Tarea 3 EL4106 - Semestre Otoño 2019

Profesor: Javier Ruiz del Solar
Auxiliar: Patricio Loncomilla
Ayudantes: Gabriel Azócar, Nicolás Cruz, Francisco Leiva, Giovanni Pais

Fecha enunciado: viernes 03 de mayo de 2019.
Plazo entrega tarea: viernes 17 de mayo de 2019.

El objetivo de esta tarea es implementar un clasificador de revisiones de restaurantes usando SVMs. Se usará un dataset que contiene revisiones de restaurantes (una secuencia de palabras), cada una de las cuales tiene una etiqueta 0 (revisión desfavorable) o 1 (revisión favorable). En total hay 1000 revisiones disponibles. El archivo original se puede bajar de la siguiente dirección:

https://www.kaggle.com/hj5992/restaurantreviews

Dado que ese archivo contiene caracteres de retornos de línea en formato Linux, se subirá un archivo que puede ser leído también en editores de textos de Windows. Salvo por eso, los archivos son iguales. El formato del dataset es el siguiente:

But I don't like it.	0
Hell no will I go back	0
The fries were great too.	1

La base de datos estará disponible en u-cursos.

Se pide utilizar un modelo *Bag of Words* (BoW) y la metodología de clasificación estadística *Support Vector Machine* (SVM) para entrenar y validar un clasificador de revisiones. El trabajo a ser realizado incluye analizar detalladamente el efecto en el rendimiento del clasificador mediante la utilización de distintos tipos de *kernel*s. En esta tarea tendrán que entrenar y calibrar 4 clasificadores SVM binarios, que utilicen distintos tipos de Kernels/grados.

Se pide:

1) Teoría:
   a) Explique de manera genérica el funcionamiento de los clasificadores SVM.
   b) Indique en qué consisten los kernels lineal, polinomial y gaussiano
   c) ¿Cuál es el efecto de mover el parámetro C?
   d) Explique en qué consiste el algoritmo Bag of Words
   e) Explique en qué consiste cross-validation y cómo influye la cantidad de folds

2) Implemente un código que lea el dataset y permita generar una lista con las líneas de texto del archivo de revisión, y otra lista con las etiquetas. Como indicación, las etiquetas están separadas del texto mediante un carácter de tabulación '\t'. Puede usar la función split('\t') para separar la revisión de la etiqueta.

3) Implemente un código que permita generar una representación BoW para cada línea de texto. Se recomienda usar la función CountVectorizer( ) de scikit-learn.

4) Dividir la base de datos en 3 conjuntos representativos: entrenamiento (60%), validación (20%) y prueba (20%). Compruebe la representatividad de éstos, verificando si la proporción de cada clase se mantiene cercana a la proporción del conjunto completo.

5) Entrenar un clasificador SVM lineal inicial que permita discriminar revisiones favorables de desfavorables. Para obtener un buen clasificador, se debe usar una grilla para buscar los mejores parámetros para el clasificador. Se recomienda usar la función GridSearchCV( ) con 5 folds, considerando distintos parámetros C. El clasificador base SVM a usar se puede construir así: svm.SVC(kernel='linear', probability=True)

6) Evaluar sobre el conjunto de validación y generar la matriz de confusión. Se recomienda usar metrics.confusión_matrix( )

7) Generar una curva ROC que muestre el desempeño del clasificador, y el área bajo la curva. Se recomienda usar metrics.roc_curve( ) y metrics.auc( )

8) Repetir los pasos (5), (6) y (7) para kernels polinomial (dos grados distintos) y rbf

9) Evaluar el mejor clasificador obtenido sobre el conjunto de prueba, indicando las métricas indicadas en (6) y (7).

Los informes y códigos deben ser subidos a u-cursos a más tardar el día viernes 17 a las 23:59. Incluir un corto archivo de texto explicando cómo se utiliza su programa. Las tareas atrasadas serán penalizadas con un punto base por cada día de atraso.

Importante: La evaluación de la tarea considerará el correcto funcionamiento del programa, la inclusión de los resultados de los pasos pedidos en el informe, la calidad de los experimentos realizados y de su análisis, la inclusión de las partes importantes del código en el informe, así como la forma, prolijidad y calidad del mismo. Se subirá a u-cursos una pauta indicando la estructura del informe y los desgloses de los puntajes.