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

Tarea 4

Instrucciones de entrega:

- Entregue un informe en formato **pdf** con sus resultados, pseudo codigo (si corresponde) y conclusiones. Además incluya su codigo en R.
- 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:** 10 de Noviembre

Ejercicio 1. En esta tarea, emplearemos el dataframe 'menu.csv', el cual hemos obtenido de Kaggle: Nutrition Facts for McDonald's Menu. Este conjunto de datos proporciona información detallada sobre los valores nutricionales de varios productos que se encuentran en el menú de McDonald's.

En todas las siguientes preguntas use **k-fold cross validation** con $k = 4$. Además, **realice un muestreo estratificado** para conservar las proporciones originales de cada clase.

1. (3 puntos) Utilice el algortimo de (kernel) svm para clasificar entre productos de tipo **Breakfast** y **productos de otro tipo**. Para este objetivo utilice todas las covariables numéricas en el dataset e implemente los modelos:

   a) vanilla svm,

   b) kernel svm, usando el kernel squared exponential dado por

   $$K_1(\boldsymbol{x}, \boldsymbol{x}') = \exp\left\{-\frac{1}{\sigma^2}\|\boldsymbol{x} - \boldsymbol{x}'\|^2\right\}, y$$

   c) kernel svm, usando el kernel locally periodic dado por

   $$K_2(\boldsymbol{x}, \boldsymbol{x}') = \exp\left\{-\frac{1}{\sigma^2}\|\boldsymbol{x} - \boldsymbol{x}'\|^2\right\} \cdot \exp\left\{-\frac{2}{\ell^2}(\sin(\pi\|\boldsymbol{x} - \boldsymbol{x}'\|/p))^2\right\}.$$

   En todos los casos encuentre parametros que maximizen las metricas de evaluación: i)Accuracy y ii)F1-score. Utilice al menos 1000 combinaciones diferentes de parámetros para cada kernel y para el parametro de regularización C (1000 en total). Basado en sus resulados, ¿Con qué modelo se quedaría usted? Discuta sobre los hiperparametros encontrados y muestre gráficamente el Accuracy y F1-score encontrados en terminos de los parámetros.

En las siguientes preguntas use todos sus datos para entrenar sus algoritmos (es decir, no es necesario hacer cross-validation).

2. (1 punto) Para los kernels anteriores (vanilla y squared exponential) considerando sus mejores parámetros encontrados en la parte anterior, encuentre las 10 observaciones que son más difíciles de clasificar. Muestre estos datos graficamente utilizando un grafico de dispersión entre las Sodio `(Sodium)` y las Proteinas `(Protein)`. Finalmente pinte cada observacion de acuerdo a su clase. Comente.

3. (1 punto) McDonal's esta preocupado de mal clasificar productos de tipo **Breakfast** en otra categoria. Modifique el algoritmo de svm y cree una categoria intermedia donde se encuentre aproximadamente el 15 % de los productos más dificiles de clasificar con el fin de que a estos productos se les haga una inspección manual. Muestre sus resultados graficamente en terminos de las variables Sodio `(Sodium)` y las Proteinas `(Protein)`. Utilice el kernel gaussiano para resolver esta pregunta.

   Hint: Utilice el threshold para definir esta nueva categoría.

4. (1 punto) Utilizando el clasificador binario, escriba e implemente un algoritmo que le permita clasificar entre productos de tipo i)**Breakfast**, ii)**Beef & Pork** y **Chicken & Fish** y iii) **otros productos** (solo puede utilizar una libreria que le permita hacer clasificación binaria). Implemente su algoritmo utilizando kernel svm con el kernel squared-exponential y todas las covariables en el dataset. Muestre la matriz de confusión obtenida y comente sus resultados