Nota: 10 / 10

Reconocimiento Estadístico de Patrones
Tarea 6

Randy Osbaldo Ibarra Cayo

B. Preguntas Cortas **Nota: 4 / 4**

Pregunta *a*

Vimos en la clase que si $x \in \mathbb{R}^2$, tipicamente el número de vectores de soporte para datos linealmente separables es 2 o 3. ¿ Qué se puede decir si $x \in \mathbb{R}^3$?

Solución

Suponemos datos linealmente separables en $\mathbb{R}^3$, entonces existe un plano que separa las dos clses.

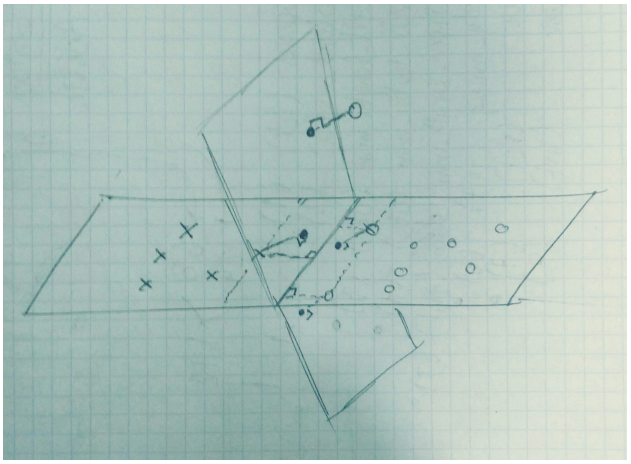
- Si nuestros datos están en una recta entonces al tomar el punto medio de los **dos** datos de distintas clases más cercanos, y cualquier vector no paralelo a la recta de los datos define un plano (ortogonal a este vector) que separa las clases. En este caso se tiene sólo dos vectores de soporte. En particular, podemos tomar el plano ortogonal a la recta de los datos que pasa por el punto medio de los dos datos de distintas clases más cercanos y así maximizar el margen.

- Si nuestros datos están en un plano, entonces cualquier plano no paralelo al plano de los datos define una recta sobre el plano de los datos. Tomamos un punto medio de **dos** datos de distintas clases más cercanos (Pueden exister varias parejas a la misma distancia) y consideremos el plano ortogonal a la recta une estas dos observaciones. Puede ser que la recta inducida por este plano no separe las clases, pero esto nos indica en que direccion debemos rotar (el plano que induce) la recta para separar las clases.

  Una vez que encontramos una recta que separa las clases puede existir un dato que está más cerca de esta recta que los dos datos de distintas clases más cercanos. Entonces podemos rotar la recta de tal forma que la distancia entre esta observacion y la recta crece, mientras que la distancia entre los dos datos más cercanos y esta recta disminuye. Dejamos de rotar la recta cuando las distancias entre estos tres datos y la recta sea la misma. Por último tomamos el plano ortogonal al plano de los datos cuya intersección con este último es la recta que construimos. En este caso se tienen **tres** vectores de soporte. Notemos que 4 vectores de soporte en el mismo plano implica que se tienen dos parejas de puntos que definen dos rectas paralelas lo cuál, se comentó en clase, es un evento de probabilidad cero.

Por último, notemos que si este plano está en un espacio Entonces todos los planos que contienen a la recta obtenida que separa en el plano las clases, separa los datos en el espacio 3D. Luego, consideremos un punto de alguna de las dos clases fuera del plano de los datos. Si la distancia de este dato al plano es menor que el margen inducido por la recta en el plano y los datos, entonces es necesario rotar (un plano no paralelo al plano de los datos) de tal forma que aumentemos la distancia entre el dato del espacio y el dato, y disminuyamos la distancia de los vectores de soporte al plano que estamos considerando. De esta forma maximizamos el margen en el espacio 3D.

Concluimos que para el espacio 3D, se pueden tener desde 2 hasta 4 vectores de soporte. Ok



Pregunta b

Supongamos que tenemos los datos bidimensionales de clasificación binaria (+, -) de la Figura 1(a). Como el área de decisión es $x_1^2 + x_2^2 - 1 = 0$ es suficiente hacer una transformación polinomial de grado 2, i.e. es suficiente trabajar con un kernel polinomial de grado 2. De qué grado mínimo debe ser el kernel polinomial para los datos de la Figura 1(b) y de la Figura 1(c)?

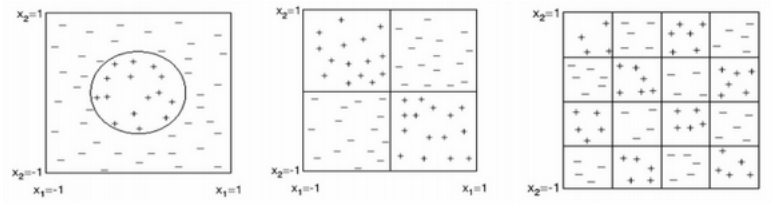


Figura 1

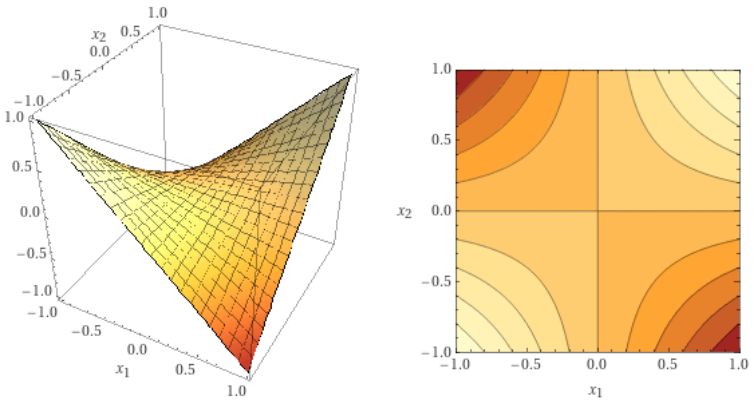
Solución

En los tres casos es facil ver que los datos no son linealmente separables, por lo que el kernel debe ser de al menos grado dos.

- Para la figura (b) buscamos una función cuyas curvas de nivel con valor 0 sean los ejes x_1 y x_2. Notemos que la clase + cumple que $sgn(x_1) = sgn(x_2)$, mientras que la clase − cumple que $sgn(x_1) \neq sgn(2)$. Por otro lado, el producto de dos numeros con mismo signo nos da como resultado un número con signo positivo, mientras que el producto de dos numeros con signodistinto nos deja un número negativo. Por lo anterior, basta observar la función

  $$x_1 \cdot x_2$$ Ok

  la cual cumple que los datos de la clase − tendrán un valor positivo, mientras que los datos de la clase + tendrán un valor negativo. Dado que la función $x_1 x_2$ es un polinomio en dos variables de grado dos, podemos concluir que el grado mínimo del kernel polinomial es 2.

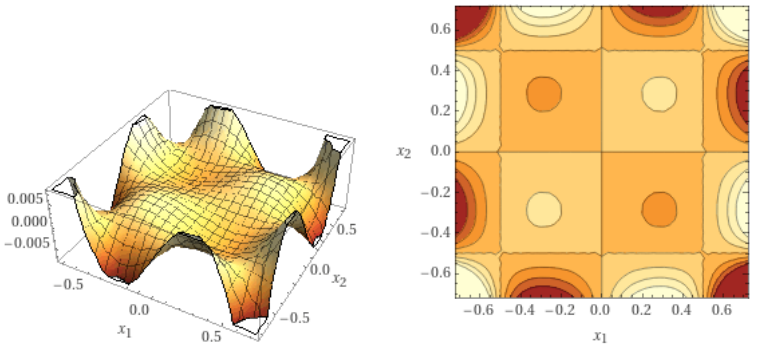


- Para la figura (b) notamos que deben existir curvas de nivel de valor 0 cuando $x_1 = 0$, $x_2 = 0$, $x_1 - 0.5 = 0$, $x_2 - 0.5 = 0$, $x_1 + 0.5 = 0$, $x_2 + 0.5 = 0$. Consideremos una funcion polinómica en dos variables. Si restringimos x_2 a un valor constante distinto de 0, 0.5 y -0.5, debemos de tener una función polinómica que tenga al menos 3 raíces, esto es, de grado al menos 3. Análogamente, si restringimos x_1 a un valor constante distinto de 0, 0.5 y -0.5, debemos de tener una función polinómica que tenga al menos 3 raíces, esto es, de grado al menos 3. Entonces la funcion polinómica debe ser de grado al menos 6, de lo contrario podemos elegir restringir x_1 o x_2 a un valor de tal forma que la funcion restringida sea un polinomio de grado menor a 3.

  Por otro lado, notemos que la función

  $$x_1 \cdot x_2 \cdot (x_1 - 0.5) \cdot (x_2 - 0.5) \cdot (x_1 + 0.5) \cdot (x_2 + 0.5)$$ Ok

  tiene como curvas de nivel de valor 0 las rectas que separan las clases y también podemos ver que esta función separa las clases en las 16 regiones que aparecen en la figura (b). Por lo tanto, el mínimo grado del kernel polinomial es 6.



Bien!

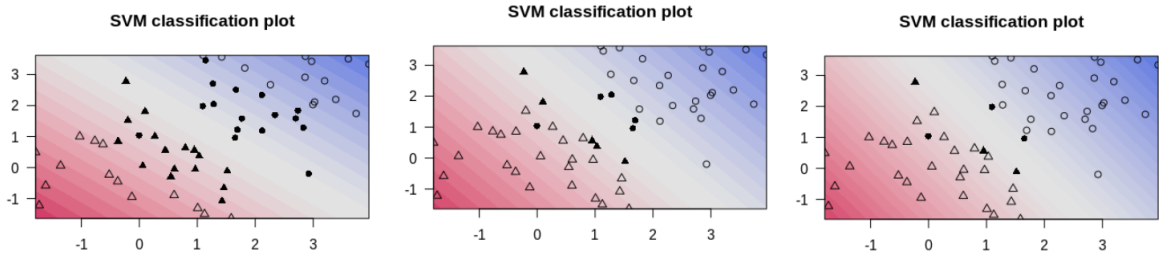
Pregunta c

Juega con el código de svm2.r, en particular lo que tiene que ver con el efecto del parámetro `cost` de la función ksvm en el resultado. Argumenta empiricamente lo que representa el parámetro `cost`: λ o γ (minuto 24 del video de la clase de 19/4).

Solución

Al variar el parámetro `cost` podemos notar que entre más grande, el número de vectores de soporte es menor, mientras que si el parámetro `cost` es menor, entonces el número de vectores de soporte es mayor. Por lo que el parámetro `cost` influye en el tamaño del margen: Si es costo es grande entonces nos conviene usar un margen pequeño, mientras que si el costo es pequeño se nos permite un margen mayor.

En las siguientes figuras se encuentran los resultados desde un costo de 0.05 (izquierda) hasta un costo de 10 (derecha).



Ok

Pregunta d

Vimos que minimizar $E(1-Yg(X))_+$ sobre g conduce al clasificador Bayesiano óptimo $\hat{y}(x) = sgn(g(x))$. ¿Se obtiene lo mismo al minimizar $E(1-Yg(X))^2$?

Solución

Dada una muestra $\{(x_i, y_i)\}_{i=1}^n$ independientes, se tiene que

$$\frac{1}{n}\sum_{i=1}^{n}\left[1-y_i g(x_i)\right]^2 \to E\left[1-Yg(X)\right]^2$$

cuando $n \to \infty$ por la Ley de los Grandes Números. Si se quiere minimizar $E\left[1-Yg(X)\right]^2$ sobre g, basta encontrar para cada x el valor $g(x)$ que minimiza

$$E_{Y|X=x}\left[1-Yg(x)\right]^2 = \left[1-g(x)\right]^2 P(Y=1|X=x) + \left[1+g(x)\right]^2 P(Y=-1|X=x)$$

Notemos que si $g(x) > 1$, entonces $2P(Y=-1|X=x) < \left[1+g(x)\right]^2 P(Y=-1|X=x)$, por lo que

$$E_{Y|X=x}\left[1-Y\right]^2 = 2P(Y=-1|X=x) < \left[1-g(x)\right]^2 P(Y=1|X=x) + \left[1+g(x)\right]^2 P(Y=-1|X=x)$$

Por lo que basta limitarnos a valores $g(x) \leq 1$ pues buscamos minimizar. Por otro lado si $g(x) < 1$, entonces $2P(Y=1|X=x) < \left[1-g(x)\right]^2 P(Y=1|X=x)$, de modo que

$$E_{Y|X=x}\left[1+Y\right]^2 = 2P(Y=1|X=x) < \left[1+g(x)\right]^2 P(Y=1|X=x) + \left[1+g(x)\right]^2 P(Y=-1|X=x)$$

Por lo que basta limitarnos a valores $-1 \leq g(x)$ pues buscamos minimizar.

Por lo anterior, para cada x buscamos valores $g(x)$ con $-1 \leq g(x) \leq 1$ que minimicen

$$\begin{aligned}
E_{Y|X=x}\left[1-Yg(x)\right]^2 &= \left[1-g(x)\right]^2 P(Y=1|X=x) + \left[1+g(x)\right]^2 P(Y=-1|X=x) \\
&= \left[1-2g(x)+g(x)^2\right] P(Y=1|X=x) + \left[1+2g(x)+g(x)^2\right] P(Y=-1|X=x) \\
&= 1+g(x)\left\{\left[-2+g(x)\right] P(Y=1|X=x) + \left[2+g(x)\right] P(Y=-1|X=x)\right\}
\end{aligned}$$

y la expresión anterior equivale a

$$1+g(x)\left\{g(x)\left[P(Y=1|X=x)+P(Y=-1|X=x)\right] + 2\left[P(Y=-1|X=x)-P(Y=1|X=x)\right]\right\}$$

tomando en cuenta que $P(Y=1|X=x)+P(Y=-1|X=x)=1$ y desarrollando un poco se tiene que lo anterior equivale a

$$1+g(x)^2+2g(x)\left[P(Y=-1|X=x)-P(Y=1|X=x)\right]$$

Denotamos por $d = \left[P(Y=-1|X=x)-P(Y=1|X=x)\right]$. Notemos que queremos minimizar la función

$$1+y^2+2dy$$

la cuál tiene un único punto crítico (mínimo debido a que el término cuadrático es positivo) en $y=-d$, por lo que para minimizar la función debemos asignar

$$g(x) = -\left[P(Y=-1|X=x)-P(Y=1|X=x)\right]$$

Por último, notemos que si $P(Y=-1|X=x)-P(Y=1|X=x) < 0$, entonces $g(x) > 0$ por lo que la clasificación es 1, mientras que si $P(Y=-1|X=x)-P(Y=1|X=x) > 0$ entonces $g(x) < 0$ por lo que la clasificación es -1, es decir asignamos x a la categoría más probable, por lo que minimizar $E\left[1-Yg(X)\right]^2$ sobre g condude al clasificador bayesiano óptimo.

Ok

B. Análisis de Datos

Ejercicio 1 **Nota: 2 / 2**

En este ejercicio trabajamos con datos de tres familias de pinguinos. Información general de dónde salieron los datos:

`https://docs.google.com/presentation/d/1DFJLXYRJ2kWw6AFkJu7MclFPr8zkqD-PHh5bs4xcr3Y`

Esta vez hacemos las cosas al revés; proporcionamos primero algunas exploraciones de los datos:

`https://education.rstudio.com/blog/2020/07/palmerpenguins-cran/`
`https://allisonhorst.github.io/palmerpenguins/articles/intro.html`

Una presentación muy interesante usando caminatas (tours) es:

`https://www.dicook.org/files/visec2020/slides_tourr#1`

Nos limitamos a los datos de las familias Chinstrap y Gentoo.

- Arma un notebook en Kaggle con un análisis exploratorio de los datos, incluye algunos experimentos con los tours.
- Busca algunas Máquinas de Soporte Vectorial adecuadas para estos datos (2 clases). Reporta y discute su desempeño. Compara su desempeño con un clasificador $k - NN$ y LDA.

Nota: 4 / 4

Descripción