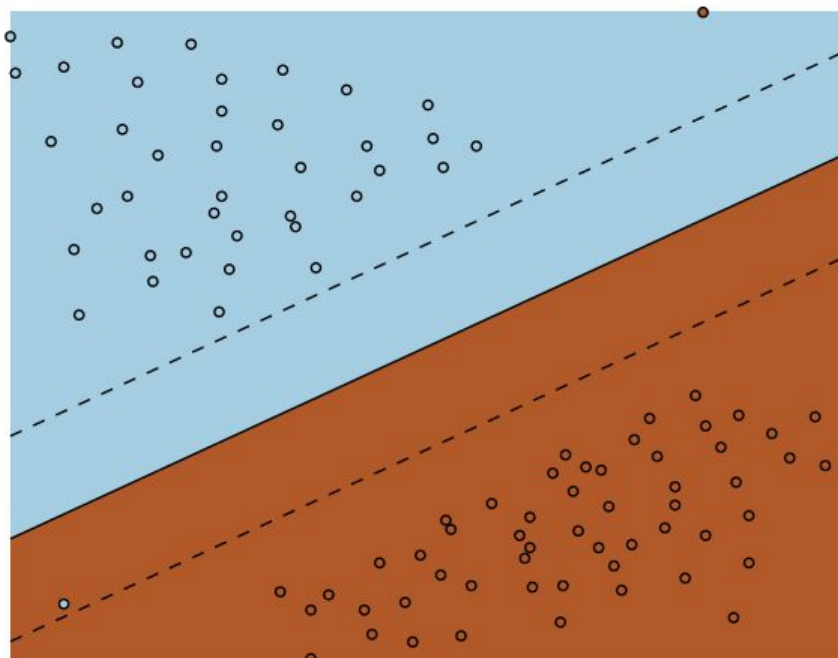


PRÁCTICA 4: MÁQUINAS DE VECTORES SOPORTE (SVM).



Víctor Monserrat Villatoro
i32moviv@uco.es
45887876R



1. Preguntas.

1.1.	Pregunta 1.	3
1.2.	Pregunta 2.	4
1.3.	Pregunta 3.	5
1.4.	Pregunta 4.	5
1.5.	Pregunta 5.	6
1.6.	Pregunta 6.	7
1.7.	Pregunta 7.	8
1.8.	Pregunta 8.	9
1.9.	Pregunta 9.	11
1.10.	Pregunta 10.	12
1.11.	Pregunta 11.	12
1.12.	Pregunta 12.	12
1.13.	Pregunta 13.	13
1.14.	Pregunta 14.	13
1.15.	Pregunta 15.	14
1.16.	Pregunta 16.	14
1.17.	Pregunta 17.	20

2.	Referencias bibliográficas.	22
----	-----------------------------	----



1.1. Pregunta 1.

Abre este script y explica su contenido. Podrás ver que se utiliza el primer dataset de ejemplo y se realiza la representación gráfica del SVM. Comenta que tipo de kernel se está utilizando y cuáles son los parámetros de entrenamiento.

En primer lugar se carga el primer conjunto de datos con 'load_svmlight_file', que carga ficheros en formato 'svmlight/libsvm' en una matriz dispersa comprimida.

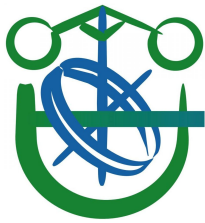
A continuación, se convierte esta en formato denso con el método 'todense' y se transforma en una matriz de 'numpy' para el entrenamiento del modelo.

Para la creación y configuración del modelo SVM se utiliza el constructor 'svm.SVC' de la librería 'sklearn'.

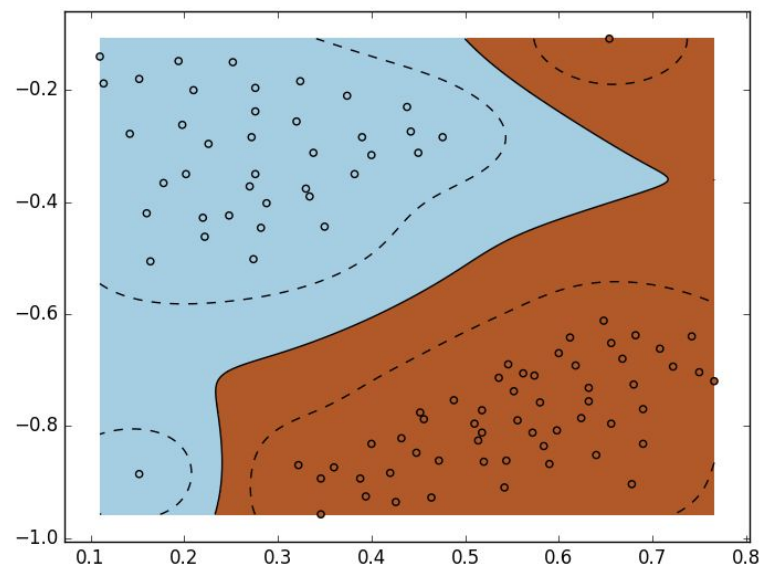
Los parámetros considerados para el entrenamiento de este han sido C (coste del kernel) igual a 1, que define cuánto vamos a penalizar el hecho de que se cometan errores en la clasificación, y gamma (ancho del kernel) igual a 100, que es el coeficiente para el tipo de kernel usado. Se ha utilizado un tipo de kernel 'rbf' o de función de base radial.

Un valor de C muy grande resultará en una SVM que cometerá el mínimo número posible de errores, mientras que un valor pequeño de C favorecerá el que se obtenga un clasificador con el máximo margen, aunque se cometan errores.

Para un kernel 'rbf', gamma es la inversa del radio. Un radio alto tiende a soluciones más suaves, con menor sobre-entrenamiento, mientras que un radio de kernel pequeño tiende a producir mayor sobre-entrenamiento.

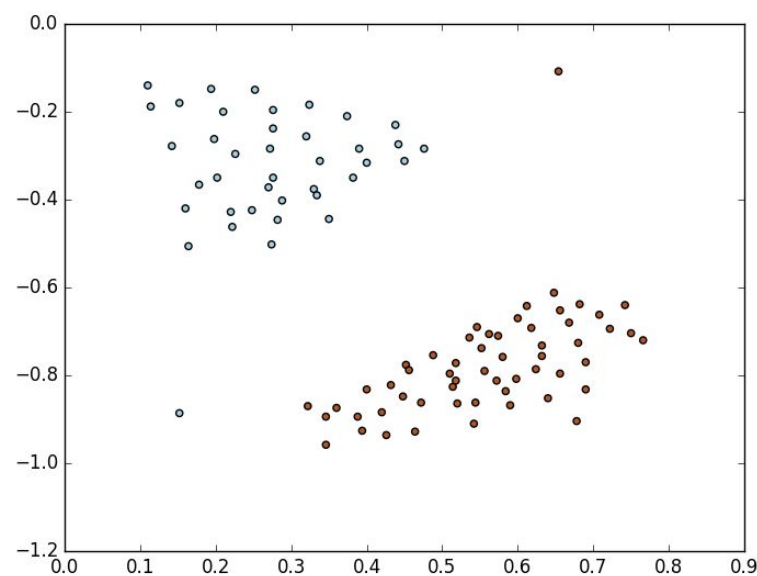


Finalmente, se representan gráficamente los patrones de entrenamiento y los valores devueltos por la SVM con colores a través de 'matplotlib.pyplot' como se observa en la siguiente figura.



1.2. Pregunta 2.

Intuitivamente, ¿que hiperplano crees que incurrirá en un menor error de test en la tarea de separar las dos clases de puntos?



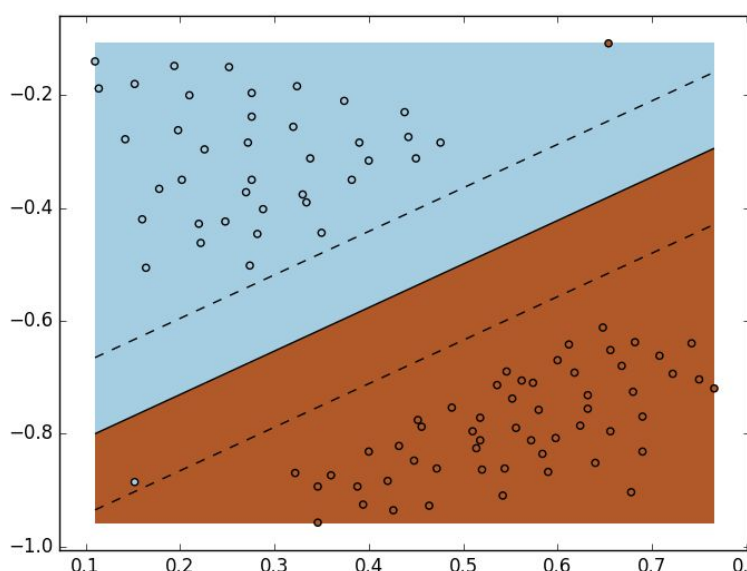
El hiperplano que incurriría en un menor error de test sería aquel de igual distancia a cada una de las nubes de puntos.



1.3. Pregunta 3.

Debes de configurar el script para que se ejecute una SVM lineal (es decir, que el kernel sea lineal).

Considerando una SVM lineal, para el primer conjunto de datos obtenemos el siguiente modelo.



Se puede observar que ahora sí se cometen errores de entrenamiento en la clasificación de algunos patrones.

1.4. Pregunta 4.

Modifica el script para que se prueben varios valores de C , en concreto, $C \in \{10^{-2}, 10^{-1}, 10^0, 10^1, 10^2, 10^3, 10^4\}$. Observa que sucede, explica por qué y escoge el valor más adecuado.

Podemos observar que a medida que se incrementa el valor del parámetro C se va obteniendo un clasificador con un margen mayor, cometiendo menor error. Esto se debe a que el parámetro C mayor penaliza más el hecho de que se cometan errores en la clasificación.

Figura $C \in \{10^{-2}, 10^{-1}, 10^0, 10^1, 10^2, 10^3, 10^4\}$

En mi opinión, el valor más adecuado para C sería de 0.1 aun cometiendo errores en la clasificación de los patrones aislados.

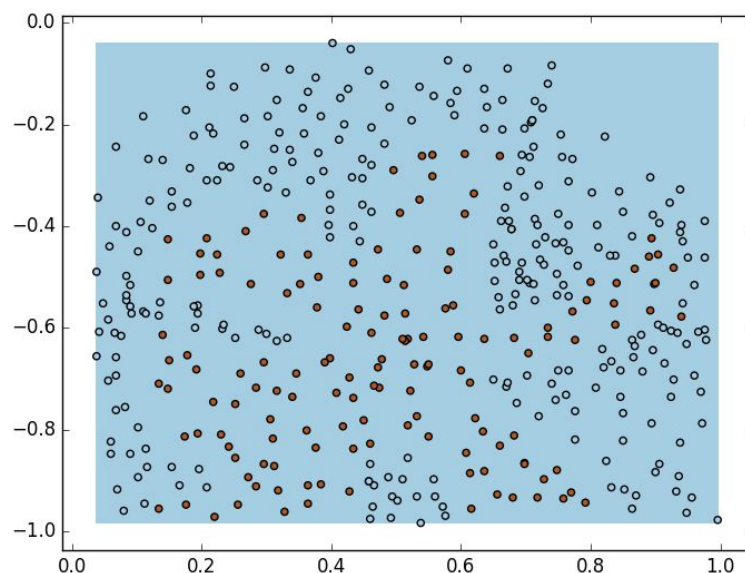


El motivo es que este es el modelo en el que el hiperplano clasificador es más “paralelo” a las nubes de patrones, además de conseguir un margen bastante grande. Me hace pensar que los patrones para los que comete un error de clasificación son casos atípicos y que la mayoría de patrones estarán próximos a alguna de las nubes por lo que el hiperplano que comete un menor errores para la mayoría de estos es aquel equidistante y “paralelo” a estas regiones.

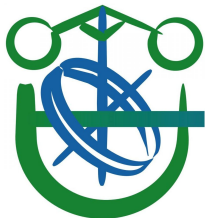
1.5. Pregunta 5.

Prueba a lanzar una SVM lineal con los valores para C que se utilizaron anteriormente. ¿Consigues algún resultado satisfactorio en el sentido de que no haya errores en el conjunto de entrenamiento? ¿Por qué?

No se consigue ningún resultado satisfactorio porque estamos usando hiperplanos de separación lineal y el segundo conjunto de datos no es linealmente separable. En otras palabras, no existe un hiperplano de separación lineal en el espacio que sea capaz de clasificar las clases para este conjunto.

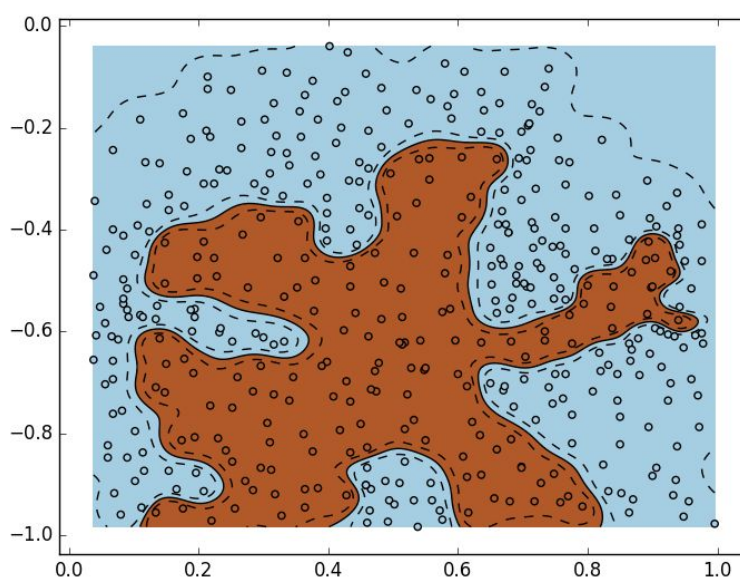


Como se puede apreciar es conjunto de datos sigue un modelo bastante más complejo que un sencillo modelo lineal. Esta es la razón de que no podamos abordarlo de esta forma.

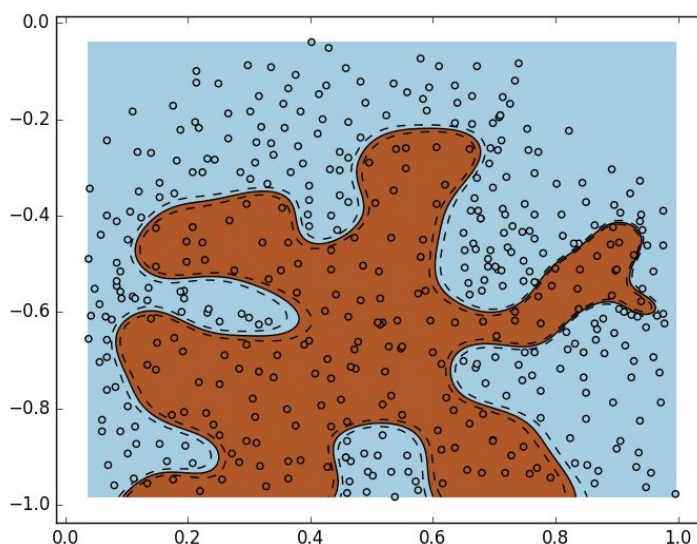


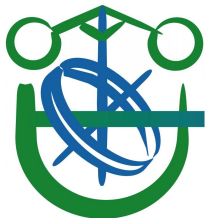
1.6. Pregunta 6.

Propón una configuración de SVM no lineal (utilizando el kernel tipo RBF o Gaussiano) que resuelva el problema. El resultado debería ser similar al de la Figura 3. ¿Qué valores has considerado para C y para γ ? Además, incluye un ejemplo de una configuración de parámetros que produzca sobre-entrenamiento y otra que produzca infra-entrenamiento.

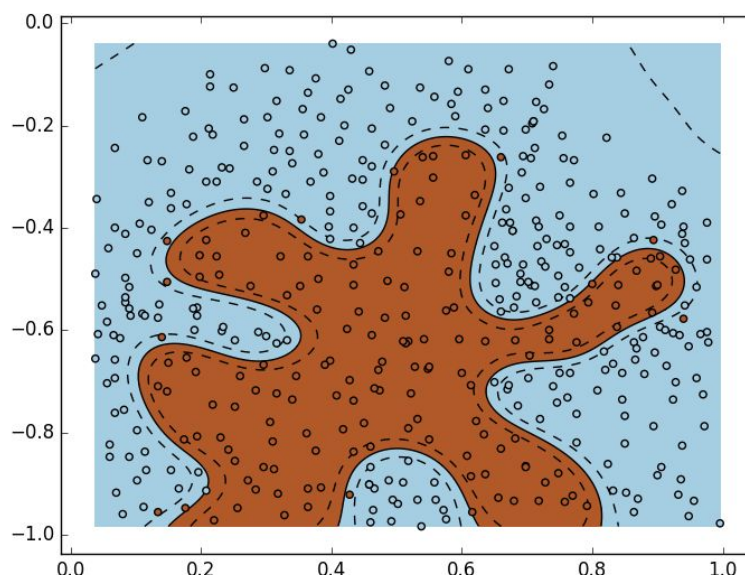


Para resolver el problema se ha utilizado kernel de tipo 'rbf' con C igual a 10 y gamma igual a 500.





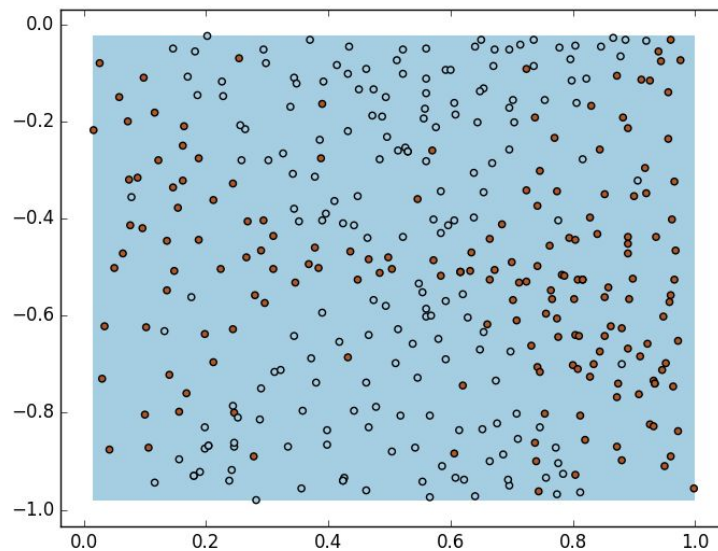
Un ejemplo que produzca sobre-entrenamiento podría ser usando un parámetro C igual a 10^6 y gamma igual a 100. Se puede apreciar como hay zonas en las que incluso ya no es posible diferenciar el margen.



Un ejemplo que produzca infra-entrenamiento podría ser usando un parámetro C igual a 1 y gamma igual a 100. Se puede apreciar como existen ya varios patrones que no han sido bien clasificados.

1.7. Pregunta 7.

En este caso, ¿es el dataset linealmente separable? A primera vista, ¿detectas puntos que presumiblemente sean outliers? ¿Por qué?

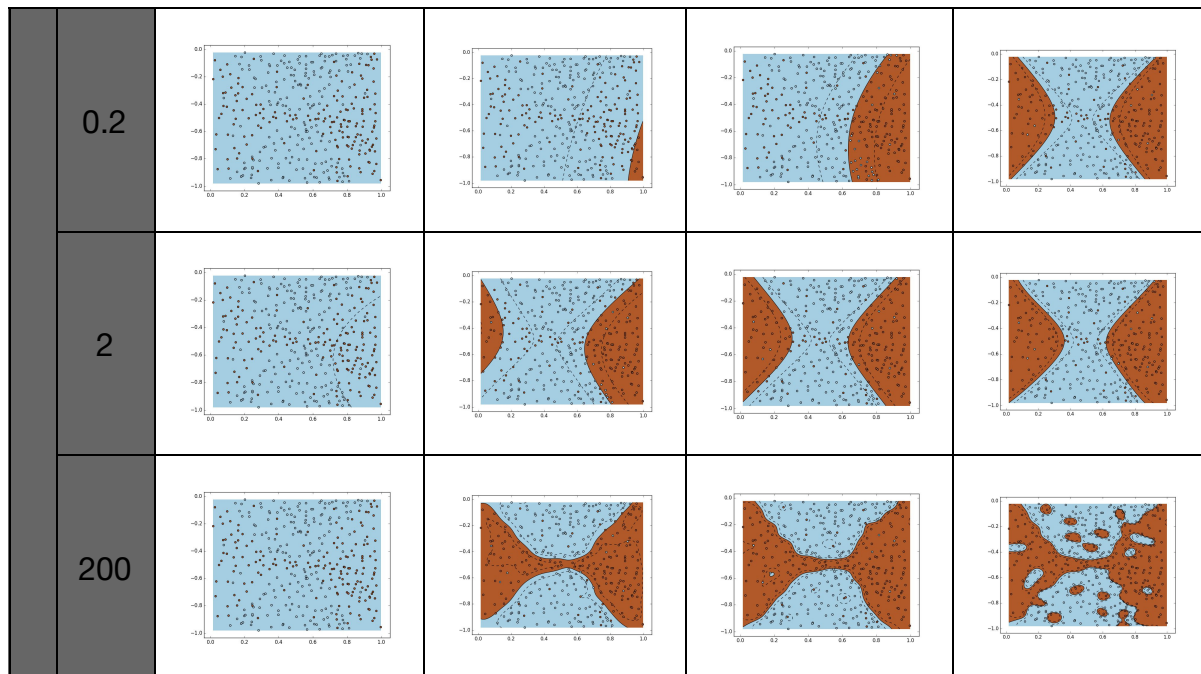


Como se puede apreciar el conjunto de datos no es linealmente separable. A primera vista, se detectan puntos que seguramente sean valores atípicos porque en las zonas donde predomina una clase también se encuentran aisladamente patrones de la otra clase.

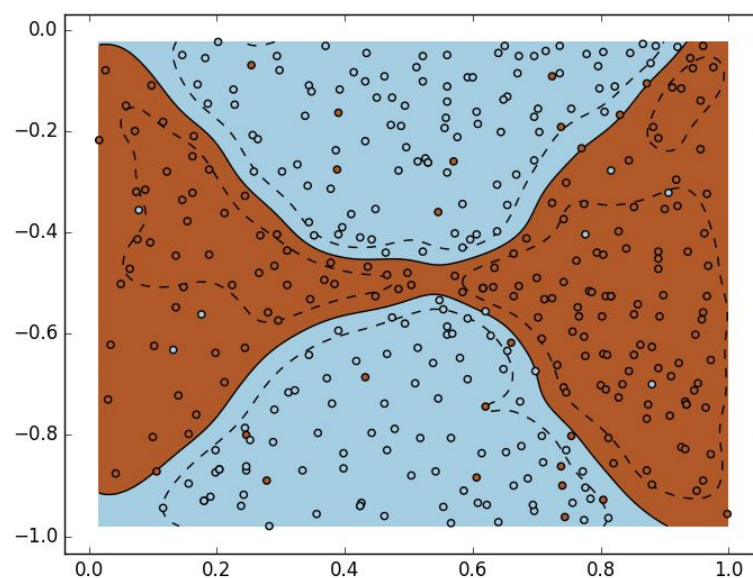
1.8. Pregunta 8.

Lanza una SVM para clasificar los datos, con objeto de obtener un resultado lo más parecido al de la Figura 5. Ajusta el ancho del kernel en el intervalo $\{0.02, 0.2, 2, 200\}$. Ajusta el parámetro de coste en el intervalo $C \in \{0.02, 0.2, 2, 200\}$. Establece el valor de los parámetros óptimos. Además, incluye un ejemplo de una configuración de parámetros que produzca sobre-entrenamiento y otra que produzca infra-entrenamiento.

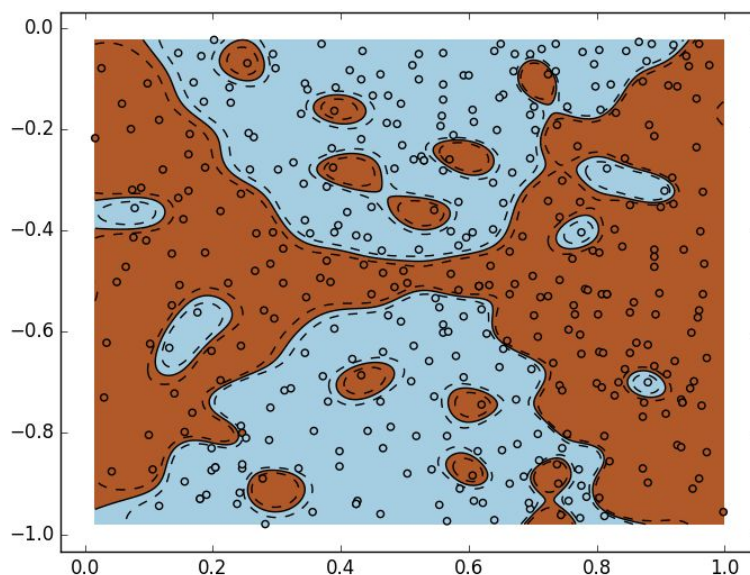
		C			
		0.02	0.2	2	200
γ	0.02				
	0.2				



Se aprecia que los parámetros óptimos para la clasificación de los datos del tercer conjunto han sido $C = 0.2$ $\gamma = 200$.



Un ejemplo de sobre-entrenamiento sería el modelo obtenido con parámetros $C = 200$ $\gamma = 200$.

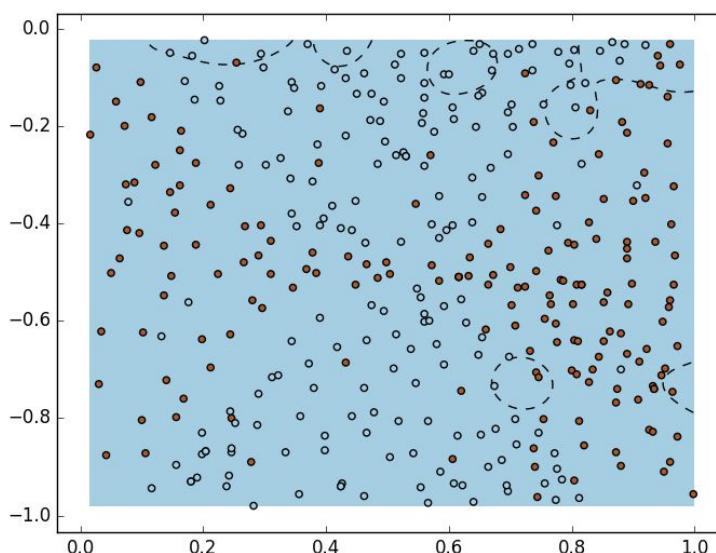


Ejemplos de infra-entrenamiento podrían ser cualquier combinación de $C \in \{0.02, 0.2, 2, 200\}$ y $\gamma \in \{0.02, 0.2, 2\}$.

1.9. Pregunta 9.

Vamos a reproducir este proceso en Python. Divide el dataset sintético `dataset3.libsvm` en dos subconjuntos de forma estratificada, con un 80% de patrones en train y un 20% de patrones en test. Realiza el proceso de entrenamiento especificado anteriormente, utilizando los valores de C y γ obtenidos en el último ejercicio. Comprueba el porcentaje de buena clasificación que se obtiene para el conjunto de test.

El porcentaje de buena clasificación que se obtiene para el conjunto de test utilizando los valores de C y γ obtenidos en el último ejercicio es de 51.35%. Al estandarizar los datos, los valores óptimos para los parámetros de clasificación han dejado de ser los especificados anteriormente.



El modelo obtenido es bastante malo para la clasificación de este conjunto de datos. Los valores óptimos para los parámetros C y γ serán diferentes a los anteriormente especificados.

Los nuevos valores óptimos para $C \in \{0.02, 0.2, 2, 200\}$ y $\gamma \in \{0.02, 0.2, 2, 200\}$ son $C = 2$ y $\gamma = 2$. Con estos valores se consigue un CCR entre 85 y 90% aproximadamente. Los datos no son exactos porque el algoritmo, ahora, tiene una parte aleatoria de selección de patrones para entrenamiento y para test.

1.10. Pregunta 10.

Amplía el código anterior para realizar el entrenamiento del ejercicio anterior sin necesidad de especificar los valores de C y γ . Compara los valores óptimos obtenidos para ambos parámetros con los que obtuviste a mano. Extiende el rango de valores a buscar si es que lo consideras necesario.

El modelo obtenido es idéntico, los valores óptimos de C y γ siguen siendo $C = 2$ y $\gamma = 2$ y se consigue el mismo CCR, de forma aproximada debido a la parte aleatoria del algoritmo.

Al extender el rango de valores, el tiempo de búsqueda de los valores óptimos para los parámetros aumenta considerablemente. En nuestro caso, hemos aumentado el rango a $C \in \{2^{-5}, 2^{-4}, \dots, 2^4, 2^5\}$ y $\gamma \in \{2^{-5}, 2^{-4}, \dots, 2^4, 2^5\}$ y los valores óptimos siguen siendo idénticos.



1.11. Pregunta 11.

¿Qué inconvenientes observas en ajustar el valor de los parámetros “a mano”, viendo el porcentaje de buena clasificación en el conjunto de test?

El principal inconveniente de ajustar el valor de los parámetros “a mano” es que es un trabajo laborioso y repetitivo. Además, hay que comparar las salidas, que tienen una parte aleatoria, lo que nos puede llevar a errores. Todo esto se puede automatizar con la búsqueda automática, que nos facilita la labor y nos protege contra posibles fallos humanos.

1.12. Pregunta 12.

En primer lugar, necesitarás convertir los ficheros tipo arff (formato de Weka) al formato de libsvm.

Los ficheros tipo arff han sido convertidos al formato de libsvm a través de Weka importando los ficheros como arff y exportando estos como libsvm.

Ahora, train_digits.arff y test_digits.arff son train_digits.libsvm y test_digits.libsvm respectivamente.

1.13. Pregunta 13.

Una vez hayas convertido los ficheros, utiliza el script que desarrollaste en el ejercicio anterior para entrenar esta base de datos. Observe el valor de CCR obtenido para el conjunto de generalización y compáralo con el obtenido en prácticas anteriores. El proceso puede demorarse bastante. Al finalizar, toma nota de los valores óptimos obtenidos para los parámetros.

Los valores óptimos de los parámetros obtenidos han sido $C = 200$ y $\gamma = 0.02$. El valor de CCR obtenido para el conjunto de generalización con estos parámetros es de 95.61%, un 0.63% superior al obtenido en prácticas anteriores (94.984%).

1.14. Pregunta 14.

Localiza donde se especifica el valor de K para la validación cruzada interna y el rango de valores que se han utilizado para los parámetros



C y γ . ¿Cómo podrías reducir el tiempo computacional necesario para realizar el experimento? Prueba a establecer $k = 3$, $k = 5$ y $k = 10$ y compara, utilizando una tabla, los tiempos computacionales obtenidos y los resultados de CCR en test.

El valor de K para la validación cruzada se especifica a través del parámetro `cv` y el rango de valores de los parámetros C y γ se especifican a través del parámetro `param_grid`, ambos en el constructor `GridSearchCV`.

```
class sklearn.model_selection. GridSearchCV (estimator, param_grid, scoring=None, fit_params=None, n_jobs=1, iid=True, refit=True, cv=None, verbose=0, pre_dispatch='2*n_jobs', error_score='raise', return_train_score=True) [source]
```

	Tiempo computacional	CCR en test	Valores óptimos para C y γ
$k = 3$	16.49s	95.61%	$C: 2$, $\gamma: 0.02$
$k = 5$	34.62s	95.61%	$C: 200$, $\gamma: 0.02$
$k = 10$	82.09s	95.61%	$C: 2$, $\gamma: 0.02$

Sólo aumenta el tiempo computacional y varían los valores óptimos.

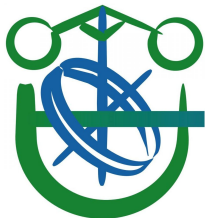
1.15. Pregunta 15.

Debes entrenar un modelo lineal de SVM con valores $C = 10^{-2}$, $C = 10^{-1}$, $C = 10^0$ y $C = 10^1$. Para ello utiliza un script similar al que usaste en la sección 2.6. Compara los resultados y establece la mejor configuración.

La mejor configuración para entrenar el modelo lineal de SVM para el conjunto de datos 'spam' es $C = 0.1$ y $\gamma = 0.01$. Se obtiene un CCR en test de 98.9% y el tiempo computacional ha sido de 246 segundos.

1.16. Pregunta 16.

Para la mejor configuración, construye la matriz de confusión y establece cuáles son los correos en los que la SVM se equivoca. Consulta las variables de entrada para los correos que no se clasifican correctamente. Ten en cuenta que para cada patrón, cuando x_i es



igual a 1 quiere decir que la palabra i-ésima del vocabulario aparece, al menos una vez, en el correo.

La mejor configuración para entrenar el modelo lineal de SVM para el conjunto de datos 'spam' es $C = 0.1$ y $\gamma = 0.01$. Se obtiene un CCR en test de 98.9% y el tiempo computacional ha sido de 246 segundos. Además, podemos interpretar los datos de una forma más individual, patrón a patrón. Para ello, identificamos cuántos, cuáles y cómo son los patrones que ha clasificado mal.

		Clase estimada	
		No	Sí
Clase real	No	684	8
	Sí	3	305

Podemos observar que han sido clasificados correctamente 989 de 1000 patrones que tenemos en el conjunto de generalización. Los patrones que se han clasificado de forma incorrecta han sido los siguientes (considerando el primer patrón como patrón número 0):

- 9: es spam pero no ha sido clasificado como spam.
- 21: no es spam pero ha sido clasificado como spam.
- 58: no es spam pero ha sido clasificado como spam.
- 73: no es spam pero ha sido clasificado como spam.
- 147: no es spam pero ha sido clasificado como spam.
- 328: no es spam pero ha sido clasificado como spam.
- 407: es spam pero no ha sido clasificado como spam.
- 526: no es spam pero ha sido clasificado como spam.
- 560: no es spam pero ha sido clasificado como spam.
- 842: no es spam pero ha sido clasificado como spam.
- 881: es spam pero no ha sido clasificado como spam.

A continuación se muestra el contenido de cada uno de estos correos en orden creciente:

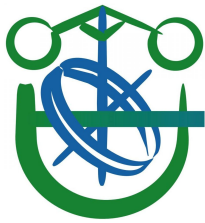


- emailaddr for group httpaddr inform irish linux list maintain subscript un user
- an and at be by call can contact email emailaddr geek heaven httpaddr if immedi is list mail messag need net not number of offic our out repres respond return sf spamassassin sponsor start talk the there thi thinkgeek to until welcom when will you your
- action address advertis all an and anti appl approv ar as at awai back bank be between bill bui busi by call can card center channel click code com congress content continu control copyright cost custom cut daili deliv digit don down driver easi edit email emailaddr emerg end energi enter enterpris error european event everi exactli expert faq fax fee feed file find firm fnumber follow for franc free from get give go good grand grant group ha half hardwar headlin here high host hp html httpaddr ibm id imag in instruct intel intern internet into intro is isp it job join juli just keep kill known land languag letter life like line linux listen mac mail make maker media messag microsoft middl more ms need net new next no number of old on onli out over password pc person plan platform plug polit price privaci product publish pull qnumber realli reg regist remot reserv right road same season secur see send septemb servic ship sign site situat softwar sort spec special sponsor staff stai standard start storag stream suffer sun sure system tablet take team tel term that the these thi ticket tm to todai told too turn uk under unit unsubscrib up updat url us visit want war we web will win window with www xp yahoo you your
- abl about abov accept access account act addit address administr advertis advis affect affili ag again against agenc agent agre air all allow also altern am among an and ani anoth ar area as associ assum at attempt author autom avail award be becom befor begin behalf below between book box brand broadcast bug busi but button by ca california can cannot capabl case cash caus center centuri chanc check choos claim clearli click co collect com commit commun compani complet



comput condit confirm connect corpor correspond countri
court dai data date dear decis declar defin deliv depend describ
direct directli director do document doe dollarnumb domain
doubl draw drive dure each educ either electron email
emailaddr employe enter enterpris entertain entri equal equip
error etc event except expens expir fair famili feder few film final
first five follow for foreign form free from full fulli game gener
govern grand grant greater ground ha had hardwar hat have
held her hi him hold home how ident if immedi improv in inc
includ independ indic individu inform instal institut instruct
integr internet is island it judg kind late later law legal like limit
link list live lo local locat log loss lost made mai mail major mani
manufactur matter meet member mention messag middl militari
must my name natur near necessari night no non not number
numberd numberth oblig occur of offer offic offici on onc onli
onlin open or order organ origin other otherwis out over oz pack
page paragraph parent part parti particip per period permiss
person place plan pleas point polici potenti prevent print privati
prize problem process product promot properti provid public
purchas purpos random reason receiv regard relat releas repli
repres request requir reserv resid respect respons result retail
return right round rule run same screen script second secur
select self send servic set sign signatur sincer site size social
softwar specif specifi sponsor standard state subject submit
such take taken tax technic telephon ten term termin texa than
thank that the their them these thi third those time tm to togeth
traffic transfer travel trip under unit up us usa valid valu ve via
visit voic wai we web week where which who whole whose will
win winner with within without won write www you your

- about accord acquir activ actual ad addit administr advertis
advic advis after again ago agre almost alon also am america
american among an and ani announc annual anoth approach ar
aren around articl artist as ask associ assum at august averag
awai back bank base be becaus been befor below better big
billion bit board both bottom brand bring bui busi but by cabl
call can capit career case cash caus charg check chief chip chri
citi civil clearli client close com commit compani complet



comput conflict connect consid construct consult consum
contact contract control copyright corpor cost could court
creativ credit cross current dai david deal debt decad declin
demand describ despit detail did differ director discount discov
discuss do doe dollar dollarnumb don doubl down dure earli
earlier earn easili econom economi edg electron els emailaddr
emerg end entertain environ equip establish estim even ever
exempl except execut exist express extens face fall far favorit
feder few file financ financi find firm first fix flow folk for form
former found friend from fund gener global go goe good govern
grant great group grow guarante gui ha had half hand have he
head health help hi high highli hold home hope hour how
httpaddr huge human if in includ industri inform initi instead
interest internet interview into invest investor involv is issu it
jame jim job john june just keep king know knowledg known
larg larger last late later latest less let like limit line littl lo loan
long look lose lost lower made mai mail make maker manag
mani march mark market master materi media men million
monei month more most mr much my name nation nearli
necessari need network new newslett no nor not note now
number numberb octob of off offer offic often old on onc onlin
oper opportun or organ other our out over own owner pacif
paid part parti partner pattern payment peopl percent perfectli
perform phone pick place plai plan pleas poor pop possibl
potenti prefer presid pretti price privat problem profit promis
prove public purchas put qualiti quarter question quickli rais re
reach realiz reason recent remain replac report repres result
retail return reveal right risk robert role sai said sale san save
search second see seed seek seem seen self sell senior sent
septemb serv servic shape share should show sinc sit site six
skeptic skill small so sold some sometim special speed spend
spot stai start step still stock structur style success suffer suit
system take taken talk team technolog texa than that the thei
their them then there these thi think third those though thought
three through ticket time to toll took top touch travel tri trip
troubl trust turn two under unless unlik until up upgrad us usual
ve veri visit vote wa want war washington watch we web week



well were west what when where which who whose why will
william with won work worth would write ye year york you your

- address advertis all alwai and announc been below bui close
cnet com comparison custom dai download each emailaddr faq
for format go have help in item juli let mail manag million more
my name nbsp new number of on our own price product provid
purchas put review sale save servic shop so subscript tech
technolog that the there thousand to todai unsubscrib up valu
ve visit you your
- about abov account address africa after also although am
amount and ani ar as assist avoid awai be befor box busi by call
can caus citi com command compani confidenti contact
content countri current date dear death democrat deposit did
direct doe due dure effect emailaddr expect expens famili father
fax five for former friend from fund got govern have he head
help hi httpaddr idea in invest is it john keep kill kingdom know
knowledg late life line list long mail map me militari million
monei mr my need not note number of on open or osdn out
over person plan pleas present presid privat profit prompt reach
remain remot republ requir respect respons sai secur seek self
should son spamassassin suggest take talk tel term that the
there these thi threat through to train transact unit urgent us
war we while who wife will with www you your
- about ad advertis after am an and as at been below bodi but by
click contain dai direct doc email emailaddr except extra fals
final geek get ha have heaven here httpaddr is it list look mail
mark net next no number off on other posit razor see sender sf
signatur so spam sponsor such talk the thi thinkgeek to turn us
user welcom where which with word
- as cours ham happi have it late never non number on pick
receiv spam to too up
- about account acquir across ad address advertis affect all allow
almost american an and ani anoth anti approach ar area as ask



asset at attack avail avoid be becaus been believ best better bill
but by california call campaign can cent ceo chanc channel
chief choic choos come commerci commun compani congress
constitut contact content could countri coverag dai decid defin
delet differ direct directli do each educ effort ensur entir equal
even everi expens experi face fact fair featur feder few field firm
first for former free from front fund futur good great group ha
had have he hi high highli him howev httpaddr if in individu
inform instead internet is it just know lack larg law least less let
level light like link list lo mai mail make manag mani market
media messag method might mike million monei month more
most must nation neg no not now number of off offer offic on
onli onlin open opt or other our out paid particularli peopl per
percent perfect perform person phone plai point polit potenti
practic prepar presid press problem provid public qualiti radio
reach read receiv recipi refin relev remain repli repres respons
seek send sender sens sent she shop should simpl social
societi softwar some spam spamassassin specif spot staff
standard state such support take taken target technolog than
that the thei their them thi those through time to todai toward
tradit tri tv unsolicit unsubscrib up us veri vote wa wai walk
want war we what when while who whose will with without won
world would written year

- about abus all also america american amount an and ani anoth
ar argument as ask attempt attent author awai babi base basi
be been befor behavior believ best better black both bui by car
carri caus center children choic choos citizen close collect com
commentari commit commun confus contain cool copi could
coupl cours current daili date did direct do drive drug due
dynam educ email emailaddr even ever exist expens explain ey
fact femal for form from further gener go govern ha have heard
hi higher home hous how howev identifi if imag import in
individu invest is it know known lack larg learn less light like live
look mai make maker male mani mayb me men might miss
more mother move my name natur neg new no not noth notic
now number of often on onli or over own pai paid part particular
pass peopl perhap place pleas poor popul prefer project
purchas rais rather refer repres request reserv respect result



said save self sell shop side singl sit societi soon sort stori
street suffer support tax than that the thei their them themselv
there these thi through to todai told true truth util visit wa warn
we wed welcom well what when while who with women wonder
world would www ye you young your

1.17. Pregunta 17.

Compara los resultados obtenidos con los resultados utilizando una red RBF. Para ello, haz uso del programa desarrollado en la práctica anterior. Utiliza sólo una semilla (la que mejor resultado obtenga).

La configuración utilizada para entrenar el modelo usando una red RBF para el conjunto de datos 'spam' ha sido número de RBFs = 100 y $\mu = 10$ ($C = 0.1$). Se obtiene un CCR en test de 87.20%, un 11.7% menos, y el tiempo computacional ha sido de 5.37 segundos, 240.63 segundos menos. Además, podemos identificar cuántos son los patrones que ha clasificado mal.

		Clase estimada	
		No	Sí
Clase real	No	690	2
	Sí	126	182

Podemos observar que han sido clasificados correctamente 872 de 1000 patrones que tenemos en el conjunto de generalización, 117 patrones bien clasificados menos que en el ejercicio anterior. Se puede ver que clasifica mejor los que no son spam (6 más, bien y 6 menos, mal) pero tiene mayor confusión con los que sí lo son (123 más, mal y 123 menos, bien).



2. Referencias bibliográficas.

- <https://en.wikipedia.org/>
- <http://scikit-learn.org/>
- <https://www.scipy.org/>
- <http://archive.ics.uci.edu/>
- <http://spamassassin.apache.org/publiccorpus/>
- <http://www.lab.inf.uc3m.es/>
- <http://www.csie.ntu.edu.tw/>



-
- <http://stackoverflow.com/questions/27264426/arff-files-with-scikit-learn>
 - <https://github.com/dat/svm-tools>
 - Auer, Peter; Harald Burgsteiner; Wolfgang Maass (2008). «A learning rule for very simple universal approximators consisting of a single layer of perceptrons». Neural Networks 21 (5): 786-795.
 - Roman M. Balabin, Ravilya Z. Safieva, and Ekaterina I. Lomakina (2007). «Comparison of linear and nonlinear calibration models based on near infrared (NIR) spectroscopy data for gasoline properties prediction». Chemometr Intell Lab 88 (2): 183-188.
 - Meshfree Approximation Methods with MATLAB, Gregory E. Fasshauser. Illinois Institute of Technology, USA.
 - Redes de Neuronas de Base Radial, Redes de Neuronas Artificiales, UC3M, RAI 2012.
 - Alba Castro, José Luis. «Máquinas de Vectores Soporte (SVM)».