el clasificador estadístico bayesiano tengan tasas de acierto similares dado que en el caso anteriormente mencionado en el que la desviación estándar aparece en las esquinas de la matriz y el resto son ceros, sabemos que el clasificador basado en la distancia Euclídea es óptimo.

```
[0.0, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0]
[[   0.   365.8  227.6  192.5  188.6  139.5  134.1  225.2  193.4  183. ]
 [ 365.8    0.   183.7  150.   237.6  200.4  225.7  291.6  149.6  225.3]
 [ 227.6  183.7    0.   108.6  224.7  191.9  127.3  185.5  147.6  183.1]
 [ 192.5  150.   108.6    0.   130.9   84.   125.2  171.8  101.5  121.7]
 [ 188.6  237.6  224.7  130.9    0.    92.6  163.1  165.6  122.1   59.6]
 [ 139.5  200.4  191.9   84.    92.6    0.   121.3  200.8   64.   105.1]
 [ 134.1  225.7  127.3  125.2  163.1  121.3    0.   188.4  133.   151.4]
 [ 225.2  291.6  185.5  171.8  165.6  200.8  188.4    0.   201.6   78.8]
 [ 193.4  149.6  147.6  101.5  122.1   64.   133.   201.6    0.   114.6]
 [ 183.   225.3  183.1  121.7   59.6  105.1  151.4   78.8  114.6    0. ]]
```

Relaciones de distancias entre centroides

En esta imagen podemos observar que clases tienen los centroides más cercanos. Por ejemplo el 1 y el 2 tienen centroides muy lejanos, mientras que el 8 y el 5 los tienen muy cercanos.



```
Matriz de confusion:
[0.0, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0]
[ 5621.     3.    22.    27.    29.    81.    67.     3.    51.    19.]
[    0.  6481.    42.    36.    12.    31.    11.     3.   112.    14.]
[   59.   203.  4905.   222.    93.    22.   141.    46.   226.    41.]
[   11.   100.   225.  5108.    17.   229.    36.    88.   175.   142.]
[   18.    50.    26.    10.  5208.    74.    44.     4.    66.   342.]
[   50.    61.    33.   206.    65.  4460.   131.    26.   233.   156.]
[  100.    73.    61.    29.    44.   158.  5356.     1.    92.     4.]
[   33.   134.    60.    74.   176.     9.     4.  5138.    32.   605.]
[   36.   343.    44.   198.    52.   322.    25.    13.  4666.   152.]
[   46.    24.    10.    76.   308.    48.     1.   315.    71.  5050.]
```

Matriz de confusión clasificador mediante distancia euclidea

Como hemos visto antes los centroides del 1 y el 2 son muy lejanos y en la imagen de la matriz de confusión podemos certificarlo ya que solo hay 3 unos clasificados erróneamente como doses. Sin embargo, dado que la distancia entre el centroide del 8 y el del 5 es muy reducida, hay 322 ochos clasificados erróneamente como cincos.

4. Evaluación del rendimiento.

Para la evaluación entrenando con el conjunto D1 y clasificando los propios datos de D1 tenemos una tasa de acierto de: 94.6 para el bayesiano y 88.6 para el Euclídeo.

Mientras que entrenando con el conjunto D1 y clasificando los datos de D2 tenemos una tasa de acierto de: 85.88 para el Euclídeo y 84.90 para el bayesiano.

La tabla de tasas será la siguiente:

Tasa de acierto (%)			
Entrenamiento	Clasificación	Bayes	D. Euclidea
Entero	Entero	89.40	86.66
D1	D1	94.6	88.6
D1	D2	84.90	85.88
1-Fold	10-Fold	94.6	88.6

Los resultados de 10-fold corresponden a la tasa media de acierto

Para la primera prueba hemos usado el conjunto entero de datos para entrenar el clasificador. En esta prueba se puede apreciar una leve mejoría en el clasificador bayesiano respecto al Euclídeo.

En la segunda prueba usamos los datos de D1 para entrenar y clasificamos estos mismos. Aquí observamos una mejoría mucho más acentuada lo cual significa que los primeros 500 datos del conjunto tienen una dispersión menor que la media del con junto entero y que se sus clases se solapan bastante menos.

En la tercera prueba entrenamos con los datos de D1 y clasificamos los de D2. Aquí vemos un suceso curioso, la tasa de acierto del clasificador bayesiano es menor que la del clasificador de distancia Euclídea. Esto probablemente está relacionado con los resultados de la prueba anterior donde podemos ver que los datos de D1 tienen una dispersión muy diferente al conjunto entero de datos.

Finalmente el 10-Fold en el que entrenamos con una fracción de los datos de D1 y clasificamos el resto iterativamente 10 veces con 10 fracciones diferentes. En esta prueba obtenemos unos resultados muy similares a la segunda prueba en la que entrenábamos con los datos de D1 y clasificábamos esos mismos datos.