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

IIC2115 – Programación como Herramienta para la Ingeniería (I/2021)

Laboratorio 2: análisis de datos con Python

Objetivo

Aplicar los conocimientos de análisis de datos en Python para procesar un conjunto de datos reales.

Entrega

- **Lenguaje a utilizar:** Python 3.6 o superior
- **Lugar:** repositorio privado en GitHub. Recuerde incluir todo en una carpeta de nombre **L2**.
- **Entrega: jueves 8 de julio a las 23:59 hrs.**
- **Formato de entrega:**
  - Archivo Python Notebook (**L2.ipynb**) con la solución de las misiones de este laboratorio. Utilice múltiples celdas de texto y código para facilitar la revisión de su laboratorio. **Deje todo ejecutado antes de realizar su commit**, se recomienda utilizar la opción de "restart and run all" disponible en jupyter notebook.
  - Archivo python (**L2.py**) con el mismo código disponible en su archivo ipynb.
  - Todos los archivos deben estar ubicados en la carpeta **L2**. No se debe subir ningún otro archivo a la carpeta. **No suba las bases de datos a GitHub o tendrá un descuento adicional inapelable de 5 décimas**.
- **Descuentos:** el descuento por atraso se realizará de acuerdo a lo definido en el programa del curso. Además de esto, tareas que no cumplan el formato de entrega tendrán un descuento de 0,5 ptos.

- **Laboratorios con errores de sintaxis y/o que generen excepciones serán calificadas con nota 1.0.**
- Si su laboratorio es entregado fuera de plazo, tiene hasta el **viernes 9 de julio a las 11:59AM hrs** para responder el formulario de **entregas fuera de plazo** disponible en el Syllabus.
- Las discusiones en las *issues* del Syllabus en GitHub son parte de este enunciado.
- El uso de librerías externas que sean estructurales en la solución del los problemas no podrán ser utilizadas. Solo se podrán utilizar las que han sido aprobadas en las *issues* de GitHub.

Introducción

Al igual que en la A4, en este laboratorio deberá realizar una serie de procesamientos y visualizaciones de datos para finalmente predecir las ventas de una serie de tiendas de la cadena Rossmann. A diferencia de la actividad, en esta ocasión tendrá a su disposición un conjunto de datos de mayor tamaño, además de un segundo conjunto de datos que deberá incluir en su análisis.

Descripción de los datos

La fuente primaria de datos para entrenar los modelos será el set "Rossmann Store Sales", que se encuentra disponible en `https://www.kaggle.com/c/rossmann-store-sales/data`. En esta ubicación podrá además encontrar información detallada sobre cada una de las variables. El set contiene información sobre las ventas de 1.115 tiendas de la cadena Rossmann. Los datos están estructuras en cuatro archivos distintos, dentro de los cuales solo utilizará los siguiente dos:

- `train.csv`: contiene 1.017.210 registros de ventas en distintos días de las 1.115 tiendas. Acá se encuentra la variable a predecir, `Sales`.
- `stores.csv`: contiene 1.115 registros (uno por tienda), que proveen información adicional para cada tienda.

Para descargar el set de datos deberá crear una cuenta gratuita en Kaggle, la que además les permitirá revisar gran cantidad de archivos de código de personas que han utilizado este set de datos.

IMPORTANTE

Para cumplir las misiones de este laboratorio, es su responsabilidad explorar inicialmente el contenido de los archivos y familiarizarse con el formato en que está almacenada la información.
Recuerde además codificar numéricamente los valores de las columnas categóricas y normalizar las numéricas cuando corresponda. El laboratorio no considera puntaje por hacer esto, pero sí descuentos cuando no es realizado o es realizado en una variable o momento incorrecto.
Para definir los conjuntos de prueba y validación, independiente del tamaño que tengan, considere siempre el x% más reciente de los registros, y no particiones aleatorias.

Misiones

M1. Entrenamiento validado (3 ptos.): entrene al menos 3 modelos distintos de regresión para predecir el logaritmo del valor de la variable `Sales`, solo utilizando el contenido del archivo `train.csv`. A diferencia de instancias anteriores, considere un set de validación que permita elegir para cada modelo el conjunto óptimo de hiperparámetros. Investigue sobre el uso de técnicas como validación cruzada (*cross-validation* o *leave-one-out* para esto y utilice al menos una de ellas. Finalmente, compare el rendimiento en el set de test de todos lo modelos, y de las versiones de estos modelos sin validar, es decir, solo entrenados en el set de entrenamiento y evaluados en el set de prueba, usando los hiperparámetros por defecto de `sklearn`.

M2. Uso de nuevas variables (1 pto.): cruce el contenido de ambos archivos y genere en base a esto nuevas variables para los registros. Repita el mismo procedimiento de la misión anterior con este nuevo conjunto de datos. Compare y analice el rendimiento, poniendo énfasis en la importancia de las nuevas variables.

M3. Visualización (2 ptos.): Investigue sobre el uso de técnicas de reducción de dimensionalidad, como `PCA` o `tSNE`. Utilice alguna de estas para visualizar en 2D el espacio de características generado en la misión anterior, destacando aquellos puntos que presentan un mayor nivel de error en la estimación, por ejemplo, pintándolos con una escala de colores correlacionada con el nivel de error. Analice los resultados y comente sobre la existencia de grupos muestras con alto error.

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.