Pautas para la evaluación práctica y teórica de la asignatura

Curso 2024-2025

Análisis de Datos No Estructurados

Ana Laguna Pradas & Cristina Puente Agueda

La siguiente guía recoge la información necesaria para que el alumno comprenda las pautas a seguir para la evaluación tanto teórica como práctica de la asignatura de Análisis de Datos No Estructurados.

Sección Teórica

La evaluación de la parte teórica de la asignatura se llevará a cabo a través de 2 cuestionarios tipo test (uno para texto y otro para imagen/audio). La prueba se realizará en horario lectivo y la duración aproximada será de 30 minutos. Los alumnos grabarán la pantalla mediante OBS y subirán la grabación al Drive, dejando un enlace a la misma en Moodle.

Ambos exámenes contarán el 40% de la nota final.

El trabajo se podrá realizar por grupos de máximo 3 personas.

Sección Práctica

La evaluación de la sección práctica de la asignatura se llevará a cabo a través de 2 prácticas entregables en las fechas determinadas por el profesorado.

La nota de ambos trabajos contribuirá al 60% de la nota final. Se dedicarán dos sesiones de clase para asegurar que todos los alumnos avanzan correctamente. El trabajo se podrá realizar por grupos de máximo 3 personas.

Nota: Dado el carácter pragmático de la asignatura, se invita al alumnado a realizar prácticas con objetivos y bases de datos que sean de su propio interés y motivación, o que tengan una aplicabilidad en la vida real con el fin de despertar el espíritu emprendedor.

En aquellos casos en los que el alumnado prefira orientación sobre las prácticas, se propone la siguiente temática por bloques de tipología de datos:

Texto: Análisis de la letra de tus propias canciones (de Spotify o cualquier otra plataforma de música, podcast, etc.) por temática, autor, época.
Tras la descarga del dataset de archivos de audio (.wav, .mp3, etc), el alumno aplicará técnicas de speech-to-text / automatic speech recognition para extraer la letra de las canciones. Una vez extraído el texto se procederá a analizar el dato siguiendo las directrices de la sección "Estructura de las prácticas".

Imagen: Análisis de tus propias fotografías de Google Fotos o cualquier otro repositorio.
Tras la descarga de fotos de Google Fotos por temáticas (montaña, ciudad…), personas, animales, etc. se procederá a analizar las imágenes tal y como se detalla en la sección "Estructura de las prácticas".

Adicionalmente cada año suele realizarse alguna *competition* o *hackaton* el último día de clase, donde el equipo ganador suele beneficiarse de un aumento de su nota final.

Este año realizaremos el siguiente *hackathon* en clase:

Audio: Análisis acústico de tus propias canciones por temática, autor, época…
Haciendo uso del mismo dataset de canciones de la práctica de texto, en este caso analizaremos el mismo dato desde la perspectiva acústica analizando directamente los ficheros de audio.
En este caso, se necesitará traer el dataset a clase y se darán 2 horas para el análisis, y 1 hora para compartir resultados contando con grupos de 6 personas. A continuación se mandarán los resultados al profesor vía email con acceso directamente al repositorio de código. Se deliberará el equipo ganador antes de finalizar la clase tras decisión del jurado.

Estructura y evaluación de las prácticas

La idea de las prácticas es que el alumno tenga la oportunidad de acercarse a un proyecto real. Para ello debería ser capaz de realizar un proyecto desde 0 de cualquier dato no estructurado siguiendo el pipeline de Data science visto en clase. Siempre se partirá del análisis exploratorio, extracción de features, Machine Learning, Deep learning generativo y discriminativo (este último incluyendo redes from scratch y modelos preentrenados) para finalmente comparar resultados con herramientas del mercado (e.g., probamos nuestro clasificador o generador de texto versus chatGPT). Las conclusiones e interpretación de los resultados comparando técnicas y/o modelos es muy importante durante todo el trabajo para valorar la adquisición de conocimientos por parte del alumnado.

A continuación, se incluyen algunas ideas del pipeline para la elaboración de la las prácticas según el tipo de dato:

TEXTO

Análisis de Datos Exploratorio (EDA), ejemplos para visualizar y percibir el tipo de problema que tenemos entre manos:

- Distribución de datos, balanceo de clases.
- Procesamiento de texto: stopwords, lematización.
- Representación de texto: TF-IDF, word frequency, bag of words, word embeddings.
- Análisis avanzado: clustering, topic modeling.

Machine Learning

- Clasificación de texto por temáticas, autores, estilos, épocas.

Deep Learning

- Text classification: Clasificación por temáticas, autores, estilos, épocas. Comparación entre modelos *from scratch* y pre-entrenados. Curvas de accuracy y loss, evaluación de overfitting.
- Question Answering: Responder preguntas sobre el contenido de las canciones de un autor, época o tema.
- Summarization: Generación de resúmenes por autor, temática o estilo.
- Text generation: Creación de nuevas canciones basadas en autor, temática o estilo, o combinando características.
- Opcional: Conversión de texto a voz (*text-to-speech*).

- Comparación de rendimiento con modelos profesionales (GPTs) como ChatGPT.

IMAGEN

Análisis de Datos Exploratorio (EDA) para asegurarnos de que técnicas pueden ser más apropiadas:

- Tipo de imágenes, tamaños, balanceo de clases.
- Histogramas de color, distribución de píxeles.

Machine Learning

- Pequeña prueba con ML clásico (*feature extraction* puede ser con Deep Learning).

Deep Learning

- Image classification: Clasificación de imágenes en categorías con modelos from *scratch* y modelos pre entrenados. Curvas de accuracy y loss, evaluación de overfitting.
- Object detection: Detección de objetos en imágenes.
- Image-to-Image: Generación de nuevas imágenes por categoría/clasificación.
- Opcional:
    - Generación de descripciones de imágenes (*Image Captioning*).
    - Conversión de imágenes a texto (*Image-to-Text*).

AUDIO

Análisis de Datos Exploratorio (EDA), para saber si tengo que normalizar, hacer data augmentation, etc.

- Cantidad de audios, balanceo de clases, sample rate.
- Visualización de espectrogramas.

Deep Learning

- Audio classification: Clasificación de audios por temáticas, autores, estilos, épocas.
- Opcional:
    - Generación de nuevas canciones (*como en la práctica 6*).

- Análisis de emociones en audio (*felicidad, tristeza, etc., similar a la práctica de Spotify*).

📌 Importante:

De cara a la evaluación se tendrá en cuenta que se haya seguido el pipeline de Data Science (EDA → Feature Engineering → ML clásico → DL generativo y discriminativo (from scratch to transfer learning)→ Comparación con modelos comerciales)

- Hay que pintar SIEMPRE curvas de train y validation para accuracy y loss, si no es difícil analizar el overfitting. También se suele mirar confusion matrix, sensitivity, specificity, etc.
- Como vimos en clase, en los modelos from scratch para clasificación (CNN or RNN) se intenta mejorar accuracy jugando con cantidad de parámetros, learning rate, drop out, data augmentation, etc. para intentar mejorar la accuracy hasta que se llega a transfer learning, pero siempre vamos paso a poso desde scratch.
- Las conclusiones son muy importantes. Cuidado con las malas interpretaciones, por ejemplo: el mejor modelo es el que tiene más accuracy, pero y qué pasa con el overfitting, ¿generaliza bien? ¿desplegarias este modelo en producción? Tenéis que saber cómo evaluar diferentes modelos y mejor la accuracy.

A nivel herramientas o entornos podéis utilizar lo que más os guste u os sintáis más cómodos. Comparto unas meras sugerencias para quien lo necesite, pero es totalmente libre elección:

- Texto: SpaCy, NLTK, Hugging Face.
- Imágenes: OpenCV, TensorFlow/PyTorch.
- Audio: librosa, torchaudio.

De manera ilustrativa, también os comparto un ejemplo de entregas de otros años para que os hagáis una idea de lo que es un buen trabajo: bien organizado, explicado, con todos los pasos del pipeline (desde el exploratorio, modelos from scratch de ML y DL, modelos preentrenados para feature extaction y fine-tuning...), gráficos, conclusiones.

Entrega de las prácticas

Se habilitará un espacio en Moodle para las entregas de las prácticas con una fecha límite (ver fechas en Fechas de evaluación). Sin embargo, es común que por cuestiones de tamaño el trabajo no pueda cargarse a Moodle. Por tanto, en caso de problemas con la subida, os recomiendo que me paséis directamente un **readme con el enlace al repositorio de código siempre indicando los miembros del equipo por email**. Toda aquella entrega tardía tanto en Moodle como vía email se verá reflejada con un impacto negativo en la nota de los trabajos.

Se insiste concienzudamente en que vuestro trabajo siempre se suba al repositorio de código con el fin de que os acostumbreis a poblarlo. En el ámbito profesional, forma parte de vuestra carta de presentación en la evaluación de candidaturas a un puesto laboral.

Fechas de evaluación 2025

- Examen texto: 6 marzo
- Examen imagen/audio : 24 abril
- Entrega texto: 13 de abril (primer finde de semana santa)
- Entrega imagen: 8 mayo
- *Hackathon* audio: 24 abril