Aprendizaje automático

Práctica 3. Regresión logística

Universidad de Zaragoza, curso 2022/2023

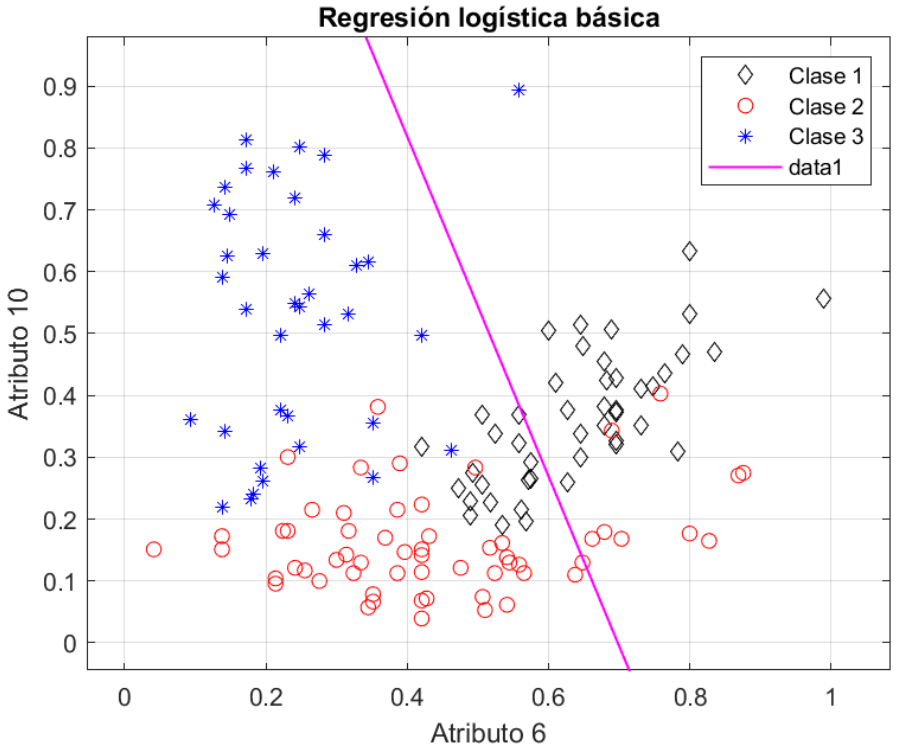
Juan Eizaguerri Serrano 816079

Regresión logística

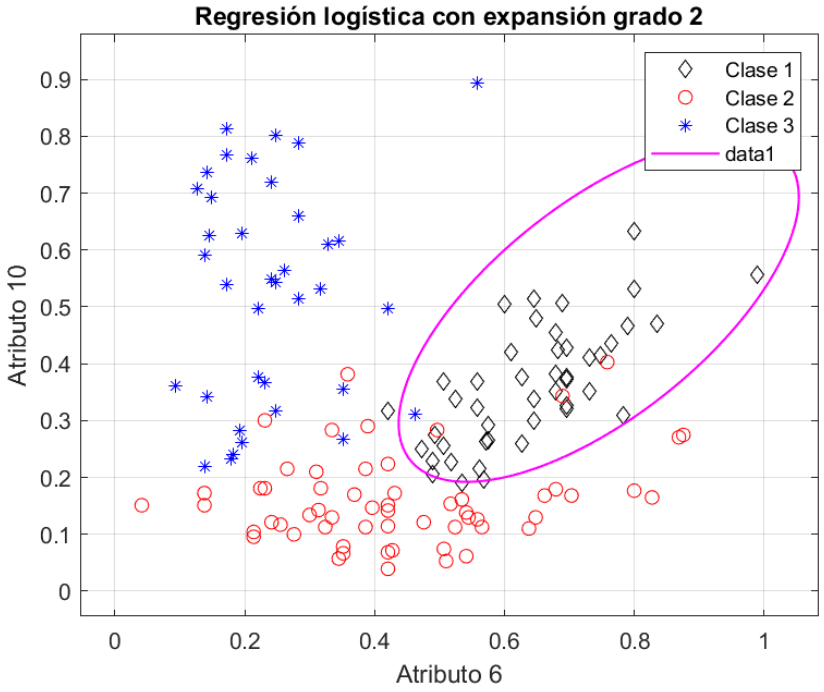
Modelos básicos

En esta práctica se va a utilizar el método de regresión logística para clasificación binaria de tipos de vinos en función de algunos de sus atributos. En concreto, se quiere detectar los vinos de clase uno en función de los atributos 6 y 10.

Realizando una regresión logística básica se minimiza la función de coste para calcular las θ de la ecuación de la recta que separa los vinos de clase 1 del resto.



En un primer vistazo a la representación de los datos, se observa que los vinos que se quiere detectar forman un cúmulo ovalado. Un buen método para encontrar un clasificador adecuado podría ser realizar expansión polinómica de los atributos con productos cruzados y grado 2, obteniendo el siguiente resultado.



Métricas de evaluación

Para evaluar los modelos desarrollados para este problema se va a utilizar la tasa de acierto A, que mide la proporción de datos clasificados correctamente, y la tasa de error E, que mide la proporción clasificada erróneamente. Está claro que A = 1-E.

Utilizando estas métricas para los dos modelos anteriores se obtienen los siguientes resultados.

		Tasa de acierto	**Tasa de fallo**
Regresión logística básica	Datos de entrenamiento	0.8099	0.1901
	Datos de test	0.8611	0.1389
Regresión logística con expansión polinómica	Datos de entrenamiento	0.9577	0.0423
	Datos de test	0.9444	0.0556

La regresión logística básica consigue una tasa de acierto del 80% con los datos de entrenamiento y 86% con los de test. Es extraño que los resultados con datos de test sean mejores que con los de entrenamiento, puede deberse a cómo se ha hecho la partición de datos, dejando datos más fáciles de clasificar en el set de prueba.

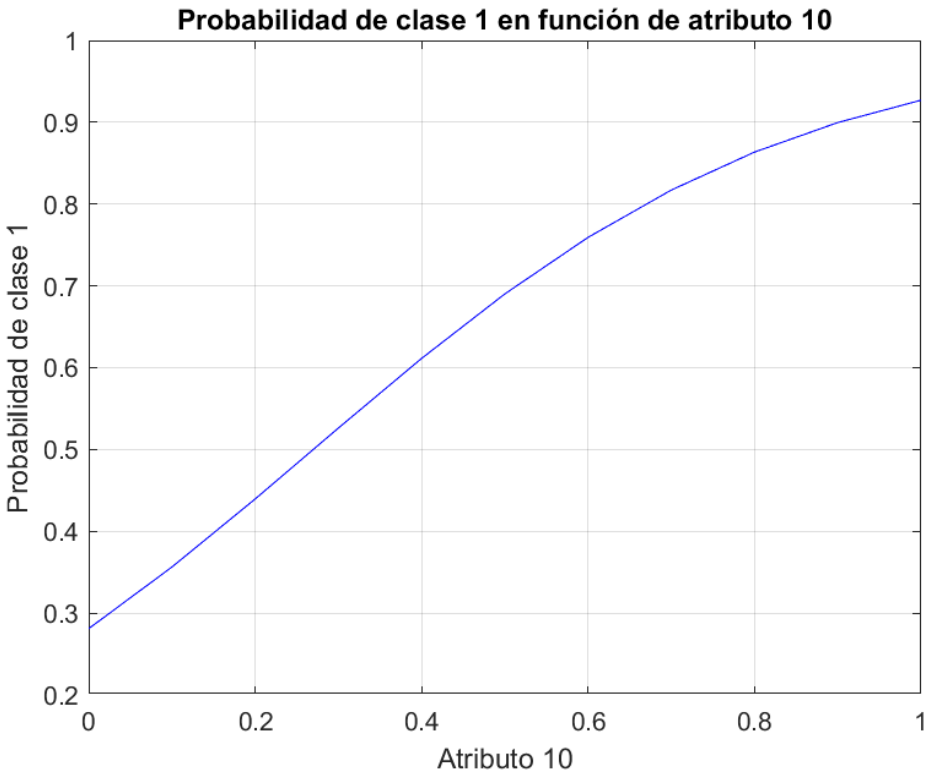
Se observa también que la regresión logística con expansión polinómica obtiene una tasa de acierto del 95% para los datos de entrenamiento y 94% para los datos de test. Aunque es un buen resultado, los grados del polinomio han sido escogidos a mano, por lo que es

interesante buscar el mejor grado o aplicar regularización como se hará en apartados posteriores de la práctica.

En ningún caso las métricas de entrenamiento son mucho peores que las de test, por lo que no hay sobre-ajuste.

Probabilidades en función de un atributo

Con el modelo desarrollado, es posible fijar uno de los atributos y calcular la probabilidad de que el vino sea de clase 1 para distintos valores del otro. Después, se visualizan las probabilidades calculadas en una gráfica. En concreto, en este apartado se fija el atributo 6 a 0.6, se calculan las probabilidades para los valores del atributo 10 en progresión de 0.1 y se obtiene el siguiente resultado.

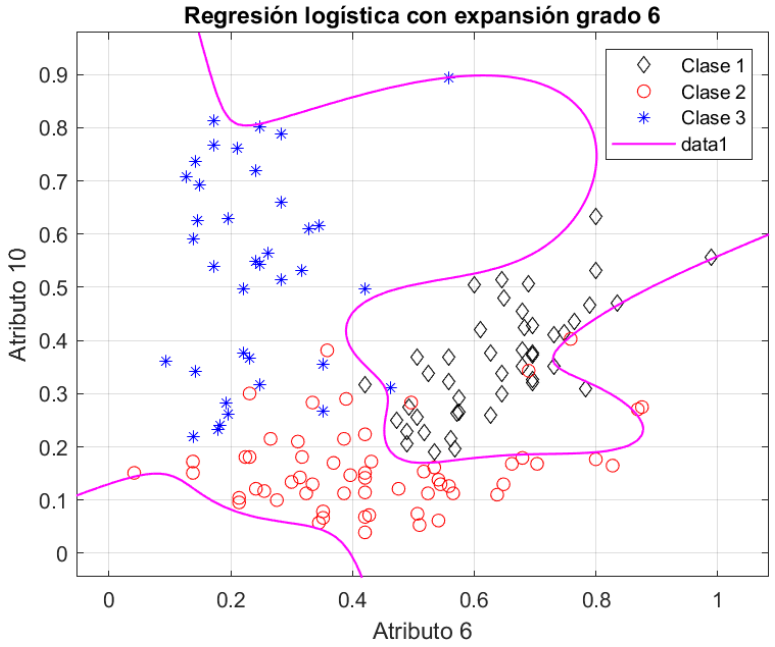


Se observa que el atributo 10 tiene una correlación positiva con la posibilidad.

Regularización

Expansión sin regularizar

En este apartado se va a utilizar expansión polinómica con productos cruzados de los atributos 6 y 10, utilizando grado 6. Si se entrena minimizando el coste logístico como en los modelos básicos se producirá un sobreajuste importante.

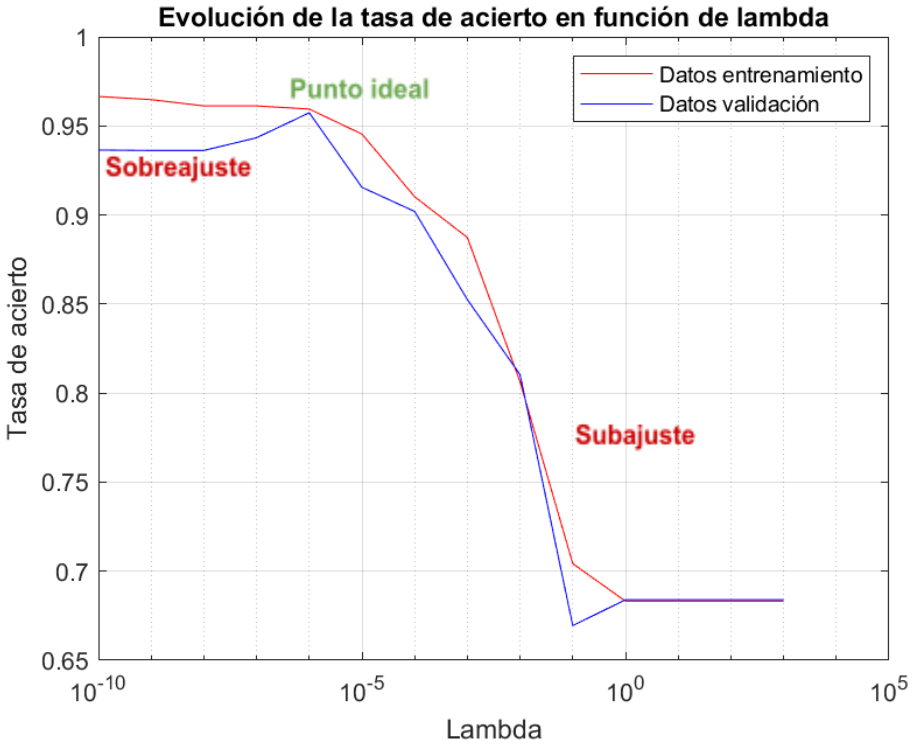


El modelo "memoriza" los datos de entrada y la superficie de separación obtenida se moldea demasiado a los datos. Como resultado las tasas de acierto sólo son buenas para los datos con los que se ha realizado el entrenamiento.

		Tasa de acierto	**Tasa de fallo**
Regresión logística con expansión grado 6	Datos de entrenamiento	0.9718	0.0281
	Datos de test	0.8055	0.1944

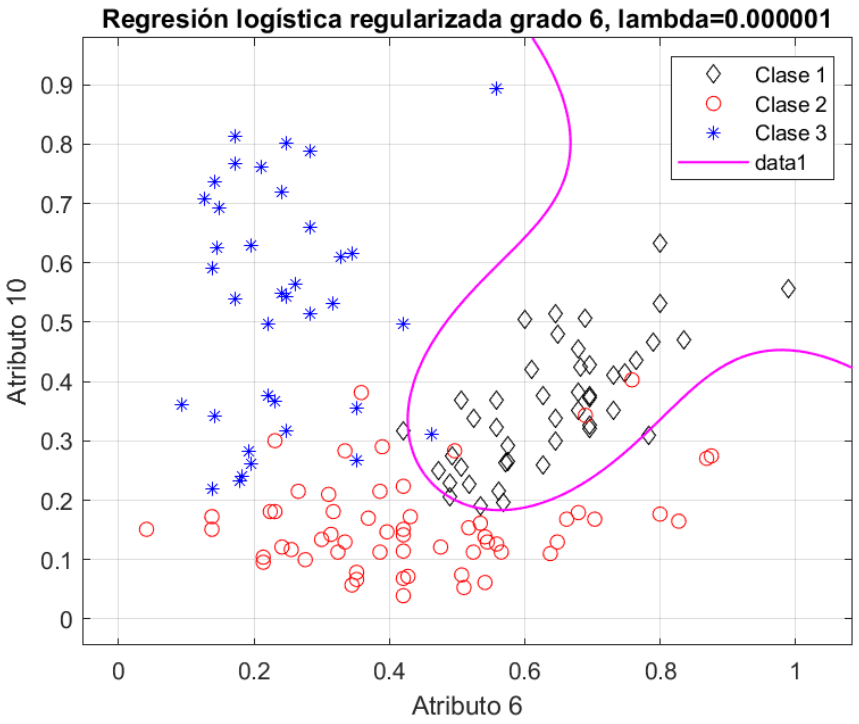
Regularización y selección de λ por k-fold

Para solucionar este problema se va a utilizar regularización para penalizar los modelos más complejos. Para encontrar un valor óptimo para el parámetro λ se va a utilizar el algoritmo de k-fold cross-validation, en el que se va minimizar el coste logístico regularizado para el entrenamiento. Utilizando k=5 y probando valores de λ en progresión geométrica entre 10^{-10} y 10^{3} se encuentra el valor 10^{-7}. Se puede comprobar la evolución de la tasa de acierto para los datos de entrenamiento y validación.



Se observa que para valores muy bajos de λ la penalización no es suficiente para evitar el sobre-ajuste, y para valores altos, se penaliza tanto las θ que el modelo entrenado no se adapta a los datos produciendo sub-ajuste. En el punto ideal hay una tasa de acierto de 0.9594 para los datos de entrenamiento y 0.9573 para los de validación.

Utilizando la λ decidida y entrenando con todos los datos de entrenamiento se obtiene un modelo que genera la siguiente frontera.



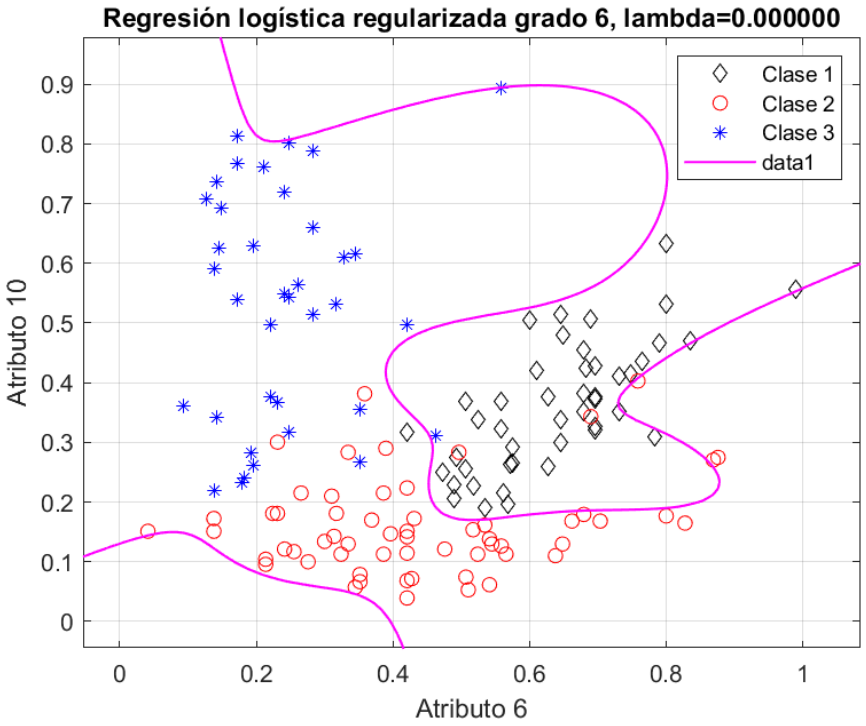
Se toman las tasas de acierto y error para los datos de entrenamiento y test.

		Tasa de acierto	Tasa de fallo
Regresión logística regularizada grado 6, λ=10^{-6}	Datos de entrenamiento	0.9577	0.0422
	Datos de test	0.9444	0.0555

Las tasas de acierto son exáctamente las mismas que para el modelo con expansión de grado 2 sin regularización, sin embargo, a diferencia del método manual, con la regresión regularizada y la selección de atributos mediante k-fold se asegura un resultado óptimo de forma automática.

Regresión logística regularizada utilizando λ=0

Entrenar el modelo utilizando lambda=0 en lugar de la mejor obtenida resulta idéntico a utilizar regresión sin regularización.



		Tasa de acierto	Tasa de fallo
Regresión logística regularizada grado 6, λ=0	Datos de entrenamiento	0.9718	0.0281
	Datos de test	0.8055	0.1944

Al tratarse de modelos idénticos se pueden extraer las mismas conclusiones. Al haber sobreajuste la frontera es demasiado cercana a los datos y se moldea demasiado a ellos.

Probabilidades en función de un atributo con regularización

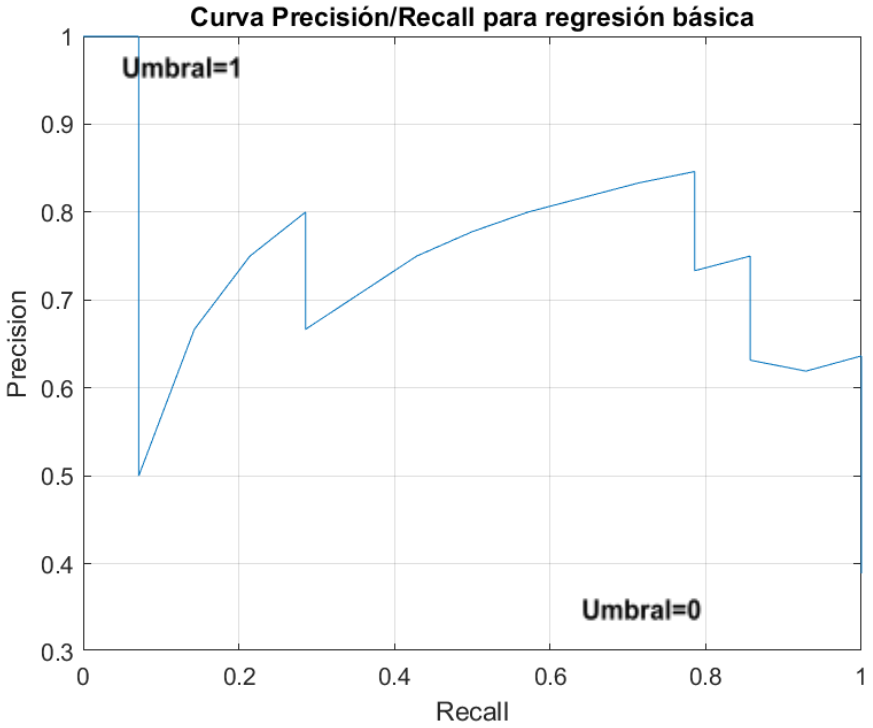
Se va a repetir esta prueba utilizando regularización y la λ seleccionada. Con el atributo 6 fijado a 0.6 se calcula la probabilidad de clase 1 para distintos valores del atributo 10 en progresión linear cada 0.1.

Al utilizarse productos cruzados para la expansión, a mayor valor del atributo 10 más interacción hay con otros atributos.

Curvas Precisión/Recall

Se ha desarrollado una función que muestra la curva precisión/recall calculando estos valores para los umbrales entre 0 y 1.

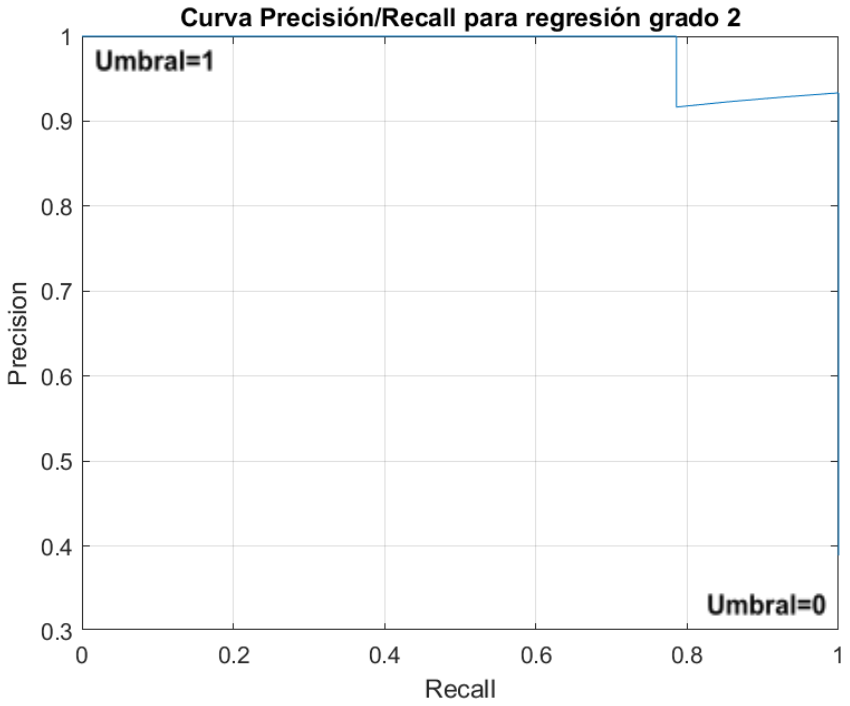
Se va a generar la curva para los distintos modelos diseñados a lo largo de la práctica con los datos de test. Para la regresión básica genera la siguiente gráfica.



Los datos de más a la izquierda los que tienen umbral=1. Donde el modelo no devuelve positivo para ninguna entrada, debido a esto no hay ningún falso positivo por lo que la precisión es máxima, pero sí que hay falsos negativos, así que el Recall será 0.

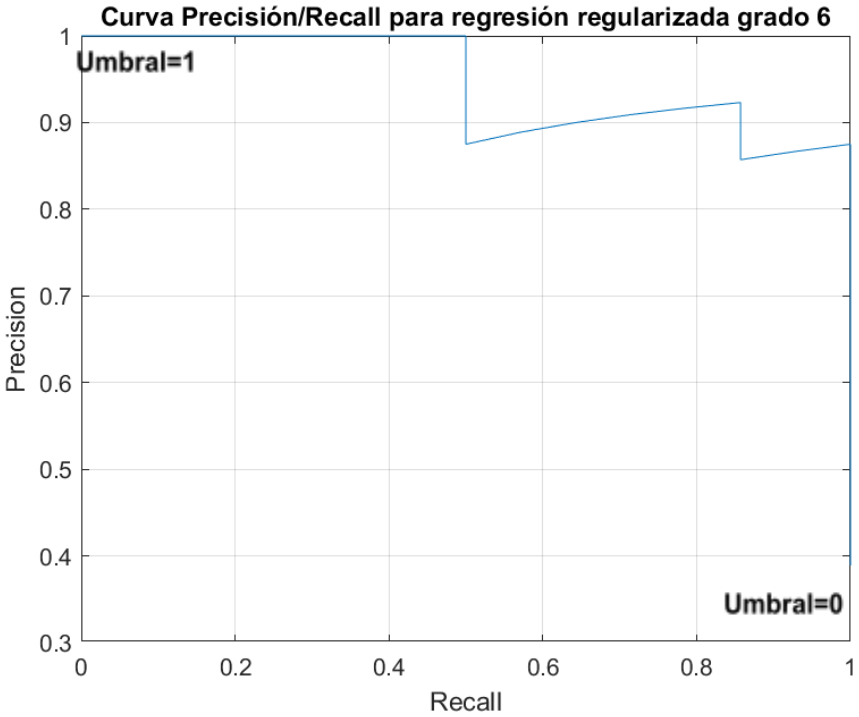
En el otro extremo está el umbral=0, donde se devuelve positivo para cualquier entrada. Debido a esto no habrá falsos negativos haciendo el recall 1, pero la precisión será 0 porque no hay ningún verdadero positivo.

Para la regresión con expansión grado 2:



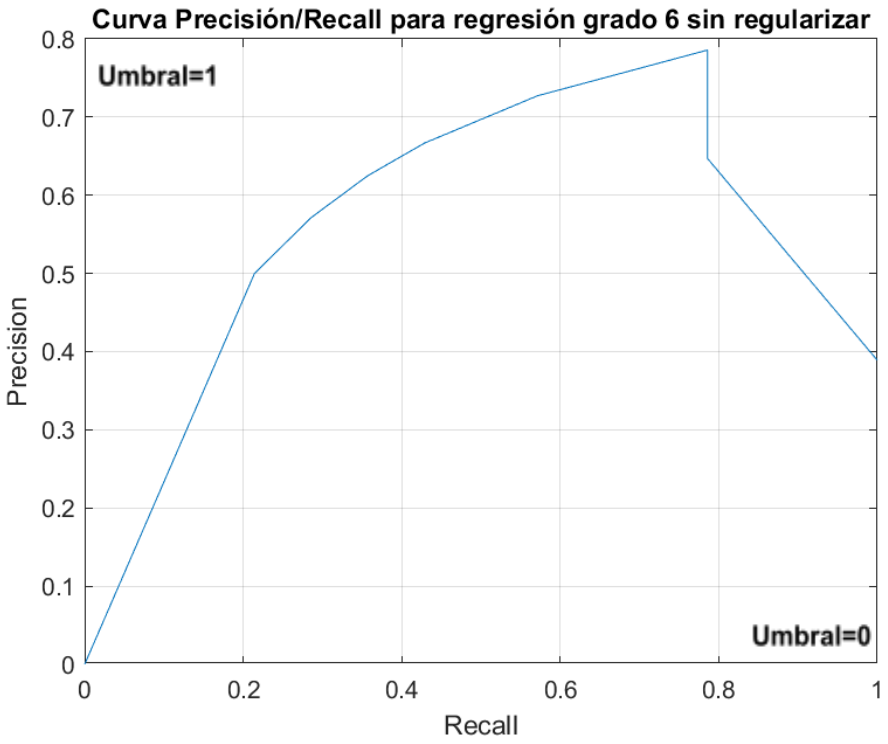
Este modelo es mucho mejor que el anterior, ya que puede mantener una precisión máxima con un recall de hasta casi 0.8.

Para la regresión de grado 6 con regularización:



Es importante destacar que, aunque este modelo tiene la misma tasa de acierto que el anterior, tienes curvas precisión/recall distintas, siendo esta segunda ligeramente peor, pues la precisión empieza a decaer antes y es peor cuando el recall es muy alto.

Por último se muestra la curva precisión/recall para la regresión logística regularizada de grado 6 y λ=0 (Equivalente a la regresión de grado 6 sin regularizar).



Se observa que debido al sobreajuste de este modelo y la cercanía de la superficie de separación a los datos se necesita un umbral muy bajo para que consiga buenos resultados.

Puede haber algunas dudas a la hora de comparar algunos de los modelos entre ellos. Para hacer esta validación cruzada se podría calcular el f_1 score, aunque está claro que el mejor modelo es el de la regresión logística de grado 2.

En el enunciado se pide asegurar que el 90% de los vinos clasificados como clase 1 realmente lo sean. Esto es que *TP / TP + FP*, es decir, la precisión, sea mayor o igual que 0.9.
El mejor modelo para esta tarea será el que tenga mayor recall para el que la precisión sea igual o mayor que la deseada. Se modifica la función *plotPrecisionRecall* para que devuelva este valor. Este es el valor encontrado para cada uno de los modelos.

	Regresión básica	**Regresión grado 2**	**Regresión regularizada grado 6**	**Regresión grado 6 sin regularizar**
Mejor recall para precisión >= 0.9	0.07142	1	0.8571	0

El mejor modelo es el de regresión logística con expansión polinómica de productos cruzados grado 2.

Conclusiones

En esta práctica se han desarrollado distintos modelos de regresión logística para la clasificación de vinos de clase 1 en función de dos atributos. Se ha utilizado regresión básica, con expansión polinómica de productos cruzados sin regularizar, y con expansión regularizando.

Se ha probado que se puede obtener un muy buen resultado manualmente, y también que utilizando regularización y buscando un λ óptimo mediante k-fold se puede llegar a un resultado igual de bueno de forma automática.

También se ha explorado de qué manera afecta en este problema un atributo en la tasa de acierto cuando se fija el otro atributo a un valor fijo.

Por último, se ha explorado la utilidad de las métricas de precisión y recall, que resuelve el problema de clases sesgadas, que no es significativo en el problema de esta práctica, pero puede ser interesante aplicarlo cuando la clase que se quiere diferenciar tiene una tasa de aparición baja frente a las demás. Se ha desarrollado una función que muestra la curva precisión/recall calculando estos valores para distintos umbrales posibles, y se ha propuesto una solución al problema de encontrar el mejor modelo que obtenga un recall de 0.9 al clasificar vinos de clase 1.

A modo de resumen, estas han sido todas las métricas de tasas de acierto y error que se han recogido a lo largo de la práctica para los distintos modelos probados.

		Tasa de acierto	Tasa de fallo
Regresión logística básica	Datos de entrenamiento	0.8099	0.1901
	Datos de test	0.8611	0.1389
Regresión logística con expansión grado 2	Datos de entrenamiento	0.9577	0.0422
	Datos de test	0.9444	0.0555
Regresión logística con expansión grado 6 (Sin regularizar)	Datos de entrenamiento	0.9718	0.0281
	Datos de test	0.8055	0.1944
Regresión logística regularizada grado 6, $\lambda=10^{-6}$	Datos de entrenamiento	0.9577	0.0422
	Datos de test	0.9444	0.0555
Regresión logística regularizada grado 6, $\lambda=0$	Datos de entrenamiento	0.9718	0.0281
	Datos de test	0.8055	0.1944

El tercer y el quinto modelo obtienen los mismos resultados porque son idénticos, mientras que el segundo y el cuarto no lo son. Las tasas coinciden pero difieren en el dibujo de las superficies que generan y en las curvas precisión recall.