UNIVERSIDAD DE GRANADA

E.T.S. de Ingenierías Informática y de Telecomunicación

UNIVERSIDAD DE GRANADA

Departamento de Ciencias de la Computación e Inteligencia Artificial

Sistemas Inteligentes para la Gestión en la Empresa

Guión de Prácticas

Práctica 2: Clasificación de Imágenes

Curso 2016-2017

Máster Profesional en Ingeniería Informática

Práctica 2

Competición en Kaggle sobre Clasificación con Múltiples Clases

1. Objetivos y Evaluación

En esta segunda práctica de la asignatura Sistemas Inteligentes para la Gestión en la Empresa continuaremos estudiando el uso de algoritmos de aprendizaje supervisado para clasificación. Abordaremos un problema propuesto en la plataforma Kaggle (`https://www.kaggle.com/`) centrado en la clasificación de imágenes médicas. Los estudiantes adquirirán destrezas en análisis de datos avanzado, ampliarán conocimientos sobre el desarrollo y aplicación de técnicas de predicción basadas en *Deep Learning* y pondrán en práctica los contenidos teóricos de la asignatura referidos a la creación de multi-clasificadores y *ensambles* de clasificadores.

La práctica se desarrollará en grupos de dos personas. La evaluación se realizará en función de: (1) la posición final que ocupe el resultado propuesto por el estudiante (posición relativa respecto al conjunto de estudiantes); (2) la calidad de la memoria presentada. Para ser evaluado no bastará con subir los resultados a Kaggle; se deberá también adjuntar un documento que describa el proceso seguido por el estudiante para resolver la práctica. La práctica calificará para el **50 %** de la puntuación de prácticas; esto es, 3 puntos sobre 6. Las prácticas de calidad excelente podrán optar a mayor calificación.

2. Descripción del Problema y Reglas de la Competición

Se trabajará sobre la competición "Intel and Mobile ODT Cervical Cancer Screening":

`https://www.kaggle.com/c/intel-mobileodt-cervical-cancer-screening`

El problema consiste en identificar el tipo fisionómico de una cérvix femenina (parte inferior del útero) a partir de imágenes obtenidas con el sistema EVA (Enhanced Visual Assessment) de Mobile ODT, con el propósito de establecer el tratamiento preventivo o inicial más adecuado

para curar el cáncer de útero. Se trata de una competición avanzada, que ofrece incluso premios en metálico para los mejores clasificados.

Como primer paso para la realización de la práctica, cada grupo de dos estudiantes deberá formar un equipo en Kaggle utilizando los usuarios creados en la práctica anterior. El nombre del usuario y del equipo usado en Kaggle deberá comunicarse al profesor de prácticas para poder realizar su seguimiento.

3. Entrega

Se podrá competir en Kaggle hasta el **14 de junio de 2017**, fecha de cierre de la competición en Kaggle. Una vez finalizada esta fase, cada grupo deberá realizar la siguiente entrega antes del **23 de junio a las 23:59** a través de la web de la asignatura en `https://decsai.ugr.es` en un único archivo `zip` con el nombre (sin espacios) de los dos miembros del equipo: `P2 - primerapellido1 - nombre1 -- primerapellido2 - nombre2.zip`. Solo será necesario hacer una entrega por grupo. El documento `pdf` con la memoria tendrá el mismo nombre.

La memoria deberá documentar en detalle el trabajo realizado, aportando tablas, gráficas y cualquier material de apoyo. Se deberán incluir, al menos, los siguientes apartados:

1. Portada: Incluirá el nombre de los estudiantes, nombre del equipo usado en Kaggle, ranking global del equipo en la competición, puntuación del equipo.

2. Exploración de datos: Descripción y discusión de las técnicas utilizadas para estudiar la estructura y la semántica de los datos y los hallazgos preliminares, así como discusión y justificación de decisiones iniciales sobre el proceso que se llevará a cabo.

3. Preprocesamiento de datos: Descripción y discusión de las técnicas de preprocesamiento utilizadas y análisis crítico de su utilidad en el problema.

    - Integración y detección de conflictos e inconsistencias en los datos: valores perdidos, valores fuera de rango, ruido, etc.
    - Transformaciones: normalización, agregación, generación de características adicionales, etc.
    - Reducción de datos: técnicas utilizadas para selección de características, selección de ejemplos, discretización, agrupación de valores, etc.
    - Aumento de datos: técnicas utilizadas para incrementar la cantidad de datos disponibles.

4. Técnicas de clasificación: Discusión de las técnicas y herramientas de clasificación empleadas, justificación de su elección. Por ejemplo:

    - *Learning from scratch* vs *fine-tuning*

- Uso de CNNs + OVO
- Post-procesamiento OVO
- Otros: *feature maps*, *ensambles*, etc.

5. Presentación y discusión de resultados: Descripción y discusión de las soluciones obtenidas, incidiendo en la interpretación de los resultados. Análisis comparativo en caso de utilizar diferentes técnicas y/o parámetros de configuración en diferentes aproximaciones.

6. Conclusiones y trabajo futuro: Breve resumen de las técnicas aplicadas y de los resultados obtenidos, así como ideas de trabajo futuro para continuar mejorando las soluciones desarrolladas.

7. Listado de soluciones: Tabla de soluciones, incluyendo una fila por cada solución subida a Kaggle durante la competición. El número de filas deberá coincidir con el número de intentos reflejado en la web de la competición. En cada fila se aportará, al menos, la siguiente información separada por columnas:

   a) número de solución,

   b) descripción breve del preprocesamiento de datos aplicado,

   c) enumeración de los algoritmos y software empleados,

   d) resultado de porcentaje de acierto sobre conjunto de ejemplos etiquetados (extraído a partir del conjunto de entrenamiento, que puede ser el valor medio si se aplica validación cruzada u otro método similar),

   e) resultado de porcentaje de acierto sobre conjunto de ejemplos no etiquetados (*public score* obtenido en Kaggle),

   f) posición ocupada en el ranking de Kaggle en el momento de subir la solución (calculada seleccionando solo esta solución para *final score*) y fecha/hora de la subida.

   En esta tabla se debe resaltar la mejor solución obtenida; normalmente será la más reciente.

8. Bibliografía: Bibliografía utilizada.