Tarea 7 Reconocimiento de Patrones

Yamil Ernesto Morfa Avalos

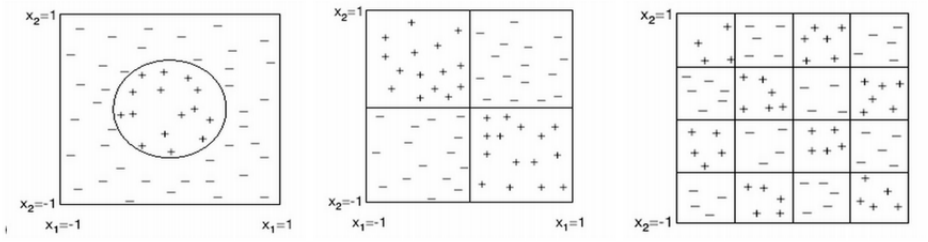
25 de abril de 2021

# B. Preguntas cortas

**a)** Vimos en la clase que si $x \in \mathbb{R}^2$, típicamente el número de vectores de soporte para datos linealmente separables es 2 o 3. Esto se debe a que dado que la frontera en una recta y una vez obtenida esta el margen se puede calcular fácilmente con la pendiente de la frontera y un punto en cada limite de este; se incluye un 3er vector de soporte dado que es posible que en uno de los limites encuentres puntos colineales, sin embargo la probabilidad de que esto suceda en ambos lados es casi nula. Análogamente para $x \in \mathbb{R}^3$ una vez obtenido el plano que separa los datos necesitamos para construir los planos que sirven para delimitar el margen necesitamos un vector normal a este y un punto que viva en el plano. Entonces en este caso sería necesario 2, 3 o 4 vectores de soporte para datos linealmente separables.

**b)** Suponemos que tenemos los datos bidimensionales de clasificación binaria $(+, -)$ de la siguiente figura. Como el área de decisión en la figura $1(a)$ es $x^2 + y^2 - 1 = 0$, es suficiente hacer una transformación polinomial de grado 2, i.e. es suficiente trabajar con un kernel polinomial de grado 2.

Figura 1:




(a) (b) (c)

Figura 1

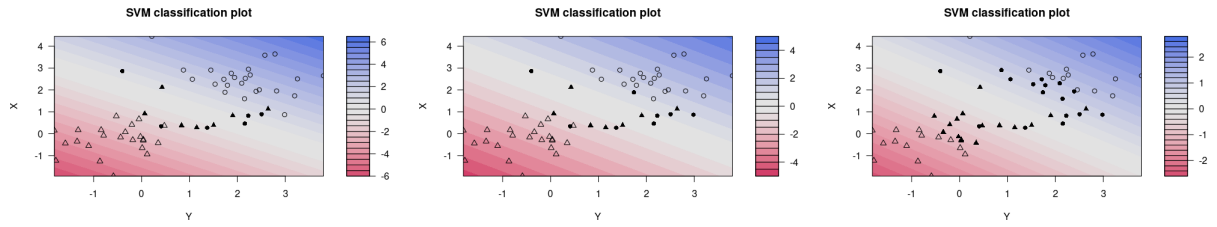
En el caso $1\,(b)$ las fronteras están dadas por las rectas $x = 0$ y $y = 0$ esto es $xy = 0$, por lo tanto necesitamos una transformación de grado 2. Análogamente para el caso $1\,(c)$ la frontera la podemos definir por $xy\left(x - \frac{1}{2}\right)\left(y - \frac{1}{2}\right)\left(x + \frac{1}{2}\right)\left(y + \frac{1}{2}\right) = 0$, luego es necesario una transformación de grado 6.

**c)** Tenemos que $ksvm$ se reduce a:

$$\min_{\beta,\beta_0} \|\beta\|^2 + \gamma \sum_i \varepsilon_i \quad S.A\ g(x_i)\, y_i \geq 1 - \varepsilon_i; \quad \varepsilon_i \geq 0\, \forall i$$

con $g(x) = sign\left(x^T\beta + \beta_0\right)$ y $\varepsilon_i$ la distancia mínima que hay que trasladar $x_i$ para que esté al lado correcto de la frontera de decisión y al menos a una distancia de $\frac{1}{\|\beta\|}$ y $\gamma$ es la penalización que se le da a cada $\varepsilon_i$ en la función de costo. Veamos como se refleja la variación de este parámetro.

Figura 2:




En la figura 2 se muestra para los datos del código $svm2.R$ las salidas para valores de $\gamma \in \{10,\ 2,\ 0{,}05\}$ respectivamente. Como se observa hay una inversamente proporcional al numero de vectores de soporte, es decir, a medida que decrece el valor de la penalización, aumenta el numero de vectores de soporte. Esto se debe a que para valores grandes de $\gamma$, para minimizar $\|\beta\|^2 + \gamma \sum_i \varepsilon_i$ se le debe dar mayor importancia al termino $\sum_i \varepsilon_i$ que multiplica a $\gamma$ y de esta manera se busca que las distancia mínima que hay que trasladar $x_i$ para que esté al lado correcto de la frontera de decisión sean menores lo que influye en la cantidad de vectores de soporte a considerar. Análogamente al $\gamma$ tener valores pequeños se le da más importancia a minimizar el termino $\|\beta\|^2$ o lo que es lo mismo maximizar la distancia $C^2 = \frac{1}{\|\beta\|^2}$ y por tanto maximizar el margen. Esto implica que se obtengan mayor numero de vectores de soporte.

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

Sea

$$E\left(1 - Yg(X)\right)_+ = \left(1 - g(x)\right)_+ P\left(Y = 1|X = x\right) + \left(1 + g(x)\right)_+ P\left(Y = -1|X = x\right)$$

Por lo que:

$$E(1 - Yg(X))_+^2 =$$
$$\left\{1 - g(x)\right\}_+^2 P^2\left(Y = 1|X = x\right) + \left\{1 + g(x)\right\}_+^2 P^2\left(Y = -1|X = x\right) + 2\left\{1 - g(x)\right\}_+ \left\{1 + g(x)\right\}_+ P\left(Y = -1|X = x\right) P\left(Y = 1|X = x\right) = *$$

Supongamos: $-1 \leq g(x) \leq 1$.

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

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

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

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

Luego sea $P\left(Y=1|X=x\right)-P\left(Y=-1|X=x\right)=\varepsilon$:

$$*=1+g^2\left(x\right)\varepsilon^2+2g\left(x\right)\varepsilon=\left(g\left(x\right)\varepsilon+1\right)^2$$

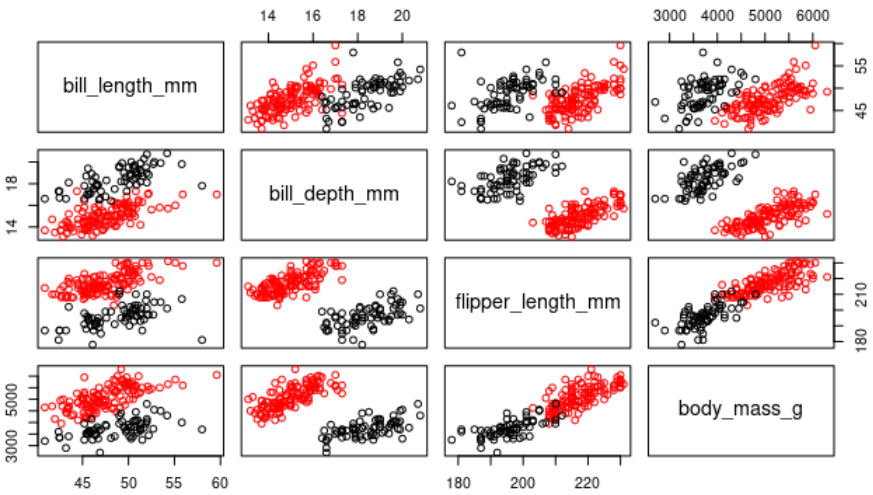
Veamos que esto es un problema cuadrático y el mínimo se alcanza en 0, entonces si suponemos que $\varepsilon<0\Longrightarrow P\left(Y=1|X=x\right)<P\left(Y=-1|X=x\right)$ tenemos que necesariamente para alcanzar el mínimo $g\left(x\right)=-1$, para otro caso $\varepsilon>0\Longrightarrow P\left(Y=1|X=x\right)>P\left(Y=-1|X=x\right)$ entonces $g\left(x\right)=1$. Luego esto es $\hat{y}\left(x\right)=sgn\left[g\left(x\right)\right]$

# C. Análisis de datos

**1.** En este ejercicio trabajamos con datos de tres familias de pingüinos. Cada observación consiste de un vector con las variables: "island", "bill_length_mm", "bill_depth_mm", "flipper_length_mm", "body_mass_g", "sex", "year" y su respectiva etiqueta en "species". Para este problema nos centraremos en las especies "Gentoo" y "Chinstrap" y en las variables: "bill_length_mm", "bill_depth_mm", "flipper_length_mm", "body_mass_g". Obteniendo así una matriz de datos de 191 observaciones de 4 variables.

**a)** Primeramente realizaremos un análisis exploratorio de los datos.

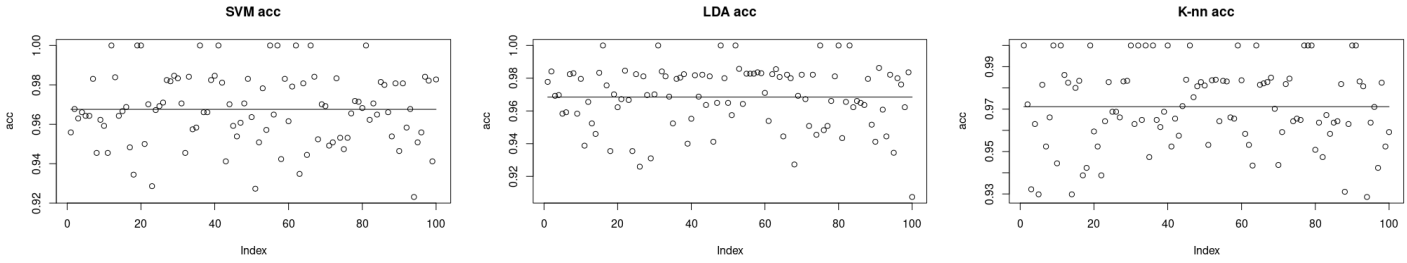
Figura 3:




En la figura 3 se muestra las relaciones entre las variables dos a dos. Podemos ver fácilmente que estas son, en su mayoría , linealmente separables.

**b)** Dado la suposición del inciso anterior apliquemos *ksvm* con kernel polinomial de grado 1. Se dividió los datos en conjuntos de entrenamiento y prueba *train*, *test* con proporción 70 % y 30 % respectivamente. Se comparó para varios juegos de datos distintos los algoritmos de $SVM$, $LDA$ y $K-nn$ y los resultados obtenidos se muestran el la figura 4 con valores de *acc* promedios de 0,9675, 0,9684 y 0,9711 respectivamente.

Figura 4:




Como podemos ver $SVM$ tuvo el peor rendimiento, esto quizás se deba a la poca cantidad de datos para entrenar.