Aprendizaje automático

Práctica 5: Clasificación Bayesiana

GAJAN Antoine (894825)

Escuela de
Ingeniería y Arquitectura
Universidad Zaragoza

Índice

Introducción 2

1. Estudio previo 3

2. Memoria de la practica 5

2.1. Cuestión 1 : Entrenamiento y clasificación con modelos Gaussianos regularizados 5

2.2. Cuestión 2 : Bayes ingenuo 5

2.3. Cuestión 3 : Covarianzas completas 10

2.4. Comparación de modelos 15

Conclusión 17

Introducción

Como parte de la asignatura .Aprendizaje automático", los estudiantes se familiarizarán con los conceptos de inteligencia artificial. Más concretamente, durante el cuatrimestre se estudiarán las bases del aprendizaje supervisado y no supervisado.

En esta quinta práctica vamos a poner en práctica los conocimientos de los cursos magistrales sobre la clasificación bayesiana. En particular, vamos a estudiar el funcionamiento de los algoritmos de Bayes completo y Bayes ingenuo. A través del aprendizaje de la distribución de datos, podremos predecir la clase a la que pertenece el dato. En particular, esta práctica nos permitirá ver las diferencias con la regresión logística estudiada en la práctica anterior.

Esta memoria pondrá de relieve el proceso emprendido para responder a las preguntas, y producirá una reflexión personal sobre los resultados obtenidos.

1. Estudio previo

En primer lugar, tuvimos que estudiar los algoritmos del curso para comprender cómo funcionaban y luego implementarlos. Así que aquí vamos a proponer los algoritmos de aprendizaje bayesiano y de clasificación bayesiana.

Algorithm 1 entrenarGaussianas(Xtr, ytr, naiveBayes, lambda)

$modelo \leftarrow \{N : [], mu : [], sigma : []\}$
{Para cada clase}
for $clase \in clases$ **do**
$Xtr\_clase = Xtr(ytr == clase, :)$
{Numero de datos en la clase}
$modelo(clase).N = size(Xtr\_clase, 1)$
{Bayes completo}
if $naivesBayes! = 1$ **then**
{Aprendizaje general con la matriz de covarianza}
$modelo(clase).mu = mean(Xtr\_clase, 1)$
$modelo(clase).sigma = 1/(modelo(clase).N - 1) * sum((Xtr\_clase - modelo(clase).mu)' * (Xtr\_clase - modelo(clase).mu)$
else
{Bayes ingenuo}
for $atributo \leftarrow 1$ to D **do**
{Aprendizaje de cada atributo separamente}
$modelo(clase).mu(atribute) = mean(Xtr\_clase(:, atribute))$
$modelo(clase).sigma(atribute) = 1/modelo(clase).N-1) * sum(Xtr\_clase(:, atribute) - modelo(clase).mu(atribute))^2)$
end for
{Matriz de covarianza diagonal porque independencia de los atributos}
$modelo(clase).sigma = diag(modelo(clase).sigma)$
end if
{Regularizacion}
$modelo(clase).sigma = modelo(clase).sigma + landa * eye(size(Xtr\_clase, 2)$
end for
return modelo

Ahora podemos implementar el algoritmo de aprendizaje bayesiano, que para un modelo dado y una entrada dada, proporciona la salida predicha.

Algorithm 2 clasificacionBayesiana(modelo, X)

$yhat \leftarrow []$

$best\_clase \leftarrow 0$

$best\_prob \leftarrow 0$
{Para cada clase}

for $clase \leftarrow 1$ to D **do**

$prob \leftarrow gaussLog(modelo(clase).mu, modelo(clase).sigma, X)$

$tab\_probs \leftarrow [tab\_probsprob]$

end for
{Get index of max value along each row}

$yhat = max(tab\_probs, [], 2);$

Para obtener la clase predicha, como en la práctica anterior, utilizaremos la estrategia de usar las funciones de Matlab para encontrar el índice del máximo según las líneas.

2. Memoria de la practica

2.1. Cuestión 1 : Entrenamiento y clasificación con modelos Gaussianos regularizados

En primer lugar, codificaremos las funciones propuestas como estudio preventivo de esta práctica. Para ello, utilizando la documentación de Matlab y las funciones anexas, podemos codificar las funciones solicitadas.

En esta fase no pensé que tuviera problemas. De hecho, al probar un pequeño conjunto de datos, todo funcionaba perfectamente. Sin embargo, cuando comencé a practicar la regularización (gran número de iteraciones, sobre los valores de lambda, y sobre los diferentes pliegues de datos), me di cuenta de que mis funciones estaban lejos de ser óptimas. De hecho, después de 15 minutos de espera, no obtenía ningún resultado. En particular, la función de clasificación no me permitía resolver los problemas en un plazo razonable. Después de buscar en mi código, esto fue debido al hecho de que calculaba las probabilidades de pertenecer a las clases de una manera inadecuada. Anteriormente, calculaba por dato, clase por clase, la probabilidad de que el dato perteneciera a la clase, y luego miraba qué clase me daba la mayor probabilidad. Fue muy largo y costoso en la funcion de tasa de acierto. La función gaussLog, que utiliza la distancia de Mahalanobis, resulta mucho más interesante, ya que moviliza el cálculo matricial.

Después de resolver estos problemas de optimización de código, mi trabajo finalmente estaba funcionando en grandes conjuntos de datos.

2.2. Cuestión 2 : Bayes ingenuo

En esta parte vamos a utilizar la clasificación Bayesiana enana. En otras palabras, supone la distribución normal de los atributos y la independencia en su distribución. Esto significa que la distribución de atributos x0 no afecta a la distribución de atributos x1 a X400 y así sucesivamente.

Usando las funciones codificadas anteriormente, obtenemos las siguientes conclusiones.

En primer lugar, aplicamos la regularización (reutilización del código de prácticas anteriores) para elegir el mejor valor de lambda. Para ello, hemos tenido en cuenta la métrica de la tasa de acierto. En efecto, para la misma justificación que en la práctica anterior, estamos ante un problema equilibrado en el sentido de que cada clase está representada uniformemente.

Hemos obtenido el siguiente gráfico.

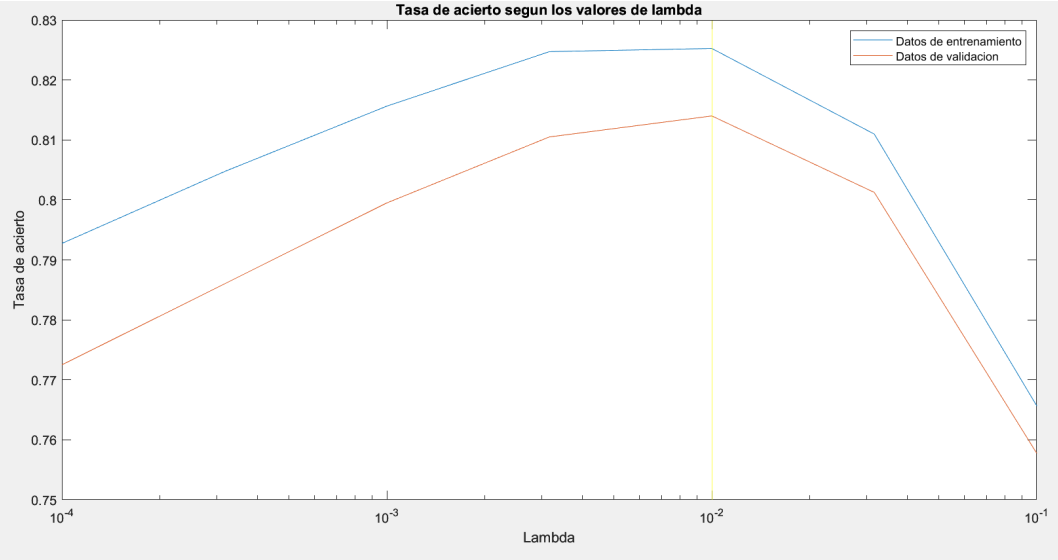

Figura 1: Tasa de acierto en función de lambda con Bayes ingenuo

El mejor valor de lambda es 10^{-2}. En efecto, es este valor el que permite obtener la mejor tasa de acierto. Por lo tanto, vamos a elegir y entrenar nuestro modelo con el conjunto de datos y este valor de lambda.

Ahora podemos evaluar el modelo y calcular la métrica adecuada asociada con el problema (tasa de acierto), ya sea con datos de entrenamiento o de test.

	Tasa de acierto
Datos de entrenamiento	0.8250
Datos de test	0.8120

Cuadro 1: Tasa de acierto con Bayes ingenuo

La tasa de certeza es bastante correcta. En efecto, es del orden del 80 %, ya sea con los datos de entrenamiento o de test. El resultado me parece bastante concluyente a la vista de las aproximaciones realizadas aquí. En efecto, suponemos la muestra gausiana y hacemos la hipótesis de independencia de los atributos. A modo de comparación, en la práctica anterior, obteníamos un índice de certeza del 97 % con los datos de entrenamiento y un índice de certeza del 87 % con los datos de test.

Hemos obtenido un resultado menos fiable aquí, pero sigue siendo razonable a la vista de las aproximaciones realizadas.

Tranquilizados por haber obtenido un modelo fiable desde el punto de vista de las aproximaciones realizadas, podemos observar algunas confusiones realizadas. Así, con los datos de entrenamiento, pudimos observar las siguientes confusiones:

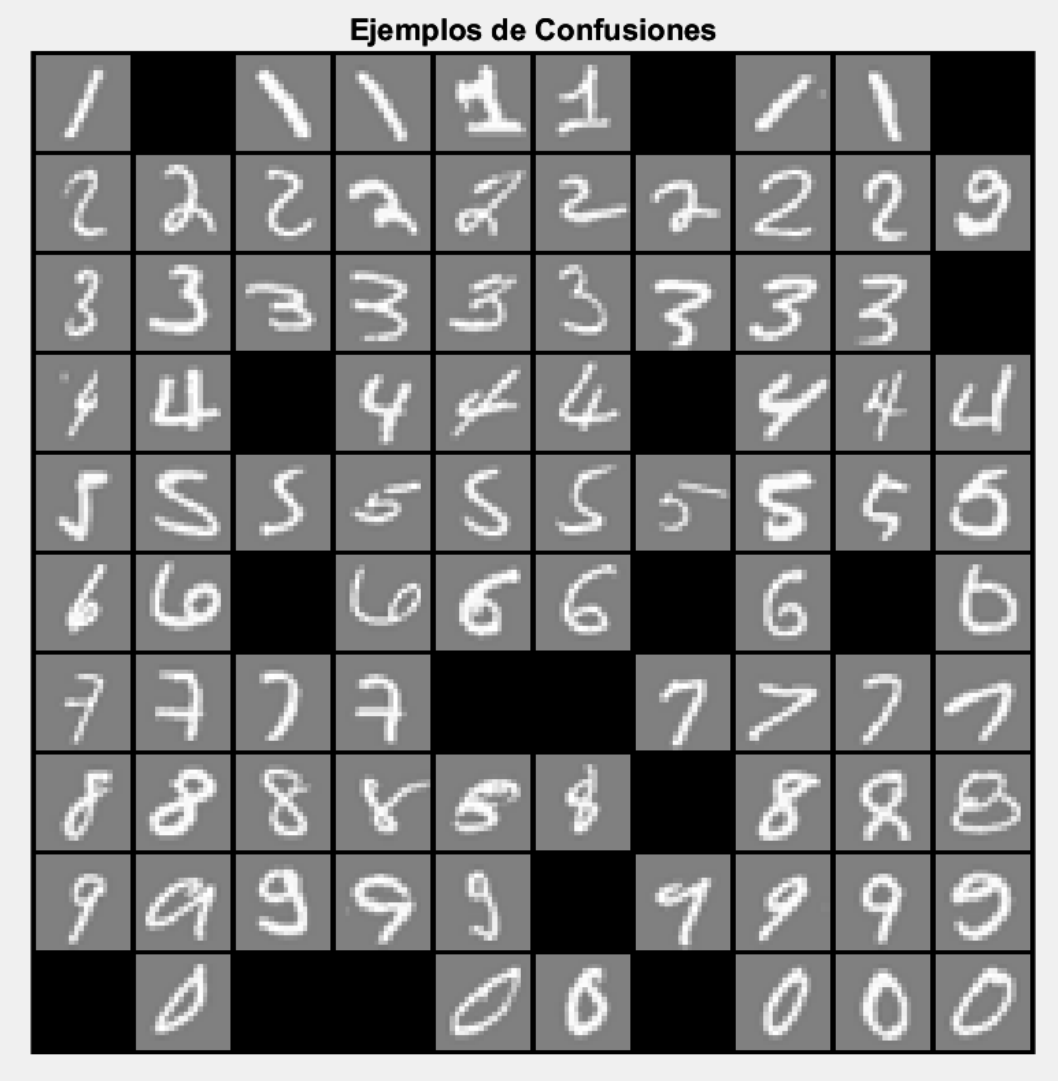

Figura 2: Ejemplos de confusiones con los datos de entrenamiento (Bayes ingenuo)

Con los datos de entrenamiento, observamos algunos errores en todas las clases.

Con los datos de test, podemos observar confusiones similares.

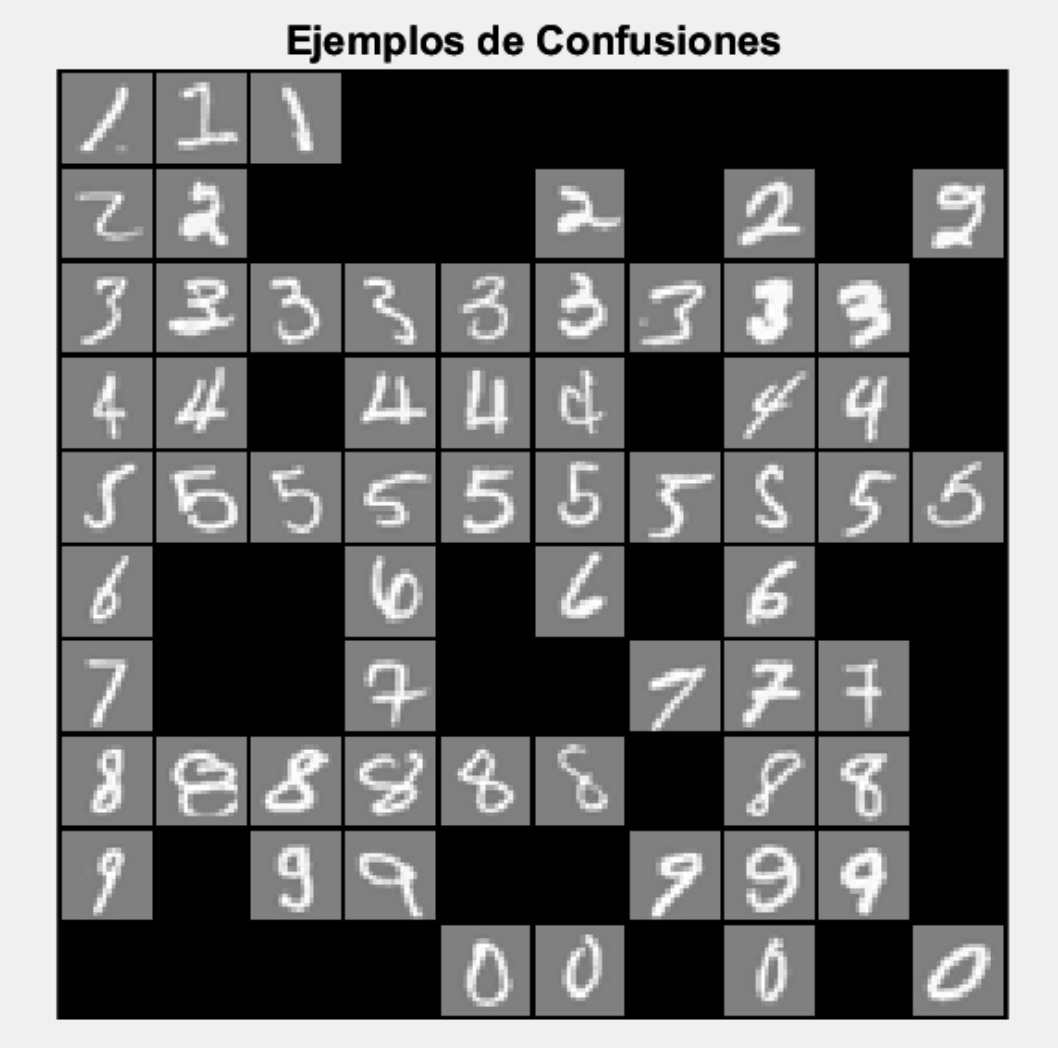

Figura 3: Ejemplos de confusiones con los datos de test (Bayes ingenuo)

Como en la práctica anterior, tenemos una idea del tipo de confusión realizada. Sin embargo, por el momento no sabemos con qué frecuencia hacemos estas confusiones. Por lo tanto, nos gustaría saber cuáles son las cifras más problemáticas para entender los posibles puntos de incertidumbre de nuestro modelo.

Para ello, vamos a construir una matriz de confusión. Para construirla, me ocupé en un primer momento de construirla mediante el cálculo por mis propios medios. Sin embargo, para obtener un resultado más agradable visualmente, he preferido para esta memoria el uso de la función confusionchart() disponible en Matlab. Esta función destaca los puntos predichos correctamente (la diagonal) y los errores cometidos. En la parte inferior de cada columna se asocia la métrica de la precisión, y a lo largo de cada fila se asocia el recall.

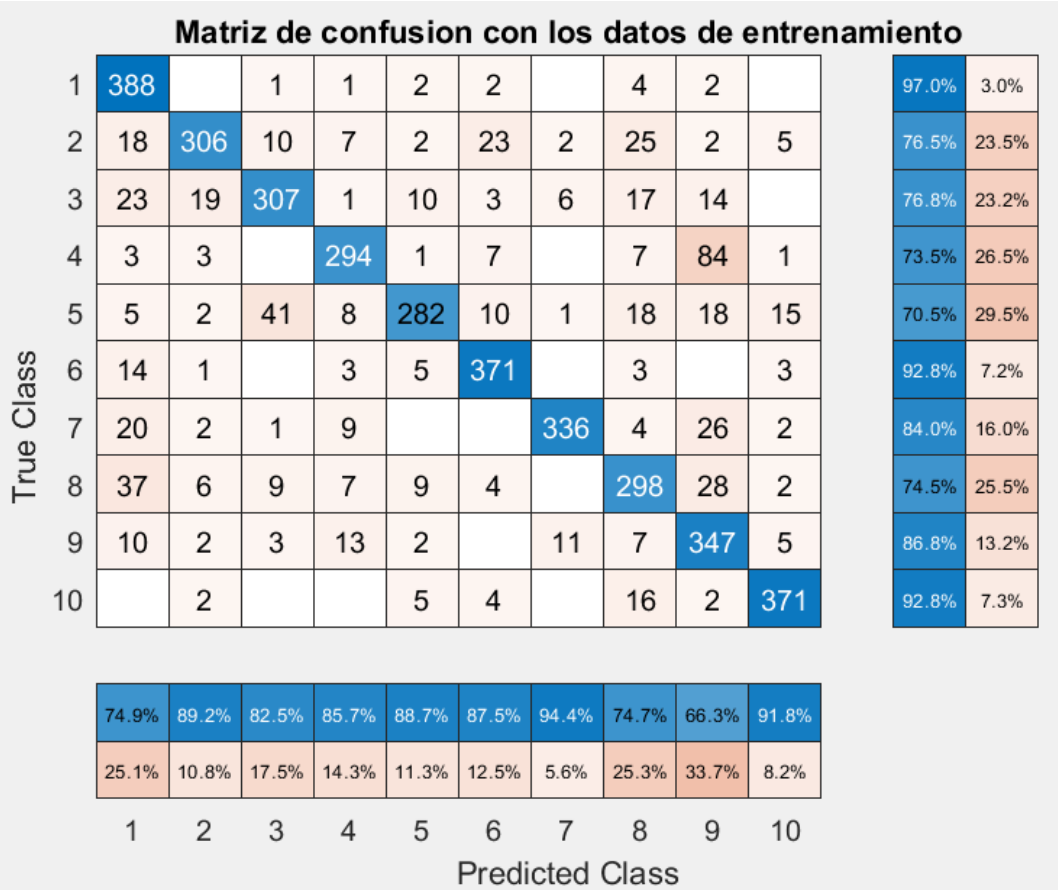

Figura 4: Matriz de confusión con los datos de entrenamiento (Bayes ingenuo)

Observamos que la mayoría de los datos se han predicho correctamente (valores altos en la diagonal). Sin embargo, la frecuencia de algunos errores me llama la atención. En particular, hemos observado 84 veces un error según el cual el modelo predijo la clase 9 en lugar de la clase 4. El segundo error más presente es la predicción de la clase 3 en lugar de la clase 5 (41 veces). Otros errores están presentes con menos frecuencia.

Ahora, vamos a ver la matriz de confusión con los datos de test para ver si estos errores frecuentes se repiten o no. Esto nos permitirá deducir las cifras más problemáticas para predecir.

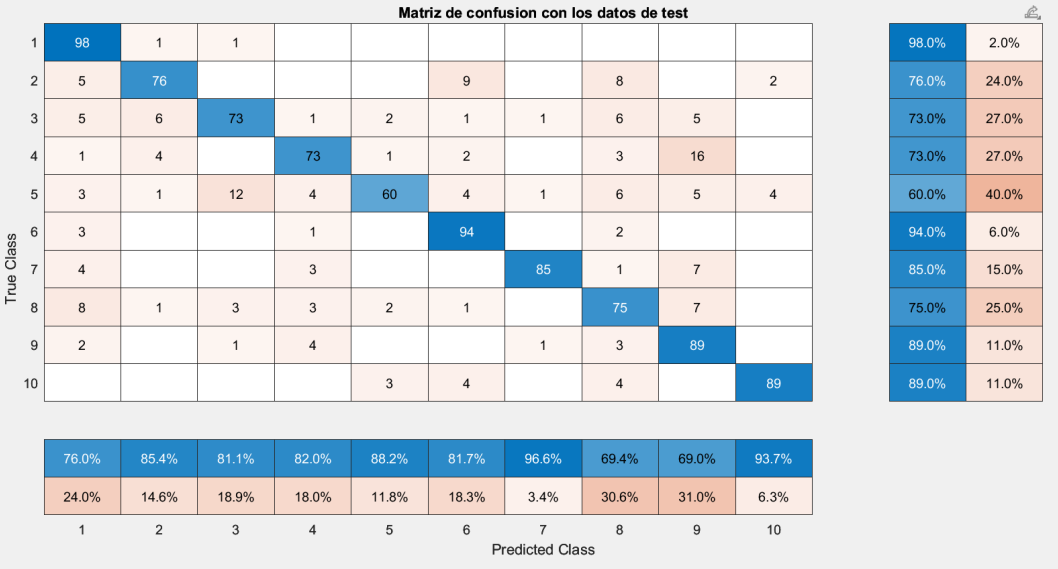

Figura 5: Matriz de confusión con los datos de test (Bayes ingenuo)

De nuevo, con los datos de test, observamos que la confusión más común es la predicción del número 9 en lugar del número 4 (16 veces). La confusión del número 5 con el número 3 también es visible de nuevo (12 veces).

Así, el estudio de las confusiones y de la matriz asociada nos permitió darnos cuenta de una confusión importante entre los números 9 y 4, y entre los números 3 y 5. La confusión entre los números 9 y 4 impacta fuertemente la precisión del modelo para la clase 9. En efecto, la precisión es sólo del 66 %, frente a más del 80 % en la mayoría de las demás clases.

Estas confusiones pueden tener varias justificaciones. La más sensata parece ser la relativa a las suposiciones para utilizar la clasificación bayesiana ingenua. En efecto, hemos supuesto una distribución independiente de los atributos, lo que no es el caso en la realidad.

2.3. Cuestión 3 : Covarianzas completas

Aquí repetiremos los pasos anteriores con el método de clasificación bayesiana completa. Este término debe entenderse en el sentido de que haremos menos aproximaciones. En particular, ya no asumiremos la independencia de la distribución de atributos, lo que parece ser una opción más realista que antes. Sin embargo, se espera que el tiempo de ejecución de la función sea mayor, ya que para calcular la matriz de covariancia, tendremos que calcular el valor de cada coeficiente, incluso fuera de la diagonal.

Usando de nuevo la función codificada en la pregunta 1 (solo hemos cambiado el parámetro Bayes ingenuo a 0 para practicar la clasificación bayesiana completa.

Comenzamos, en primer lugar, con la regularización. Esto nos permite saber cuál

es el mejor valor de lambda, para optimizar nuestro modelo.

Así conseguí el gráfico de la tarifa de acierto (métrica seleccionada) según el valor de lambda :

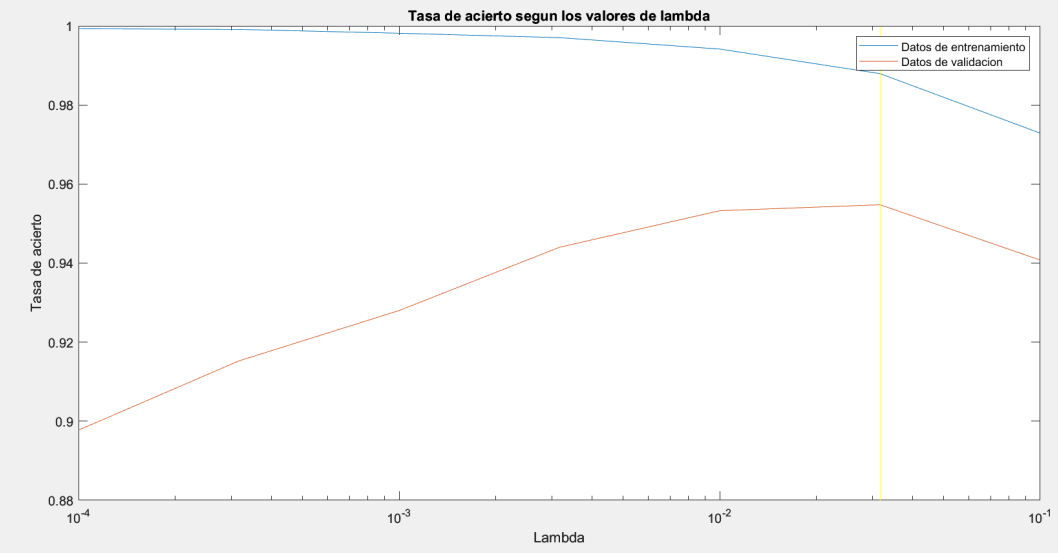

Figura 6: Tasa de acierto en función de lambda con Bayes completo

Así, conseguimos que el mejor valor de lambda sea $10^{-1,5}$. De hecho, con este valor obtenemos la tasa de acierto más alta con el conjunto de datos de validación.

Con el mejor modelo, ahora podemos evaluar la calidad de nuestra clasificación. Para ello, vamos a calcular la tasa de acierto con los datos de entrenamiento y validación.

	Tasa de acierto
Datos de entrenamiento	0.9865
Datos de test	0.9590

Cuadro 2: Tasa de acierto con Bayes completo

El resultado es mucho más concluyente que antes. En efecto, hemos obtenido una tasa de acierto más elevada, del orden del 95 %, con los datos de entrenamiento y de test. Esta clara mejora se explica por el hecho de que hacemos menos aproximación. En particular, ya no asumimos que la distribución de atributos es independiente.

Ahora queremos ver, como antes, cuáles son las confusiones más frecuentes de nuestro modelo. Por lo tanto, primero vamos a llamar a la función de la práctica anterior para ver los errores cometidos.

Con los datos de entrenamiento, obtenemos las siguientes confusiones:

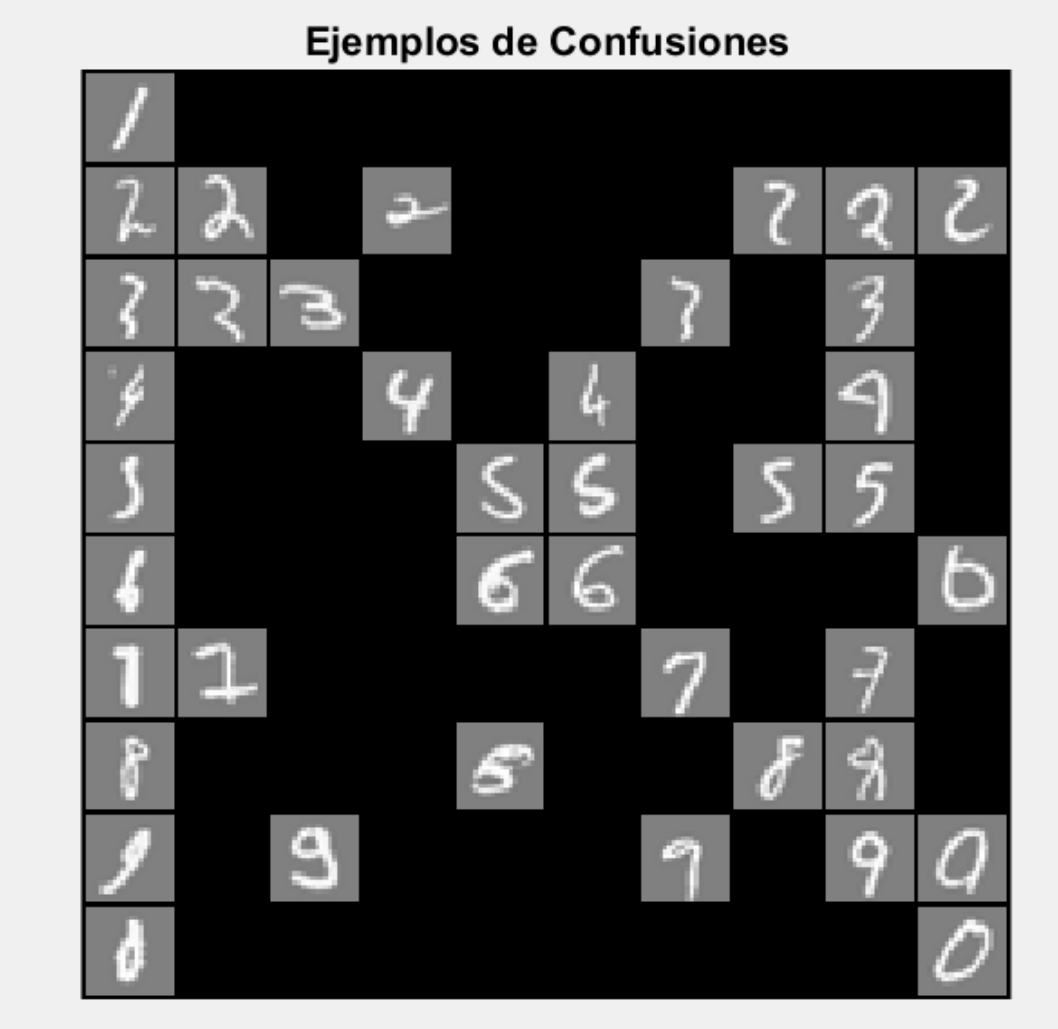

Figura 7: Ejemplos de confusiones con los datos de entrenamiento (Bayes completo)

Con los datos de test, obtenemos los siguientes errores:

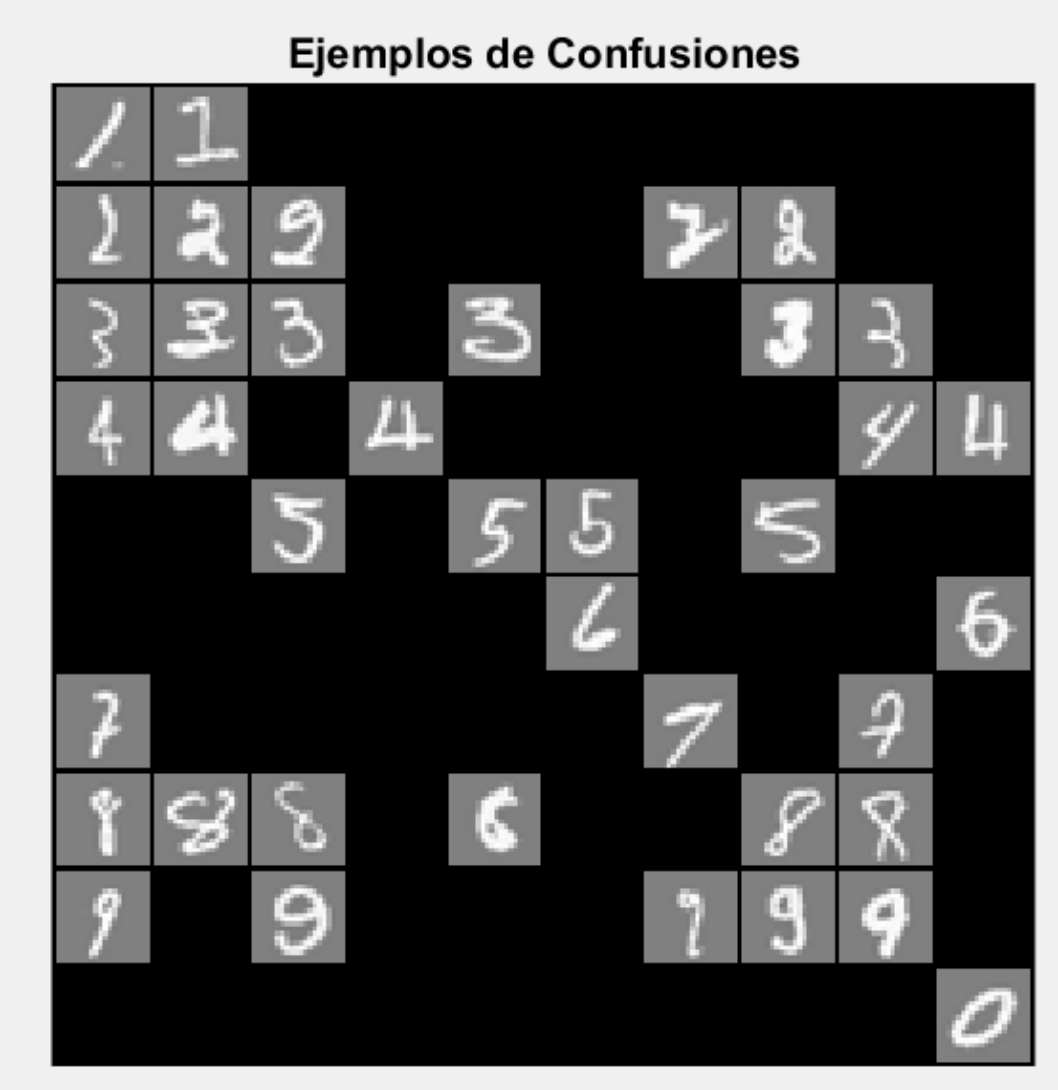

Figura 8: Ejemplos de confusiones con los datos de test (Bayes completo)

Ya sea con datos de entrenamiento o de test, notamos que nuestro modelo parece tener dificultades para predecir el número 1. De hecho, la primera columna está llena de imágenes, lo que significa que hemos cometido confusiones con cada una de las otras 9 clases.

Como antes, ahora queremos saber con qué frecuencia hemos cometido confusiones. Para ello, volveremos a dibujar la matriz de confusión, con los datos de entrenamiento y de test. Hemos obtenido las siguientes matrices:

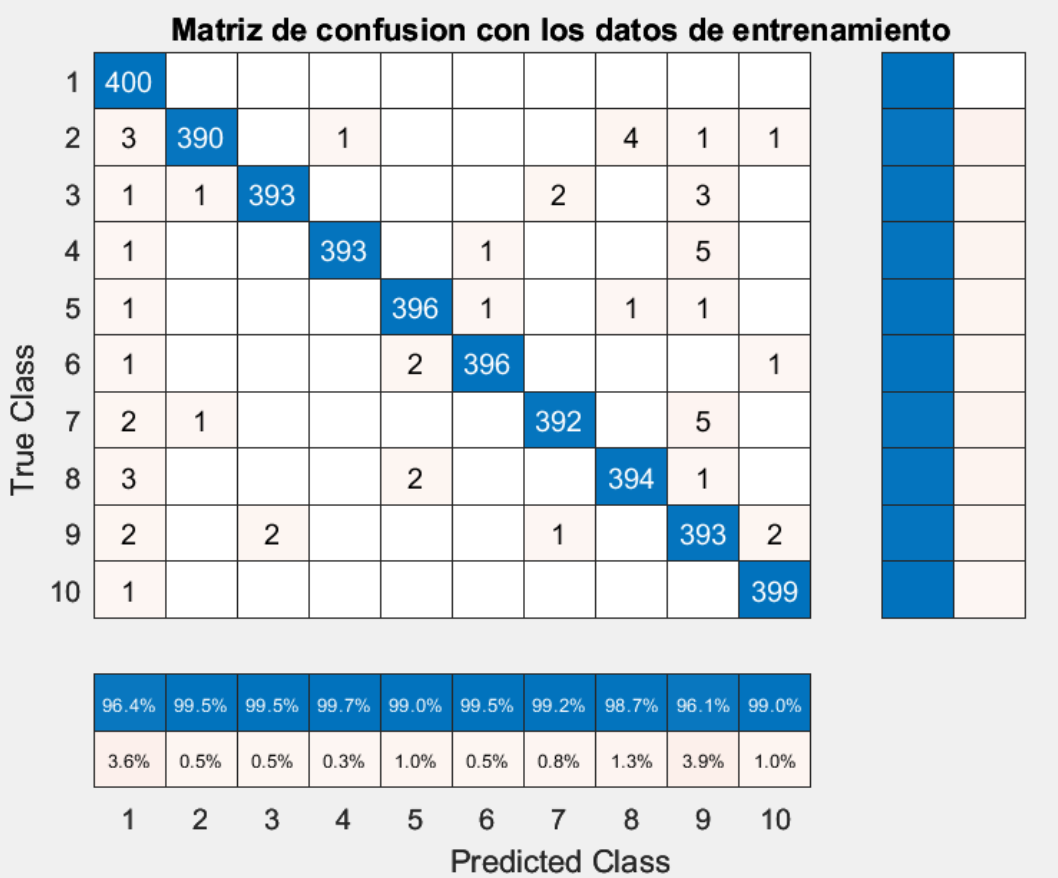

Figura 9: Matriz de confusión con los datos de entrenamiento (Bayes completo)

Con los datos de entrenamiento, el resultado es concluyente. De hecho, se cometen muy pocos errores, y con muy poca frecuencia. Esto se traduce en métricas de precisión y recall que son muy altas (por encima del 96 %).

Figura 10: Matriz de confusión con los datos de test (Bayes completo)

Para confirmar la validez de nuestro modelo, podemos hacer lo mismo con los datos de test. De nuevo, las conclusiones son similares. Se han cometido muy pocos errores, y las métricas de precisión y recall son muy altas (por encima del 96 %).

2.4. Comparación de modelos

Durante esta práctica hemos podido comprender y analizar las similitudes y diferencias de los métodos de clasificación. En particular, en la primera parte de esta práctica hemos podido establecer una comparación entre el ingenuo Bayes y el Bayes completo. En segundo lugar, ahora queremos comparar con los resultados de la práctica anterior. Esto nos permitirá comprender las ventajas y desventajas de los diferentes modelos estudiados.

En el cuadro siguiente se resumen las conclusiones extraídas de esta práctica y de la anterior:

	Datos de entrenamiento	Datos de test
Regresión logística	0.9745	0.8730
Bayes ingenuo	0.8250	0.8120
Bayes completo	0.9865	0.9590

Cuadro 3: Resumen de las conclusiones extraídas de los diferentes modelos

Observamos que el modelo con la tasa más baja de acierto es el propuesto con el ingenuo Bayes. De hecho, como se explicó anteriormente, su tasa de acierto más baja (82 %) se explica por el hecho de que hacemos muchas aproximaciones sobre los datos, en particular, la distribución gaussiana y la independencia de los atributos.

El segundo mejor modelo es el establecido con la regresión logística. Con una tasa de acierto elevada (87 %), este método de aprendizaje se basaba en la instauración de una frontera de decisión propia de cada cifra.

El mejor modelo es el establecido con el Bayes completo. De hecho, con una tasa de acierto de más del 95 %, este modelo se impone como el mejor. El método de predicción se establece según el procedimiento siguiente. Primero aprendemos la distribución de los atributos de cada clase (utilizando la media y la matriz de covariancia).

La obtención de estos modelos se hace a expensas de determinados costes. De hecho, he notado que el modelo con Bayes ingenuo es calculado por el ordenador en un tiempo mucho más limitado que el Bayes completo. Esto se debe a que Bayes ingenuo sólo calcula los términos diagonales de la matriz de covariancia. En cuanto al tiempo de ejecución, es similar entre la regresión logística y el Bayes completo. Se deduce que el Bayes completo parece más interesante aquí. En efecto, en presencia de un gran número de datos, se comete un error muy marginal al suponer la distribución gaussiana de los atributos (según el teorema central límite). Además, este método se basa únicamente en promedios y varianzas (valores para los que tenemos poca incertidumbre debido al gran número de datos), lo que reduce la incertidumbre de nuestro modelo.

Conclusión

Así pues, durante esta práctica hemos podido poner de relieve las ventajas y desventajas de los diferentes métodos de clasificación.

El ingenuo Bayes, basado en un mayor número de suposiciones como la distribución independiente de atributos, permite obtener un resultado potable en un tiempo muy bajo.

En cuanto al Bayes completo y a la regresión logística, más costosa debido a la complejidad de los cálculos solicitados y menos supuestos, son más precisos. Ambos métodos son equivalentes y permiten predecir con gran precisión la probabilidad de pertenecer a una clase.

El estudio de estos diferentes modelos de clasificación nos será útil en nuestra vida futura de ingeniería, ya que nos veremos obligados a manipular un gran número de datos y a clasificarlos de manera óptima, con el fin de mejorar el rendimiento de la empresa.

Índice de figuras

1. Tasa de acierto en función de lambda con Bayes ingenuo 6
2. Ejemplos de confusiones con los datos de entrenamiento (Bayes ingenuo) 7
3. Ejemplos de confusiones con los datos de test (Bayes ingenuo) 8
4. Matriz de confusión con los datos de entrenamiento (Bayes ingenuo) . 9
5. Matriz de confusión con los datos de test (Bayes ingenuo) 10
6. Tasa de acierto en función de lambda con Bayes completo 11
7. Ejemplos de confusiones con los datos de entrenamiento (Bayes completo) 12
8. Ejemplos de confusiones con los datos de test (Bayes completo) . . . 13
9. Matriz de confusión con los datos de entrenamiento (Bayes completo) 14
10. Matriz de confusión con los datos de test (Bayes completo) 15

Índice de cuadros

1. Tasa de acierto con Bayes ingenuo 6
2. Tasa de acierto con Bayes completo 11
3. Resumen de las conclusiones extraídas de los diferentes modelos . . . 16