Temas selectos de ciencia de datos. Proyectos 2023.

Sobre los proyectos.

- Los proyectos consistirán en una o varias de las siguientes actividades: desarrollo, implementación, aplicación y análisis, de alguna temática relacionada con el curso.
- Los proyectos podrán realizarse en equipo de máximo 3 personas. La exigencia en cuanto al desarrollo del proyecto será en función del número de integrantes del equipo.
- Deberá especificarse el trabajo desarrollado por cada uno de los integrantes, en todos los entregables que se realicen.
- La dinámica de las revisiones y presentación final de los proyectos se harán según lo mencionado al inicio del curso.
- La evaluación final del proyecto consistirá en una presentación y la elaboración de un artículo sobre el proyecto.
- Se creará un repositorio de GitHub para cada proyecto, con acceso para cada integrante del equipo y el profesor.

Proyectos

1. **Análisis y modelación de señales a partir de datos FITS**

   Flexible Image Transport System (FITS) es un formato estándar para almacenar datos astronómicos. El objetivo de éste proyecto es abordar una tarea relacionada con éstos datos, entender el fenómeno que se está representando y las señales que se generan a partir del mismo, la forma en que se almacena ésta información, y utilizar métodos de machine learning para analizarlos desde una perspectiva de ciencia de datos.

2. **Image captioning para radiografías**.

   Image captioning (IC, subtitulado de imágenes) es una tarea que involucra un modelo de generación de lenguaje para "transcribir" el contenido visual de una imagen en texto. En éste proyecto, se abordará ésta tarea en un contexto específico para diagnóstico asistido en radiografías de rayos X de tórax. Se tienen 2 objetivos: el primero es entender y adaptar un modelos de IC para rayos X, incluido uno basado en transformers. El segundo es explorar la base de datos MIMIC y la posibilidad de adaptar los modelos para diagnósticos asistidos en español.

3. **Métodos de DL para análisis de tractografía mediante datos de fMRI**.

   El objetivo de éste proyecto es abordar el análisis de datos obtenidos mediante fMRI y su aplicación en tractografía del cerebro. Hay dos objetivos principales. Primero, entender la problemática, los tipos de datos de fMRI y sus aplicaciones, así como hacer una revisión de los métodos tradicionales para abordar éstas problemáticas. Segundo, se abordarán las propuestas basadas en DL para ésta tarea, haciendo una adaptación/implementación de un modelo, analizando su desempeño y las ventajas o desventajas que presenta respecto a los modelos propuestos en la literatura.

4. **Análisis de señales de alta frecuencia para aplicaciones industriales**.

   Con el inicio y adopción del esquema de Industria 4.0, se motivó el análisis de datos de alta frecuencia generados por la sensorización de las líneas de producción y dispositivos relacionados, con aplicaciones principalmente en aseguramiento de calidad y mantenimiento predictivo. El objetivo de éste proyecto es abordar éstas aplicaciones desde una perspectiva de machine/deep learning para una tarea de predicción. Hay dos objetivos principales: explorar y obtener bases de datos relacionadas con ésta aplicación y explorar representaciones adecuadas de los datos, tanto en tiempo como en frecuencia, para aplicar métodos de aprendizaje en una tarea específica, por ejemplo, predicción de tiempo de falla.

5. **Análisis de emoción y sentimiendo en música desde una perspectiva multimodal**.

Una tarea muy popular en el área de recuperación de información musical (MIR) es la detección de emoción (mood) y sentimiento en obras musicales. En éste proyecto, se abordará ésta tarea bajo una perspectiva multimodal, explorando modelos que incluyan representaciones de información acústica, de texto y editorial.

6. **Reconocimiento automático de instrumentos en audios musicales**.

   En inteligencia artificial (AI), una tarea muy etudiada en MIR es la identificación automática de instrumentos en música polifónica. Gran parte de los métodos propuestos se basa en el análisis y extracción de características de la señal de audio para posteriormente, usar algún método de ML para identificar el instrumento. En éste proyecto, el objetivo es dar una revisión de los métodos usados para abordar ésta tarea, y adaptar o implementar, un método basado en DL en un conjunto de datos apropiado.

7. **Speech translation**.

   El objetivo en éste proyecto, es explorar e implementar métodos para la transcripción automática de audio a texto, específicamente, para el idioma español. Aunque hay métodos disponibles comercialmente, el objetivo es explorar alternativas eficientes para ésta tarea.

8. **Wavelets para series de tiempo**

   Una alternativa para obtener representaciones en tiempo-frecuencia de datos con dependencia temporal, como las series de tiempo, es la transformada wavelet. En éste proyecto, el objetivo es comprender la transformada wavelet y sus ventajas/desventajas respecto a otras representaciones en tareas de pronóstico o caracterización de series de tiempo. El énfasis principal es explorar métodos de Machine/Deep learning para éste tipo de datos basado en ésta transformación.

9. **Mecanismos y medidas de importancia de variables en textos.**

   Ya vimos en el curso, cómo los textos pueden considerarse como secuencias de tokens o "señales" que tienen un significado al considerarse en forma global. En éste proyecto, se explorarán representaciones contextuales de palabras y textos basadas en arquitecturas de DL tipo

transformers. Particularmente, se analizarán los mecanismos de autoatención como medida de importancia de las palabras respecto a alguna tarea a resolver, como clasificación de sentimiento, perfilado de autor o detección de agresividad.

10. **Redes neuronales para la generación de música**

    Una de las tareas más populares en MIR, después quizá de identificación de género musical, es la generación de música. Los modelos utilizados para ésta aplicación han evolucionado sustancialmente en los últimos años, siendo los más populares aquellos basados en DL, como redes recurrentes LSTM o GRU, y transformers. El objetivo de éste proyecto es hacer una revisión de éstas metodologías, explorar las bases de datos disponibles para el entrenamiento, y las diferentes representaciones usadas.