Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Departamento de Ciencia de la Computación

IIC2115 – Programación como Herramienta para la Ingeniería (I/2018)

Tarea 4 - Análisis de datos

Objetivos

- Utilizar librerías de cálculo científico y análisis de datos en Python.
- Extraer y procesar datos a partir de un *dataset* real.
- Entrenar modelos predictivos y aplicarlos para guiar el comportamiento de un robot.

Entrega

- **Lenguaje a utilizar:** Python 3.6
- **Lugar:** GitHub
- **Fecha:** 29 de Junio
- **Hora:** 23:59
- **Desarrollo en parejas o individual**
- Las inscripciones para las parejas son hasta el día 18 de Junio a las 23:59 y solo se aceptan las inscritas por este FORMULARIO.

Introducción

En esta tarea, deberán aplicar todo el conocimiento que han adquirido en modelos predictivos para construir un sistema que permita realizar predicciones en base a información visual. En particular, deberán utilizar técnicas de aprendizaje de máquina supervisado para construir un sistema que permita, mediante análisis de imágenes, tomar decisiones.

"Estoy perdido en un supermercado"

En un futuro cercano, el agente robótico autónomo HANS9000[1] decide ayudar a su *amo* Hugo, realizando las compras semanales en el supermercado[2]. Lamentablemente, HANS9000 no se preocupó de cargar sus baterías y llegó al supermercado con la carga mínima. Así, al momento de entrar, HANS9000 perdió el conocimiento (sea lo que sea esto) y despertó con sus baterías cargadas a medias, en un lugar arbitrario del supermercado. Al sólo poseer un sistema de localización rudimentario basado en GPS, HANS9000 no fue capaz de ubicarse correctamente en el supermercado (al encontrarse bajo techo) por lo que debió pedirle a su *amo*, mediante un correo electrónico, que lo vinieran a buscar.

Muy avergonzado ante esta situación, HANS9000 decide mejorar su sistema de localización en supermercados, agregando un módulo de procesamiento visual. Para esto recurre a los alumnos del curso "Programación Como Herramienta para la Ingeniería" con el requerimiento de construir un sistema de clasificación de imágenes, que permita no depender del GPS. Con el fin de colaborar en la construcción de un prototipo, HANS9000 hace disponible las siguientes dos fuentes de información de un supermercado en particular:

- Base de datos con el listado de los productos existentes en los pasillos de un supermercado. El formato de esta base de datos es un archivo de texto simple, donde cada fila indica el nombre del pasillo y los nombres de los productos existentes en él. Todos los pasillos tienen la misma cantidad de productos.
- Base de datos con imágenes para todos los productos disponibles en los pasillos. Para cada producto, existirán múltiples imágenes, con distintas condiciones (punto de vista, oclusión, etc.).

Producto de un error de consistencia en la programación del robot, ambas bases de datos presentan errores, por lo que se deberá realizar una etapa de procesamiento previo al entrenamiento de sus modelos. En el siguiente enlace podrá encontrar los *datasets*.

[1] Predecesor de HAL9000.

[2] Se esperaría que a estas alturas ya no fuera necesario ir al supermercado, pero para efectos de esta tarea obviaremos eso.

Código auxiliar

En el archivo adjunto `features.py` se encuentra la función `get_features()`, que recibe como *input* la ruta a la carpeta con las imágenes, y retorna un diccionario donde la llave es el nombre de la imagen, y el valor es un diccionario con atributos útiles de la imagen: su categoría, la imagen en sí, y descriptores visuales para el entrenamiento de su modelo predictivo.

Preprocesamiento y entrenamiento

Como ya se dijo, las bases de datos presentan problemas. En particular:

1. La base de datos de pasillos posee líneas corruptas sin sentido alguno. Se deben identificar dichas líneas, y eliminarlas.

2. La base de datos de imágenes presenta algunos elementos con extension `.hans`. Estas imágenes son duplicados (claramente no tiene sentido conservarlas como tal), por lo que deberán ser reemplazadas[3] por la media del total de las fotos de la categoría en cuestión.

Deberán hacer uso del código auxiliar para el desarrollo del punto 2 y el entrenamiento de sus modelos. Se exige que la información del diccionario sea insertada a un DataFrame de Pandas, y luego leída desde ahí.

Predicciones

Para evaluar el sistema, HANS9000 recibirá el archivo adjunto `test.txt`, donde cada línea representa un nuevo pasillo, compuesto por los nombres de las imágenes de los productos que lo conforman. A continuación se muestra un ejemplo de este formato, con el contenido de un archivo que describe dos pasillos con cuatro imágenes cada uno:

```
2213.jpg,147.jpg,20.jpg,3101.jpg
12.jpg,456.jpg,1423.jpg,2556.jpg
```

Para cada línea se debe retornar la categoría predicha para cada imagen, así como también el pasillo al que más se parece de los vistos en `pasillos.txt`. Está permitido utilizar cualquier algoritmo de aprendizaje supervisado, no sólo los que hemos revisado en clases, siempre y cuando se justifique brevemente su elección. Se espera que usted pruebe cada descriptor por separado, analice los rendimientos, y se quede con el mejor para su predictor final. No se castigará bajo rendimiento, aunque sí existirá *bonus* por buen desempeño.

[3] Conservando la extensión de archivo.

Entrega

Deberá entregar su trabajo en formato *.ipynb*, comentando con *markdown* su desarrollo en detalle, el análisis de los resultados obtenidos, y también un breve *feedback* respecto a la tarea. **Para esta tarea en especial**, se espera que los alumnos de la Escuela de Ingeniería mantengan un comportamiento acorde al Código de Honor de la Universidad.

Política de Integridad Académica

"Como miembro de la comunidad de la Pontificia Universidad Católica de Chile me comprometo a respetar los principios y normativas que la rigen. Asimismo, prometo actuar con rectitud y honestidad en las relaciones con los demás integrantes de la comunidad y en la realización de todo trabajo, particularmente en aquellas actividades vinculadas a la docencia, el aprendizaje y la creación, difusión y transferencia del conocimiento. Además, velaré por la integridad de las personas y cuidaré los bienes de la Universidad."

En particular, se espera que mantengan altos estándares de honestidad académica. Cualquier acto deshonesto o fraude académico está prohibido; los alumnos que incurran en este tipo de acciones se exponen a un procedimiento sumario. Ejemplos de actos deshonestos son la copia, el uso de material o equipos no permitidos en las evaluaciones, el plagio, o la falsificación de identidad, entre otros. Específicamente, para los cursos del Departamento de Ciencia de la Computación, rige obligatoriamente la siguiente política de integridad académica en relación a copia y plagio: Todo trabajo presentado por un alumno (grupo) para los efectos de la evaluación de un curso debe ser hecho individualmente por el alumno (grupo), sin apoyo en material de terceros. Si un alumno (grupo) copia un trabajo, se le calificará con nota 1.0 en dicha evaluación y dependiendo de la gravedad de sus acciones podrá tener un 1.0 en todo ese ítem de evaluaciones o un 1.1 en el curso. Además, los antecedentes serán enviados a la Dirección de Docencia de la Escuela de Ingeniería para evaluar posteriores sanciones en conjunto con la Universidad, las que pueden incluir un procedimiento sumario. Por "copia" o "plagio" se entiende incluir en el trabajo presentado como propio, partes desarrolladas por otra persona. Está permitido usar material disponible públicamente, por ejemplo, libros o contenidos tomados de Internet, siempre y cuando se incluya la cita correspondiente.