Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Departamento de Ingeniería de Transporte y Logística

ICT3115 – Sistemas Urbanos Inteligentes (I/2021)

Tarea 1: procesamiento de datos tabulados

Introducción

En esta tarea tendrá la oportunidad de experimentar con el uso de redes neuronales para el procesamiento de datos tabulados. En particular, deberá entrenar MLPs para regresión, con entradas continuas y categóricas, utilizando *embeddings* cuando corresponda. Para el desarrollo se utilizará el framework Pytorch sobre la plataforma Colab de Google. El objetivo de este proceso será estimar las ventas de una serie de tiendas de la cadena Rossmann (ejemplo visto en clases). Dado que existe abundante material disponible en línea relacionado con el desarrollo de la tarea, se espera que todo recurso externo utilizado, sea este código o librerías, esté debidamente indicado.

Set de 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 crear sets de entrenamiento, validación y test independientes, utilice las funcionalidades de

`scikit-learn`, como han sido indicadas en los ejemplos de código. Todas las variables, con la excepción de `Sales`, pueden ser utilizadas para el entrenamiento.

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.

Modelos

Para esta tarea, debe utilizar modelos de redes profundas como los descritos en el capítulo 2 del curso, con la exigencia de que debe considerar el uso de *embeddings* para las variables categóricas. Se recomienda revisar bibliografía relacionada para esto y además diseñar las capas teniendo en consideración el tipo de dato que procesará (por ejemplo, en el caso de las variables categóricas, el número de categorías es relevante). Considere además preprocesar las entradas y salidas numéricas (revise el ejemplo de código disponible en el sitio del curso). No hay problema en basarse completamente en algún modelo propuesto previamente en la literatura o en tutoriales. En cualquier caso, debe justificar su elección de modelo.

Tareas a realizar

Para la tarea se espera se espera que realice al menos las siguientes tareas:

- Entrenamiento y selección de modelos: entrene distintas arquitecturas utilizando solo las variables disponibles en el archivo `train.csv` y evalúe sus rendimientos en un set de test independiente . Seleccione el modelo que mejor rendimiento entrega y justifique su elección en base a las características de la arquitectura.
- Entrenamiento y selección de modelos parte 2: repita la tarea anterior, pero esta vez incorporando las variables definidas en el archivo `stores.csv`. Compare el rendimiento y sobreentrenamiento con el modelo elegido para la tarea anterior.
- Visualización de `embeddings`: Seleccione tres variables categóricas y grafique para cada una las transformaciones realizadas por sus respectivas capas de `embedding`. Analice e interprete los resultados en base a la semántica de las variables elegidas.

Desarrollo y entrega

La tarea puede desarrollarse de manera individual o en parejas, utilizando el framework Pytorch para Python. Se recomienda utilizar la plataforma Google Colab con el fin de facilitar la instalación de

librerías. Esta plataforma permite utilizar gratuitamente una GPU para el entrenamiento por intervalos de 12 horas continuos. En el *notebook* desarrollado debe ir tanto el código como un informe (preferiblemente intercalados), donde se expliquen los pasos realizados, se analicen los resultados y se planteen conclusiones. La entrega de la tarea tiene como fecha límite el viernes 14 de mayo a las 23:59, a través del buzón que se habilitará en el sitio del curso. Para fines de corrección, se revisará la última versión entregada.

Política de Integridad Académica

Los alumnos de la Escuela de Ingeniería deben mantener un comportamiento acorde al Código de Honor de la Universidad:

"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.