Aprendizaje automático

Práctica 5. Clasificación Bayesiana

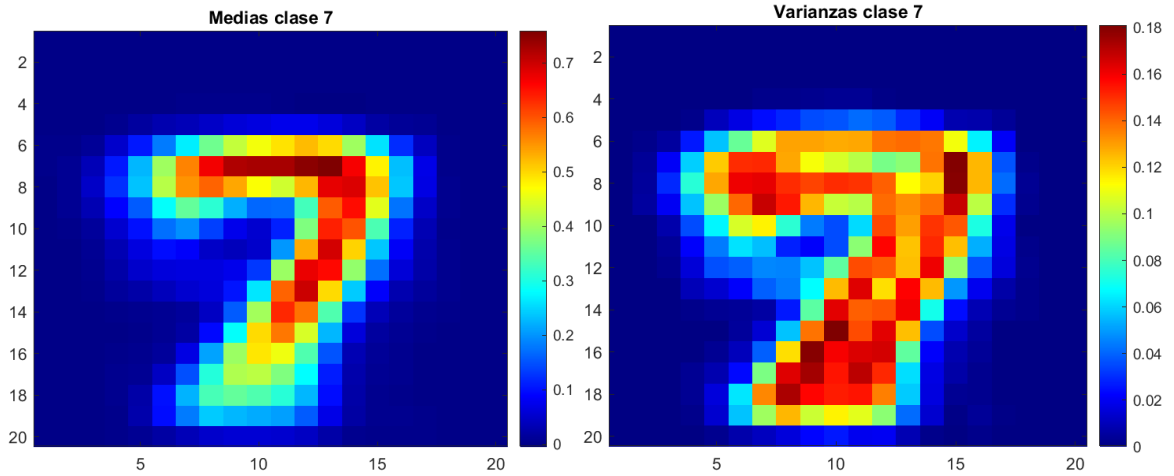
Universidad de Zaragoza, curso 2022/2023

Juan Eizaguerri Serrano 816079

Clasificación con modelo Gaussiano

En esta práctica se va a utilizar modelos generativos de clasificación Bayesiana para realizar una clasificación multiclase. Se va a utilizar el dataset MNIST con 5000 datos de 400 atributos que corresponden a la intensidad de cada píxel de la imágen 20x20 que representa. Se va a utilizar una partición de datos de 4000 para entrenamiento y 1000 para test.

Se ha preparado la función *entrenarGaussianas* para generar el modelo Gaussiano regularizado aprendiendo de unos datos de entrada X y salida Y, y parámetro regulador lambda. El entrenamiento consiste en calcular la media (μ) y covarianza (Σ) para los atributos de los datos de cada clase.



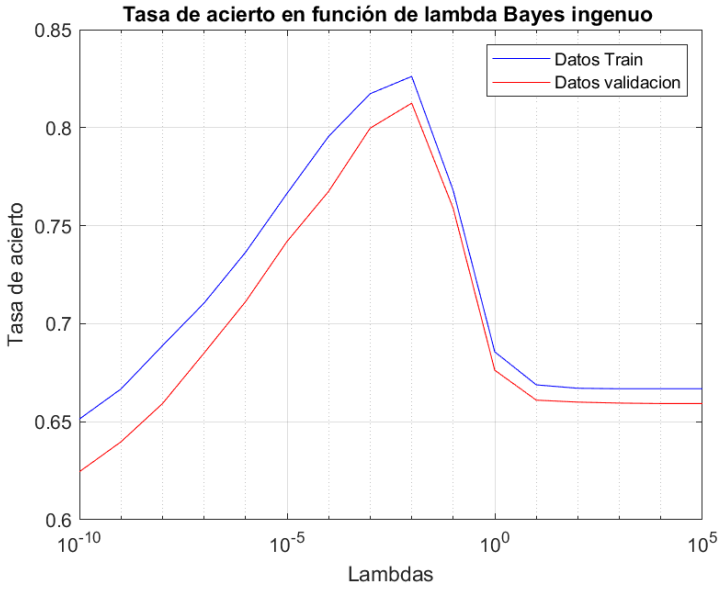
La primera imágen representa la media de los atributos para los datos de entrada de clase 7. La segunda es la varianza de dichos atributos, es decir, en qué medida varían respecto a la media. Se observa que los píxeles del centro de la figura son los de mayor media, sin embargo, la varianza es máxima en el contorno de la figura, ya que es el área en el que los datos tienen más variación.

También la función *clasificacionBayesiana*, que dado un modelo y una entrada de datos calcula la probabilidad logarítmica de que el dato pertenezca a cada clase en base a las μ y Σ del modelo y utiliza el método one-vs-all para decidir cuál de ellas es la más probable.

Bayes ingenuo

El primer modelo que se va a probar es el de bayes ingenuo, que asume la independencia condicional de los atributos y convierte su matriz de covarianzas en una matriz diagonal.

Se utiliza el algoritmo k-fold cross-validation para seleccionar el mejor parámetro de regularización λ. Se han probado distintos valores en progresión geométrica desde 10^{-10} y 10^{-5}. Se puede mostrar la evolución de la tasa de acierto en función de λ.

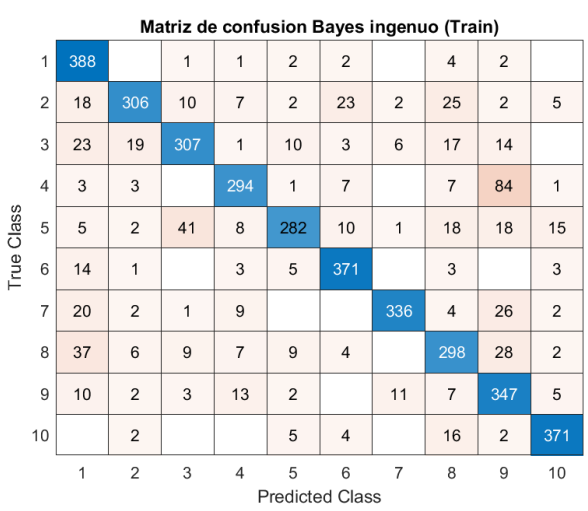


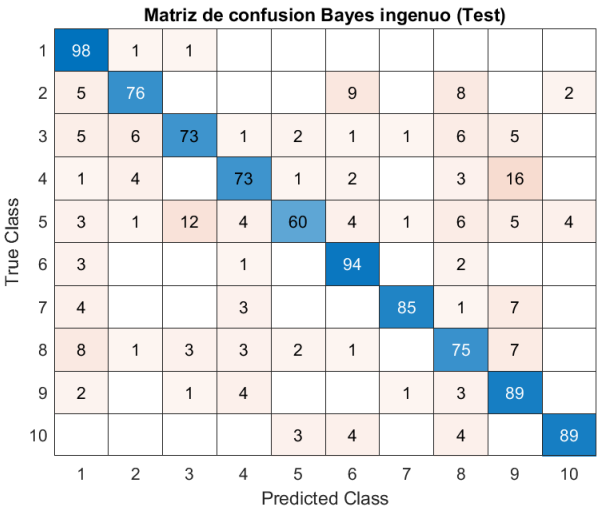
Para valores altos de λ se penalizan demasiado la estimación de las covarianzas y se produce su-ajuste, mientras que para valores demasiado altos el modelo no es suficientemente bueno como para producir buenas predicciones ni siquiera para los datos de entrenamiento.

El mejor λ encontrado es 10^{-2}, que obtiene una Ta media de 0.8261 para datos de test y 0.8125 para validación. Se entrena el modelo con todos los datos y usando este parámetro de regularización, obteniendo los siguientes resultados.

	Ta train	**Ta test**
Ingenuo reg	0.8250	0.8120

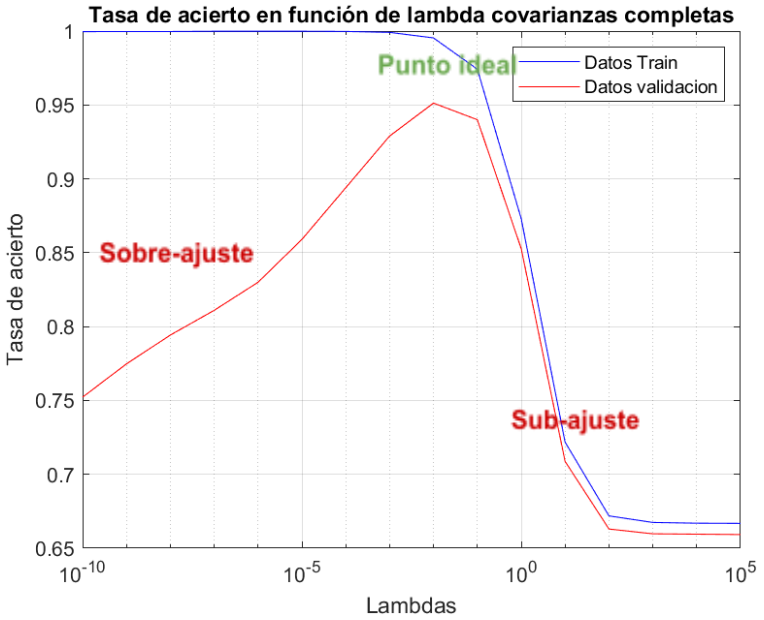
Se muestra también la matriz de confusión para ver qué clases se tienden a confundir entre sí.





Covarianzas completas

En este apartado se van a tener en cuenta las dependencias condicionales de los atributos, siendo Σ una matriz completa y no una diagonal como al utilizar Bayes ingenuo. Esto significa que se aprende de qué manera los atributos varían conjuntamente respecto a la media. De nuevo se utiliza el algoritmo k-fold cross-validation para seleccionar el mejor parámetro λ.

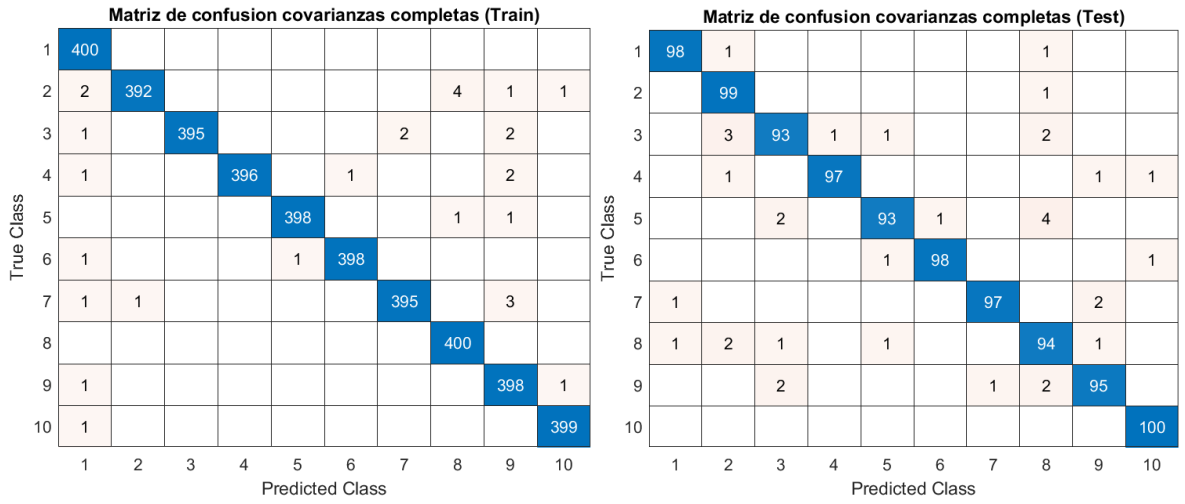


Para valores altos de λ se regulariza demasiado la estimación de las covarianzas y se produce un sub-ajuste en el que la tasa de acierto es mala tanto para los datos de entrenamiento como para los de validación. Sin embargo, para λ bajos la tasa de acierto para los datos de entrenamiento mejora hasta aproximarse al 1 mientras que la de los datos de validación disminuye, es decir, se produce sobreajuste.

De nuevo, la mejor λ encontrada es 10^{-2}, valor para el que se obtiene una Ta media de 0.9953 para entrenamiento y 0,9512 para validación. Se entrena el modelo con esta λ y todos los datos de entrenamiento.

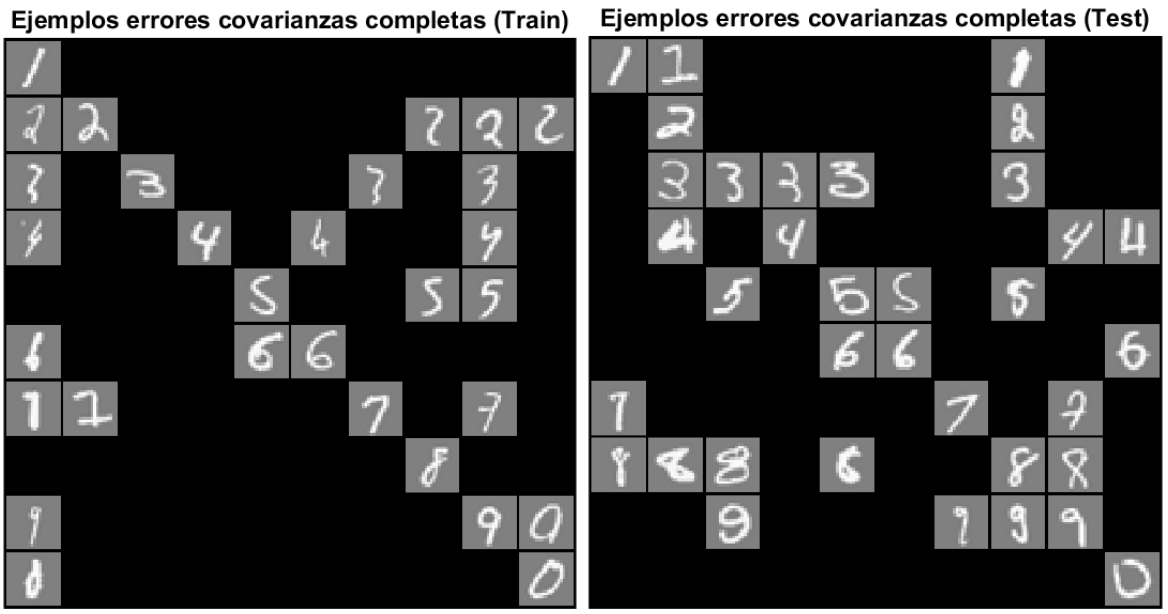
	Ta train	**Ta test**
Covarianzas completas	0.9927	0.9640

Se observan unos resultados mucho mejores que los obtenidos con Bayes ingenuo. A continuación se muestran también las matrices de confusión para este modelo.

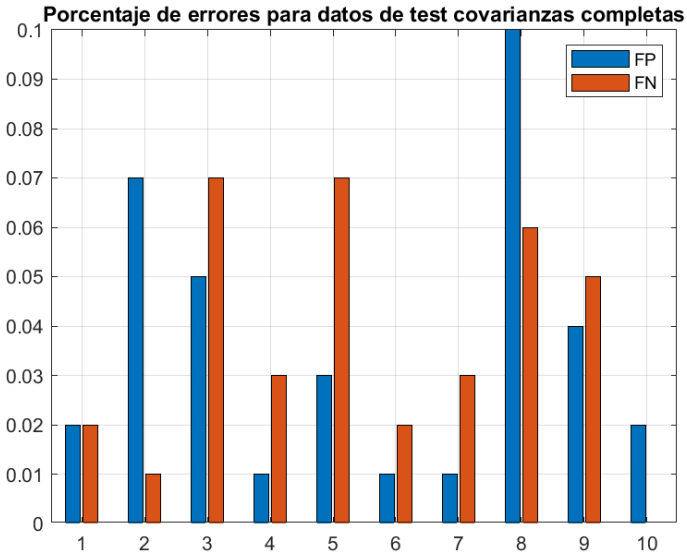


Se observa que hay muchos errores para la clase 8, se confunde especialmente con el 2 y el 5. En los datos de entrenamiento hay muchos falsos positivos para la clase 1 y 9, y entre los datos de entrenamiento también hay una cantidad considerable de falsos positivos para la clase 2.

A continuación se muestran también algunos ejemplos para cada celda de la matriz de confusión.



Se pueden representar la tasa de errores en un gráfico de barras.



Se corrobora la cantidad de falsos positivos para las clases 2 y 8. También hay muchos falsos negativos para las clases 3 y 5, y en menor medida 8 y 9.

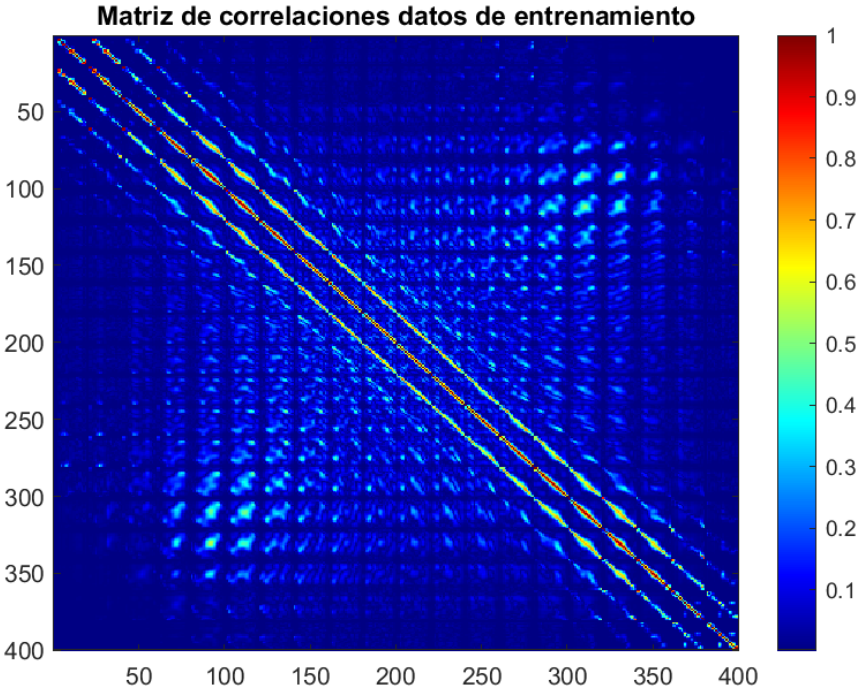
Se calcula también la precisión y el recall para los datos de entrenamiento y test, junto con la cantidad de datos de test clasificados como cada una de las clases para comprobar si el modelo tiene tendencia a predecir ciertas clases.

	pTest	**rTest**	**C**
1	0.9800	0.9800	100
2	0.9339	0.9900	106
3	0.9480	0.9300	98
4	0.9897	0.9700	98
5	0.9687	0.9300	96
6	0.9898	0.9800	99
7	0.9897	0.9700	98
8	0.9038	0.9400	104
9	0.9595	0.9500	99
0	0.9803	1	102

Las clases con más falsos positivos como el 2 y el 8 dan lugar a una precisión menor y las clases 3 y 5 que tienen más falsos negativos tienen el recall más alto.

Comparación de modelos

Se ha comprobado que para el problema de clasificación de imágenes de números utilizar clasificación Bayesiana con covarianzas completas da lugar a modelos mucho mejores que utilizando regresión logística o Bayes ingenuo. Esto se debe a que no se puede asumir la independencia condicional de los atributos. Podemos comprobarlo dibujando la matriz de correlaciones entre atributos.



Esta matriz muestra en qué medida cada uno de los atributos del eje x influye sobre el resto de atributos y sobre sí mismo en el eje y. Si los datos fuesen condicionalmente independientes la matriz de correlaciones sería una diagonal, pero se observa que efectivamente el valor de unos atributos dependen de otros.

Sabiendo esto, tiene sentido que Bayes ingenuo no dé lugar a buenos resultados para este problema. La regresión logística podría llegar a darlos utilizando expansión polinómica de atributos con productos cruzados, aunque el coste en tiempo de su entrenamiento sería considerablemente mayor que utilizando clasificación Bayesiana.

Por último se muestra una tabla con un resumen de las tasas de acierto recogidas.

	Ta train	**Ta test**
LogReg	0.9282	0.8900
LogReg Exp grado 3	0.9537	0.9090
Bayes ingenuo	0.8250	0.8120
Covarianzas completas	0.9927	0.9640