

Tarea 05 - Reconocimiento de patrones
Giovanni Gamaliel López Padilla

Problema 01

Explora en <https://colab.research.google.com/drive/1pwCqLdvxeqChzDG3MoFIgR7m6lsInKs5?usp=sharing> el efecto de cambiar los parámetros en una SVM y convéncete que es de acuerdo a (congruente con) el funcional de costo de una SVM.

Experimento 1

¿Cuál es el efecto de cambiar el parámetro λ/γ (cost)?

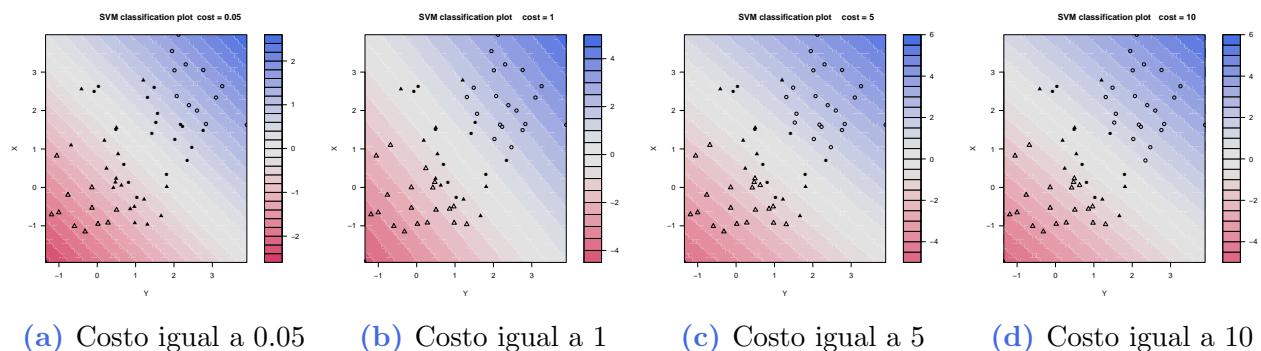


Figura 1: Diferentes valores de λ/γ para el kernel lineal.

En la figura 1 se observa que conforme se aumenta el parámetro λ/γ el número de vectores de soporte disminuye.

Experimento 2

¿Cuál es el efecto de aumentar el grado de polinomio?

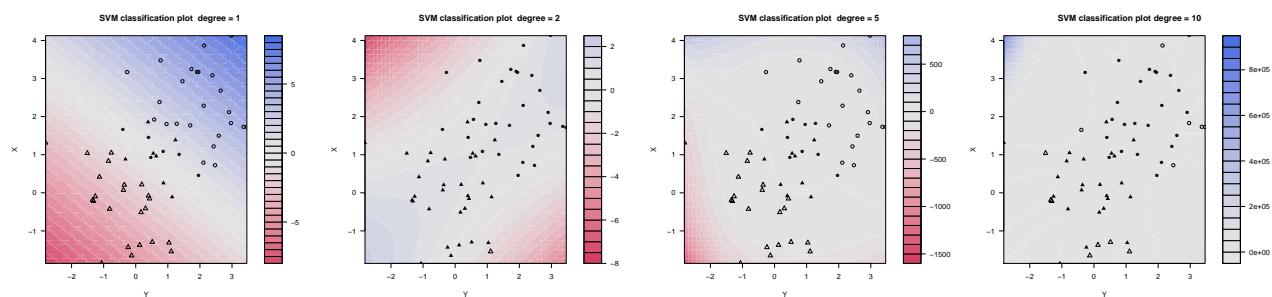


Figura 2: Diferentes grados para el kernel polinomial.

En la figura 2 se observa que conforme se aumenta el grado del kernel polinomial, la región de clasificación para cada conjunto de datos se disuelve llegando a que todos los datos conforman un mismo conjunto.

Experimento 3

¿Cuál es el efecto de cambiar el parámetro σ de kernel de base radial?

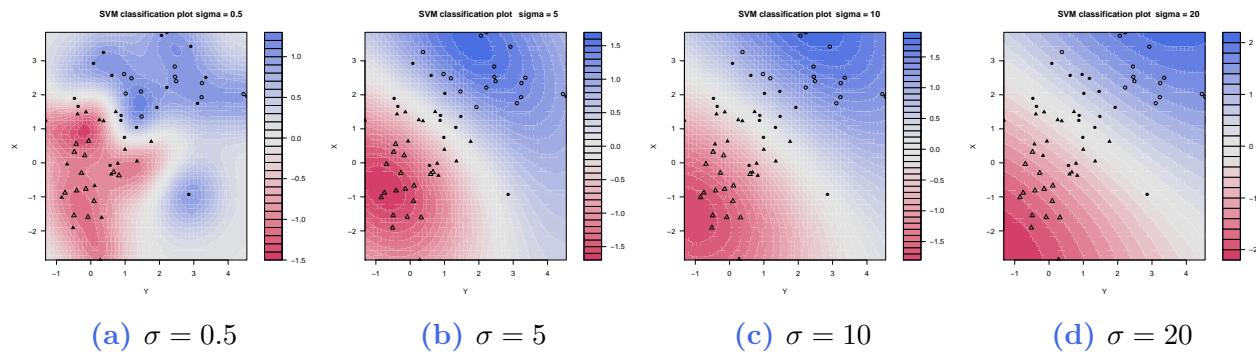


Figura 3: Diferentes valores para el parámetro σ para el kernel radial.

En la figura 3 se observa que conforme se aumenta el valor de σ , el modelo de SVM con kernel radial se aproxima a tener un kernel lineal.

Experimento 4

¿Cuál es el efecto de cambiar el parámetro σ del kernel de base radial?

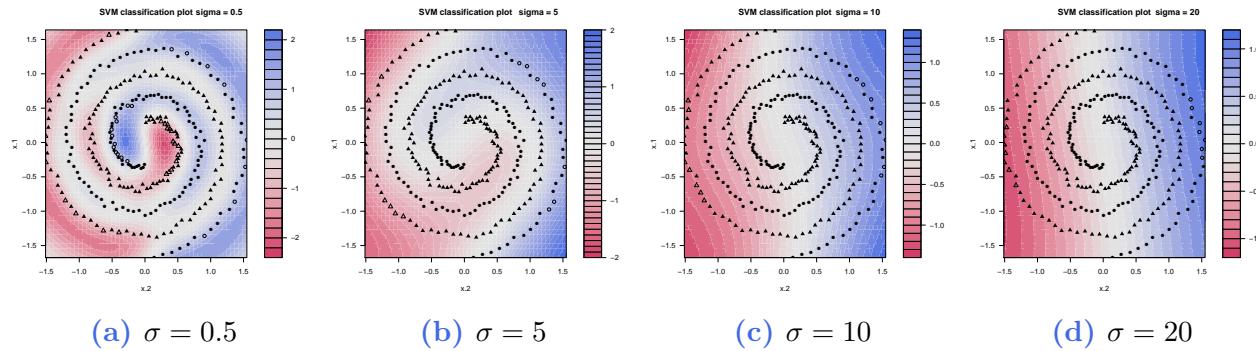


Figura 4: Diferentes valores para el parámetro σ para el kernel radial.

En la figura 4 se obtienen los resultados para los diferentes valores de σ con una base de datos dada. En esta se observa el mismo resultado que la figura 3. Conforme se aumenta el valor de σ , el kernel radial se aproxima a un kernel lineal.

Problema 2

Vimos que minimizar $E(1 - Y(g(X))_+)$ sobre g conduce al clasificador óptimo $\hat{y}(x) = \text{sgn}(g(x))$. Usando el mismo camino, muestra que se obtiene el mismo resultado para $E(\exp(-Yg(X)))$

El problema se encuentra planteado en la ecuación 1.

$$\min_{g(x)} E(e^{-yg(x)}) \quad (1)$$

Por el teorema de números grandes, el problema de la ecuación 1 se puede resolver encontrando el valor de $g(x)$ para cada x dada. Esto puede ser visto en la ecuación 2.

$$\min_{g(x)} E_{Y|X=x}(e^{-yg(x)}) \quad (2)$$

Realizando el calculo explicito de la ecuación 2 se obtiene lo siguiente:

$$E_{Y|X=x}(\exp(-yg(x))) = e^{-g(x)}P(Y=1|X=x) + e^{g(x)}P(Y=-1|X=x)$$

por lo tanto:

$$\begin{aligned} \min_{g(x)} E_{Y|X=x} &= \frac{\partial}{\partial g(x)} (e^{-g(x)}P(Y=1|X=x) + e^{g(x)}P(Y=-1|X=x)) \\ &= -e^{-g(x)}P(Y=1|X=x) + e^{g(x)}P(Y=-1|X=x) \end{aligned}$$

Encontrando el valor crítico del resultado anterior se obtiene que $g(x)$ es:

$$\begin{aligned} -e^{-g(x)}P(Y=1|X=x) + e^{g(x)}P(Y=-1|X=x) &= 0 \\ -P(Y=1|X=x) + e^{2g(x)}P(Y=-1|X=x) &= 0 \\ e^{2g(x)}P(Y=-1|X=x) &= P(Y=1|X=x) \\ e^{2g(x)} &= \frac{P(Y=1|X=x)}{P(Y=-1|X=x)} \\ 2g(x) &= \log\left(\frac{P(Y=1|X=x)}{P(Y=-1|X=x)}\right) \\ g(x) &= \frac{1}{2}(\log(P(Y=1|X=x)) - \log(P(Y=-1|X=x))) \end{aligned}$$

por lo tanto, $g(x)$ tiene esta descrito por la ecuación 3.

$$g(x) = \frac{1}{2}(\log(P(Y=1|X=x)) - \log(P(Y=-1|X=x))) \quad (3)$$

Tomando el caso cuando $P(Y=1|X=x) > P(Y=-1|X=x)$, se tiene que $g(x) > 0$, por lo tanto $\hat{y}(x) = 1$. Por otro lado cuando $P(Y=-1|X=x) > P(Y=1|X=x)$, entonces $g(x) < 0$, por lo tanto $\hat{y}(x) = -1$. El comportamiento que presenta $g(x)$ es el mismo que tiene el clasificador óptimo bayesiano. Por lo que el problema planteado en la ecuación 1 conduce al clasificador óptimo bayesiano.

Problema 3

Supongamos que (X, Y) cumplen los supuestos del clasificador binario LDA. Sin embargo, a partir de una muestra de (X, Y) , alguien decide usar QDA (el clasificador bayesiano óptimo para el caso donde $X|Y = Y \sim \mathcal{N}(\mu_y, \Sigma_y)$, o sea no aprovechar que las covarianzas son iguales) y no LDA. ¿Cómo se comparan el error de entrenamiento de QDA con el de LDA para este caso? No hay que hacer cálculos formales sino dar argumentos intuitivos

El error de entrenamiento puede incrementar al usar QDA. Esto es debido a que el número de parámetros a estimar es mayor cuando se usa QDA en vez de LDA. Puede que al final se obtenga el mismo resultado que LDA, pero se tendrá un mayor costo computacional.

Problema 4

Este ejercicio es sobre el uso de métodos de clasificación para detectar billetes falsos:

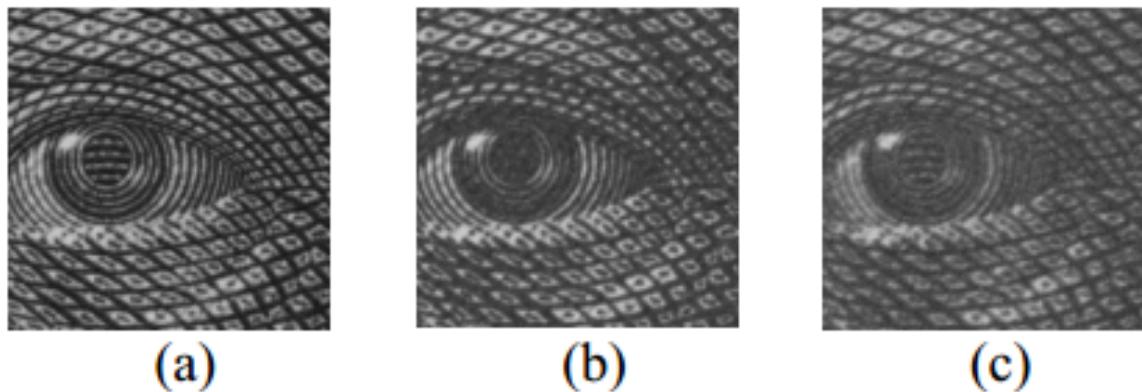


Figura 5: (a) (parde de) un billete de verdad, (b) billete falso de alta calidad, (c) billete falso de baja calidad.

En el paper que se anexa a la tarea se resume cada billete con cuatro características (varianza, skewness, curtosis y entropía) extraídas de la forma del histograma de los coeficientes de la transformación de Wavelet. Los histogramas a continuación muestran como cambia la forma cuando el billete ya no es auténtico.

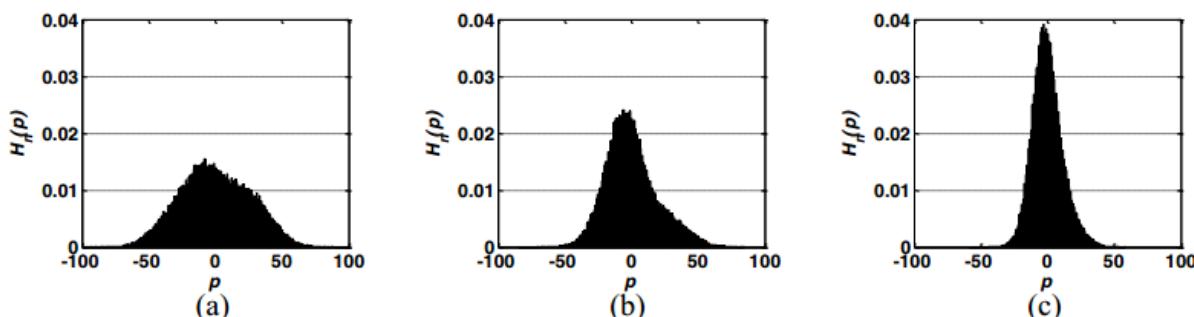


Figura 6: (a) histograma de los coeficientes de un billete de verdad, (b) billete falso de alta calidad, (c) billete falso de baja calidad.

Se anexo el conjunto de datos. La última columna indica si el billete es falso o no.

- Resume, visualiza y analiza los datos
- Construye algunos clasificadores interesantes basados en SVM (explora diferentes kernels). Estima su poder predictivo, para eso divide muchas veces los datos en conjunto de prueba y de entrenamiento y cuenta falsos positivos y falsos negativos. Las instrucciones básicas de SVM para R y Python están al final de [recpat4b.pdf](#)

Visualización de datos

En la figura 7 se visualizan los datos de varianza, skewness y curtosis por medio de planos.

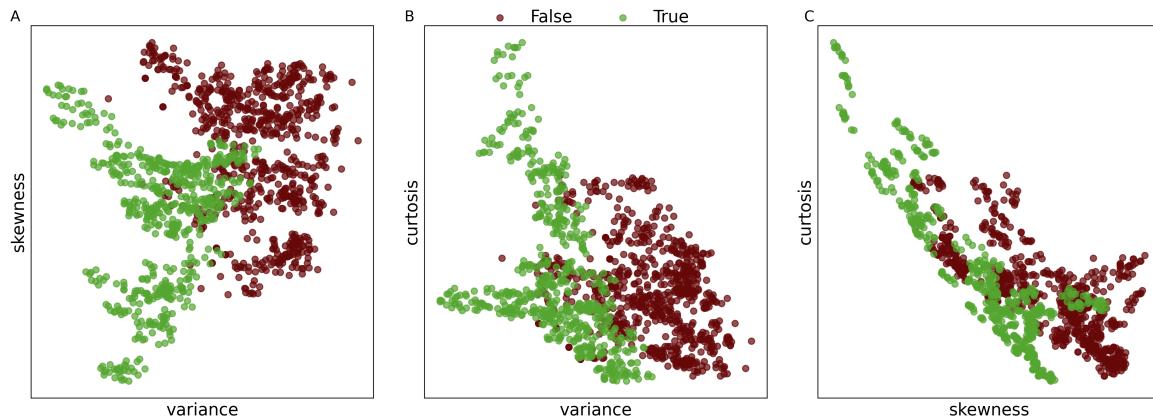


Figura 7: Visualización por planos de los datos de varianza, skewness y curtosis.

La configuración donde se obtiene una mayor diferencia entre cada conjunto de billetes es en el plano de la varianza y la curtosis. En la figura 8 se visualizan los datos usando cada característica para los planos. En esta visualización se logra apreciar de mejor manera que existe una separación entre los dos tipos de billetes.

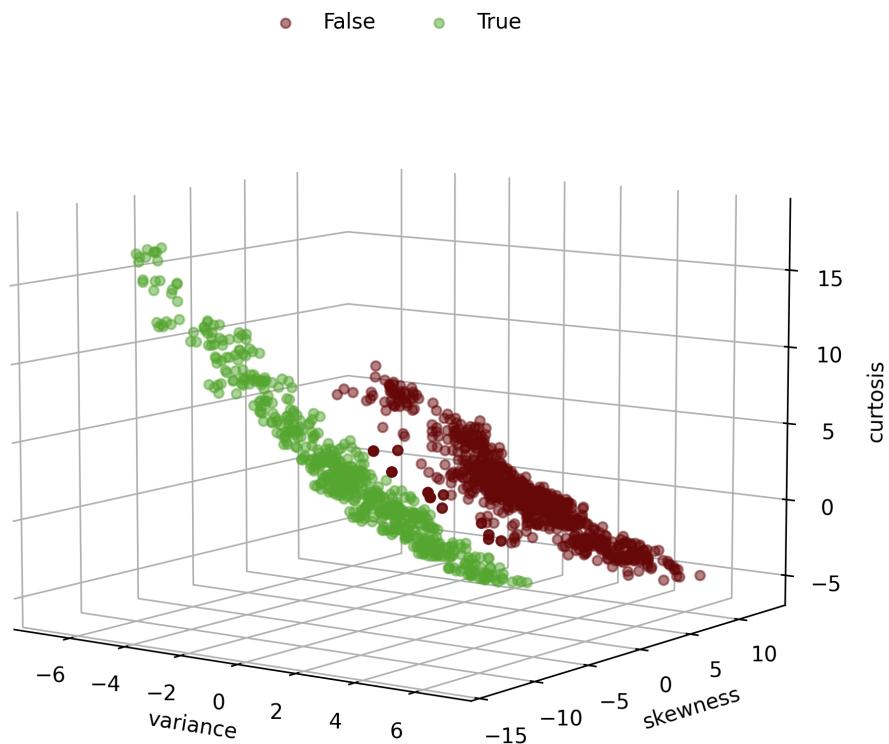


Figura 8: Representación gráfica de los datos usando cada característica como un plano.

Sea X una matriz con las características de cada billete donde $X \in R^{n \times 3}$ donde n es el número de billetes caracterizados. Se utilizó la matriz XX^T para iniciar el método de PCA con kernel lineal para obtener un conjunto de datos de dimensión $n \times 3$. En la figura 9 se visualizan las tres componentes principales obtenidas usando las combinaciones posibles en un plano

bidimensional.

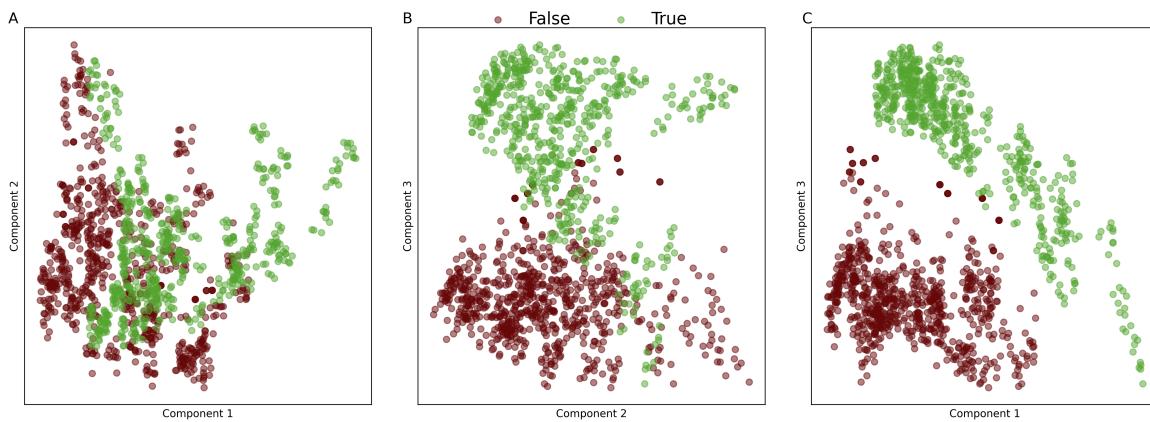


Figura 9: Representación gráfica de las tres componentes principales obtenidas de PCA usando la matriz XX^T .

Con esta representación se observa que existe una mayor distinción cuando la primer y tercer componentes son usadas como planos. En la figura 10 se visualizan las tres componentes obtenidas del método de PCA, cada componente usando un plano.

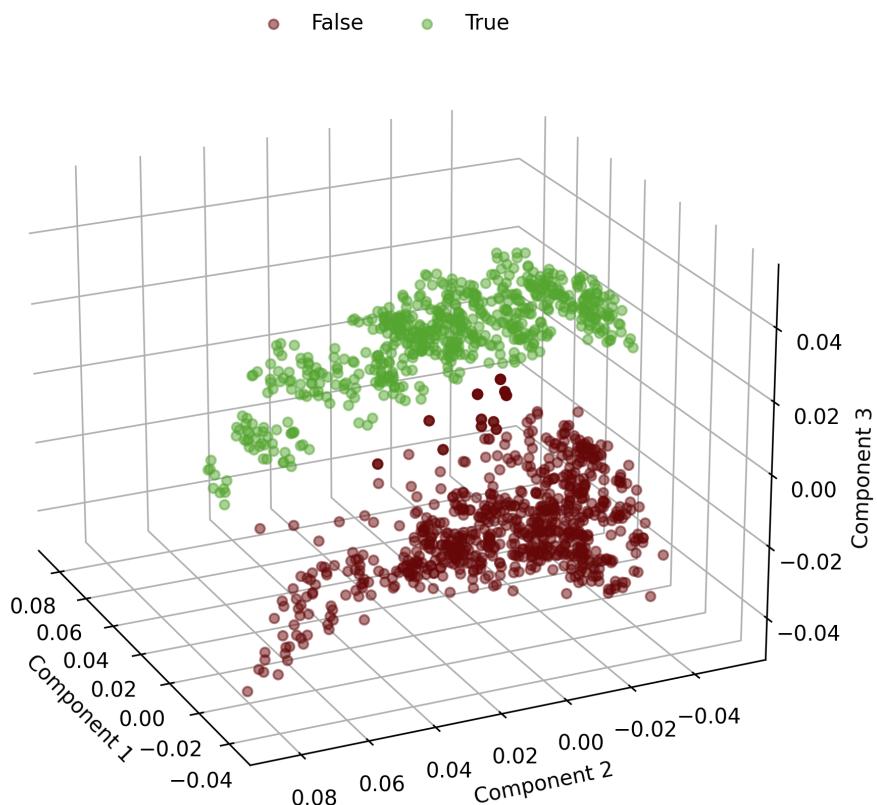


Figura 10: Representación gráfica de las tres componentes principales obtenidas de PCA usando cada componente como un plano.

Con esta representación se obtiene de igual manera una distinción entre cada conjunto de billetes. En comparación con la figura 8, se observa una mayor dispersión en los datos.

Para el método de SVM se usaron el kernel lineal, polinomial, rbf y sigmoide. En la tabla 1 se encuentran los parámetros de cada kernel usado.

Kernel	Grado	γ	r
Lineal	-	-	-
Polinomial	3	$\frac{1}{n\sigma^2}$	0
RBF	-	$\frac{1}{n\sigma^2}$	-
Sigmoide	-	$\frac{1}{n\sigma^2}$	0

Tabla 1: Parámetros usados para cada kernel. El simbolo – indica que no es necesario el parámetro en el kernel.

Se realizo una partición del conjunto de datos X . En esta partición de considero el 90 % como datos de entrenamiento y 10 % como datos de validación. Con los datos de validación se obtuvo un puntaje de accuracy. En la tabla 2 se encuentran los resultados de aplicar SVM y los datos X como se muestran en la figura 7.

Figura	Lineal	Polinomial	RBF	Sigmoide
A	0.674	0.645	0.819	0.558
B	0.927	0.906	0.899	0.811
C	0.978	1.000	1.000	0.877

Tabla 2: Puntajes de accuracy para las configuraciones mostradas en la figura 7.

Estos resultados pueden ser visualizados en la figura 11.

Con los datos obtenidos con PCA se obtuvieron los puntajes de accuracy mostrados en la tabla 3.

Figura	Lineal	Polinomial	RBF	Sigmoide
A	0.601	0.754	0.760	0.565
B	0.927	0.848	0.920	0.826
C	0.964	1.000	1.000	0.862

Tabla 3: Puntajes de accuracy para las configuraciones mostradas en la figura 9.

Estos resultados pueden ser visualizados en la figura 12.

Tratando a los conjuntos de datos mostrados en las figuras 8 y 10. Se realizo una comparación de los puntajes de accuracy, porcentajes de falsos positivos y porcentajes de falsos negativos. Los resultados que se obtuvieron son mostrados en la tabla 4.

Kernel	Originales			PCA		
	Accuracy	FN	FP	Accuracy	FN	FP
Lineal	0.971	0.014	0.014	0.978	0	0.022
Polinomial	0.978	0	0.022	1.0	0	0
RBF	1.0	0	0	1.0	0	0
Sigmoide	0.565	0.21	0.225	0.841	0.101	0.058

Tabla 4: Puntajes de accuracy para las configuraciones mostradas en la figura 8 y 10.

Con estos resultados se obtiene que por medio de los parámetros de varianza, skewness, curtosis y las primeras tres componentes obtenidas de PCA obtenidas de los parámetros se pueden crear clasificadores eficientes para la tarea de diferenciar billetes verdaderos y billetes falsos. El mejor kernel para la tarea es el RBF con los parámetros usados en la tala 1. Para obtener una mejor clasificación es usar todos los datos para generar el kernel. Usar PCA como entrada para el modelo de SVM produce mejores resultados, sin embargo para datos de mayor longitud esto puede ser perjudicial por el tiempo de computo.

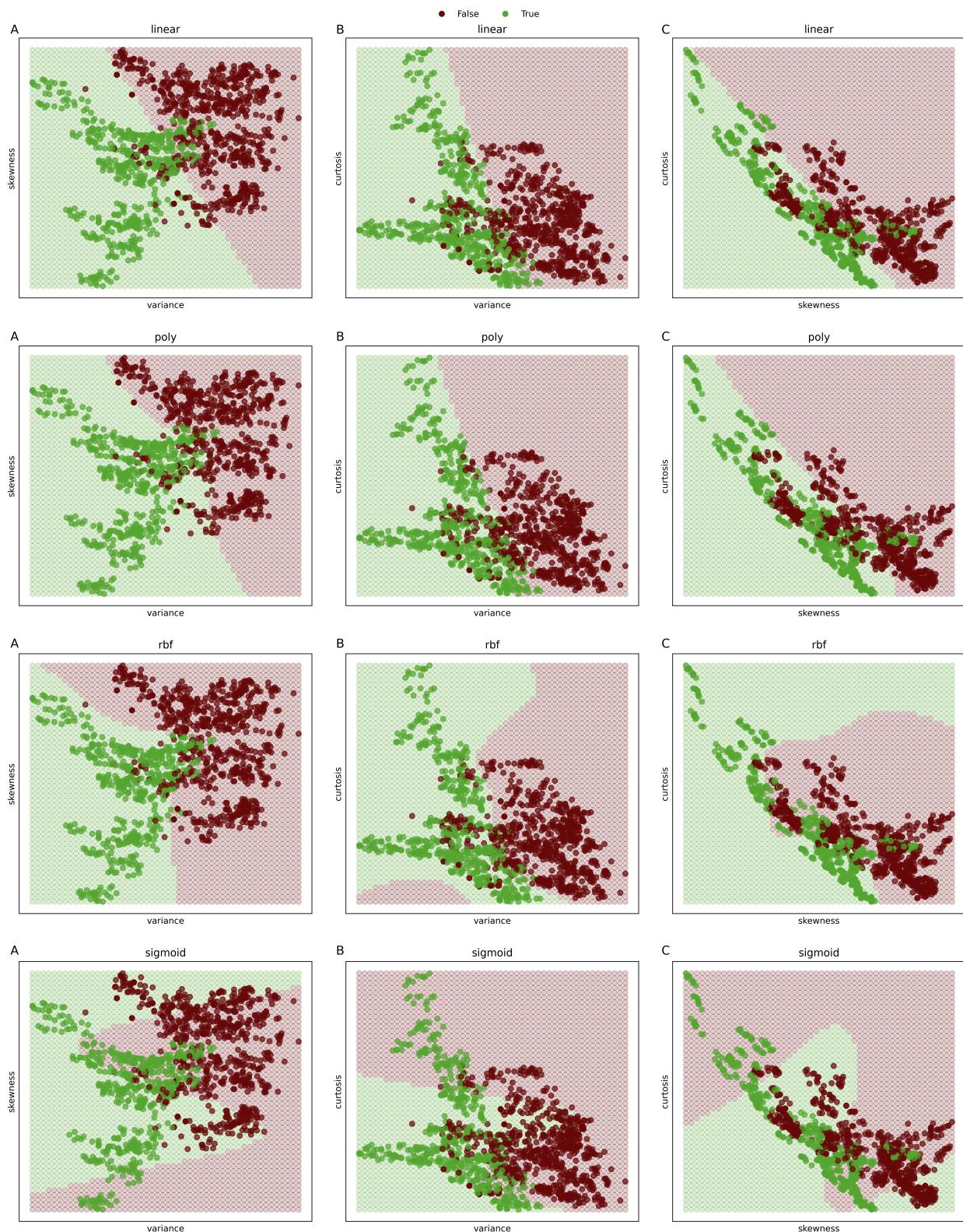


Figura 11: Representación grafica de la clasificación de los datos de varianza, skewness y curtosis como un plano de visualización bidimensional.

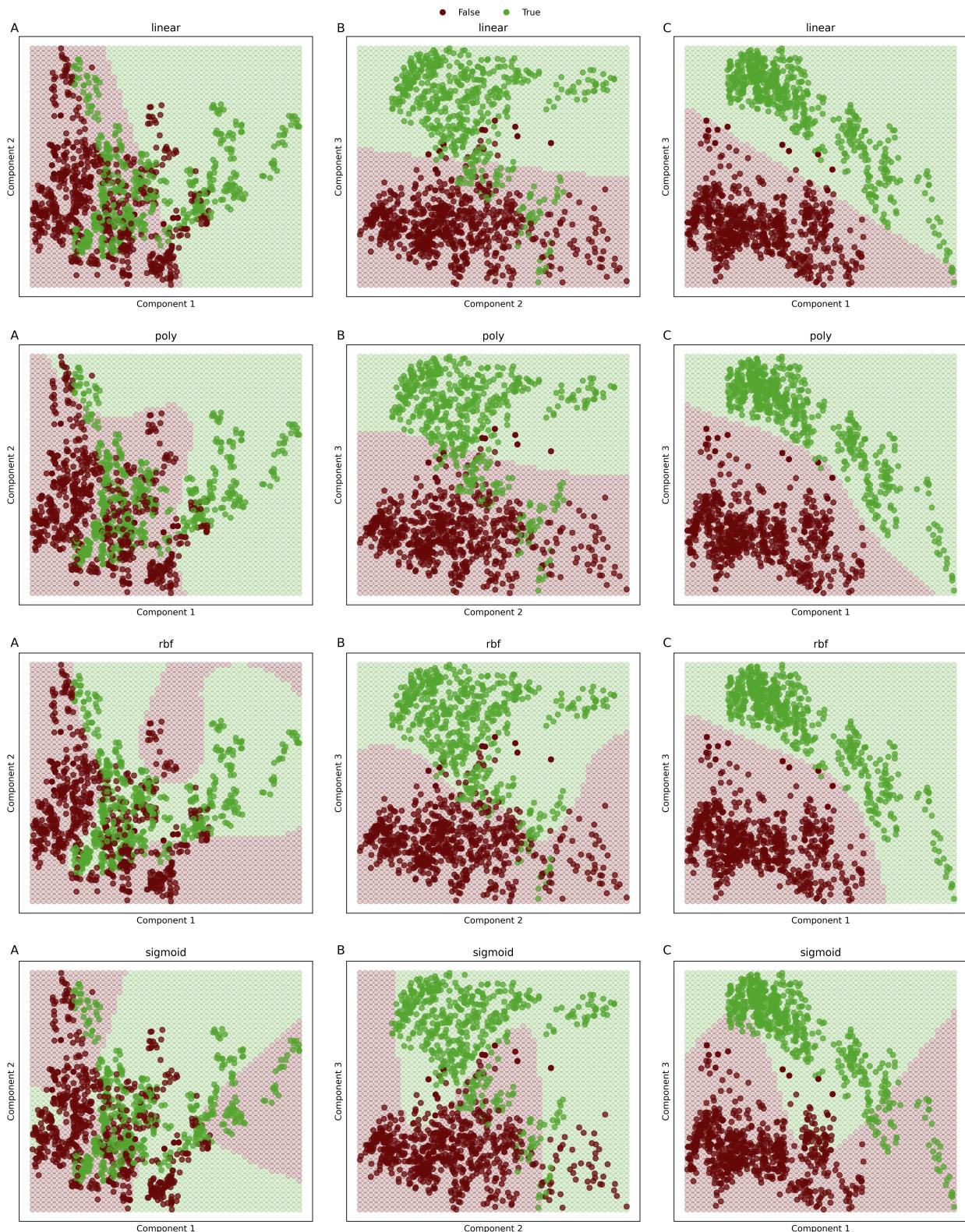


Figura 12: Representación grafica de la clasificación de los datos de las primeras tres componentes principales de PCA como un plano de visualización bidimensional.