**Curso:** Métodos basados en Kernels para aprendizaje automático
**Profesora:** Tamara Fernández

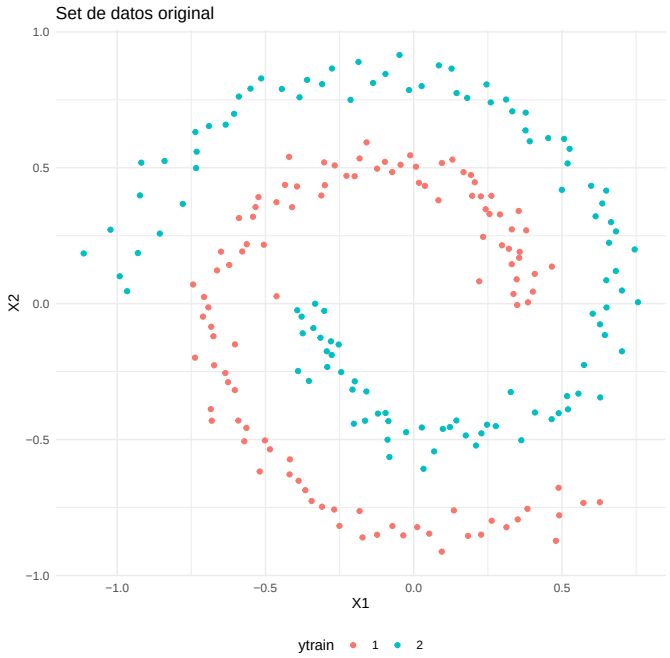
---

# Tarea 5

**Instrucciones de entrega:**

- Entregue un informe en formato **pdf** con sus resultados, pseudo codigo (si corresponde) y conclusiones. Además incluya su codigo en R o Python.
- Comente todas sus soluciones, y utilice herramientas gráficas, por ejemplo use el paquete `rgl` (u otro), si es que estima que esto le ayudará a presentar de mejor manera sus resultados.
- **Fecha de entrega:** 5 de Diciembre

**Ejercicio 1.** Considere los conjuntos de datos `simTrain.txt` y `simTest.txt`, los cuales representan los conjuntos de entrenamiento y prueba, respectivamente, para los datos mostrados a continuación:




1. Implemente vanilla SVM para construir un clasificador que permita distinguir entre los grupos 1 y 2. Reporte sus resultados en terminos de las medias de desempeño i)Accuracy, ii)Recall y iii)Precision. Su implementación debe considerar estandarización. El parámetro de regularización $C$ debe ser escogido de forma óptima para cada medida de desempeño. Para la búsqueda de este parámetro considere una grilla de al menos 1000 valores.

2. Repita el item anterior utilizando kernel PCA en conjunto con vanilla SVM. Para la implementación de kernel PCA utilice el kernel Gaussiano

$$K(x, y) = \exp\left\{-\frac{\|x - y\|^2}{\ell^2}\right\},$$

y optimize (puede buscar en una grilla de valores) el parámetro $\ell$ para maximizar cada una de las métricas mencionadas anteriormente. También optimize las medidas de desempeño respecto al parámetro de costo $C$ en vanilla SVM.

- Realice un análisis detallado que le permita elegir el menor número de componentes con las que usted debe quedarse para tener un buen desempeño (en terminos del Accuracy, Recall y Precision).
- Compare el desempeño de su algortimo (en terminos del Accuracy, Recall y Precision) cuando se estandarizan previamente los datos y cuando no. ¿Qué puede concluir?
- Compare sus resultados con los resultados con el modelo de clasificación vanilla svm.

**Ejercicio 2.** Considere el conjunto de datos `starbucks` del paquete `openintro`, que reune datos nutricionales para 77 productos vendidos por Starbucks. Esta base de datos contiene la siguiente información:

- `item`: Nombre del producto vendido.
- `calories`: Número de calorias.
- `fat`: Grasas (gramos).
- `carb`: Carbohidratos (gramos).
- `fiber`: Fibra (gramos).
- `protein`: Proteinas (gramos).
- `type`: Tipo de producto: `bakery, bistro box, hot breakfast, parfait, petite, salad,` y `sandwich`

1. En este ejercicio se le pide estudiar la distribución conjunta de la información nutricional dada por: `x=(calories, fat, carb, fiber, protein)` para los datos de tipo bakery y de otro tipo. Para esto implemente un test de hipótesis basado en el MMD (maximum mean discrepancy) utilizando un nivel de significancia $\alpha = 0{,}05$. Su informe debe considerar:
   - Detallar de forma explícita la hiótesis nula y alternativa a considerar.
   - Utilice el kernel gaussiano con length-scale $\ell = 0{,}5$. Su implementación debe considerar la estandarización de los datos.
   - Utilice Wild-Bootstrap usando $M = 5000$ muestras para encontrar la región de rechazo.
   - Muestre en un histograma las muestras generadas por Wild-Bootstrap y añada a este histograma el estadístico de prueba encontrado.
   - Obtenga una aproximación del p-valor y concluya.