Reconocimiento de patrones y aprendizaje automático
Tarea 2: Métricas de clasificación

Fecha de entrega: Viernes 10 de diciembre de 2021

Profesor: Andrés Aldana Gonzáles
Ayudante: Felipe Navarrete Córdova

1. Ejercicios

1. **Máquinas de Soporte Vectorial (3 pt).**

   a) Describa detalladamente qué son y cómo funcionan las Máquinas de Soporte Vectorial (SVM)

   b) Describa el proceso de entrenamiento de una SVM

   c) ¿Qué es el *kernel* de una SVM?

   d) ¿Cuáles son las ventajas y desventajas de usar kernels polinomiales de alto y bajo grado?

   e) Explique detalladamente qué es el *kernel trick*

   f) Explique detalladamente el kernel basado en funciones de base radial gaussianas (Gaussian RBF Kernel)

2. **MNIST y regresión logística (7 pts).**

   El conjunto de datos MNIST contiene 70,000 imágenes de dígitos escritos a mano. Cada imagen de 28x28 píxeles está representada por un vector de 784 píxeles. Cada píxel puede tomar uno entre 256 tonos de gris, donde 0 es negro absoluto y 255 es blanco absoluto. Todos los vectores tienen asociada una etiqueta que indica el dígito escrito en la imagen.

   El objetivo de este ejercicio es construir un clasificador que permita identificar el dígito al que corresponde la información de cada imagen.

   a) Descargue el archivo de datos *mnist-original.mat* desde Kaagle (click aquí) y agréguelo a su directorio raíz.

   b) Cargue los datos y etiquetas en python. El archivo *MNIST_load.pynb* contiene un ejemplo de cómo cargar los datos de MNIST.

   c) Divida el conjunto de datos en Entrenamiento (50 %), Calibración (20 %) y Prueba (30 %).

d) Utilizando regresión logística, construya un clasificador para MNIST, para ello debe considerar un clasificador binario capaz de reconocer cada dígito. Use el conjunto de entrenamiento para entrenar cada sub-clasificador.

e) Dado un clasificador que reconoce al dígito i, la regresión logística indica la probabilidad de que los datos de entrada x pertenezcan a la clase i, $P_i(x)$. Dado un umbral de probabilidad θ_i, se considera que x pertenece a la clase i si $P_i(x) \geq \theta_i$. Use el conjunto de calibración para evaluar el rendimiento de cada clasificador y determinar el umbral θ_i que hace el mejor compromiso entre la taza de verdaderos positivos y la taza de falsos positivos. Para ello, en cada clasificador:

1) Calcule una matriz de confusión usando los datos de calibración.
2) A partir de la matriz de confusión, grafique en un solo panel las curvas de Precisión y Recall en función del umbral (al menos 100 valores).
3) Grafique una curva ROC mostrando el compromiso entre TPR y FPR para cada valor del umbral (al menos 100 valores).
4) Obtenga el valor del umbral que simultáneamente maximice la TPR y minimice la FPR.

f) Evalúe el rendimiento del clasificador MNIST completo usando los datos de prueba:

1) Obtenga la matriz de confusión final para el clasificador MNIST considerando las 10 clases.
2) Calcule las métricas Micro-F1, Macro-F1, Weighted F1 y Average Accuracy del sistema.

2. Entregables

La tarea se debe entregar en un notebook de Jupyter con los resultados y las gráficas correctamente discutidos. Los archivos de datos deben estar en el mismo nivel de directorio del notebook para facilitar la ejecución de los programas.

3. Bibliografía recomendada

- Aurelien Geron - Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow-O'reilly (2019). Chapter 3 - Classification