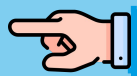
Preprocesamiento de Datos

Clase Sincrónica

{desafío} latam_

Implementar modelos de aprendizaje automático por medio de técnicas estadísticas, adecuando los diferentes algoritmos debidamente a la situación y requerimientos necesarios

- Unidad 1: Introducción al Machine Learning
- Unidad 2: Aprendizaje Supervisado y No Supervisado
  (Parte I: No supervisado)
  (Parte II: Clasificación)
  (Parte III: Clasificación)
  (Parte IV: Regresión)
  (Parte V: Series de tiempo)
- Unidad 3: Aplicando lo aprendido
  (Parte I: Preprocesamiento de datos)
  (Parte II: Modelamiento)



Te encuentras aquí

{desafío} latam_

¿Qué aprenderás en esta sesión?

En esta sesión aprenderás se hará un repaso general sobre el proceso de un proyecto de data science, aplicando todo lo aprendido en la sesión y poniendo foco en el preprocesamiento de los datos.

{desafío} latam_

¿Qué es una serie de tiempo? ¿Cuáles son sus componentes?

{desafío} latam_

¿En qué consiste un modelo ARIMA? ¿Cómo se determinan los parámetros?

{desafío} latam_

¿Cómo evaluamos un modelo de serie de tiempo?

{desafío} latam_

/* Metodología de Proyectos de Data Science */

{desafío}
latam_

Metodología de proyectos Data Science

CRISP DM

¿Por qué utilizar un marco de trabajo como CRISP DM?

1. Estructura Clara
2. Orientación al Negocio
3. Flexibilidad

{desafío} latam_

Entendimiento del Negocio

CRISP DM

Definición de los objetivos	Estudio del Contexto	Levantamiento de Recursos	Definición de éxito	Planificación inicial
Se establece el problema de negocio. Se definen los objetivos que se van a abarcar con ayuda del proyecto.	Investigación de la industria y procesos involucrados. Identificar las partes interesadas y consultarles sobre su opinión.	Se evalúa los recursos necesarios para el proyecto como personas, datos, hardware, etc.	Se establecen criterios claros para evaluar el proyecto, Se considera el impacto financiero y estratégico del proyecto.	Se crea un plan de alto nivel, describiendo las principales características. Se definen las métricas de evaluación a lo largo del proyecto.

/*Entendimiento y preparación de datos*/

Entendimiento de los datos

CRISP DM

Recopilación de datos	Calidad de datos	Exploración de datos	Muestreo de datos
Identificar las diferentes fuentes de datos. Adquisición de datos	Identificar valores faltantes, valores duplicados u otros problemas de datos. Identificar outliers. Concluir si la calidad de los datos permite modelar y primer filtro de variables.	Análisis univariado y multivariado. Análisis entre variables y correlaciones. Definir puntos importantes para la siguiente etapa y se itera.	En caso de ser necesario se puede muestrear los datos para trabajarlos de mejor forma. Asegurar que la muestra sea significativa.

Identificando Outliers

Métodos

1. **Método IQR (Inter Quartil Range)**: consideramos solo los datos entre Q1 - 1.5*IQR y Q3 + 1.5*IQR, donde IQR = Q3 - Q1

2. **Z - Score:** calcula la puntuación Z para cada punto de datos, que determina a cuántas desviaciones estándar se encuentra un valor del promedio (se suele considerar outlier un valor por sobre 2 o 3 desviaciones estándar

3. **Métodos basados en ML:** hay algoritmos como **dbscan** que detectan outliers, o KNN, que buscan valores con una distancia significativa que se consideran outliers.

4. **Otras:** algoritmos de detección de anomalías como **svm one class** o **isolation forest.**

Correlaciones

Métodos de cálculo y detección

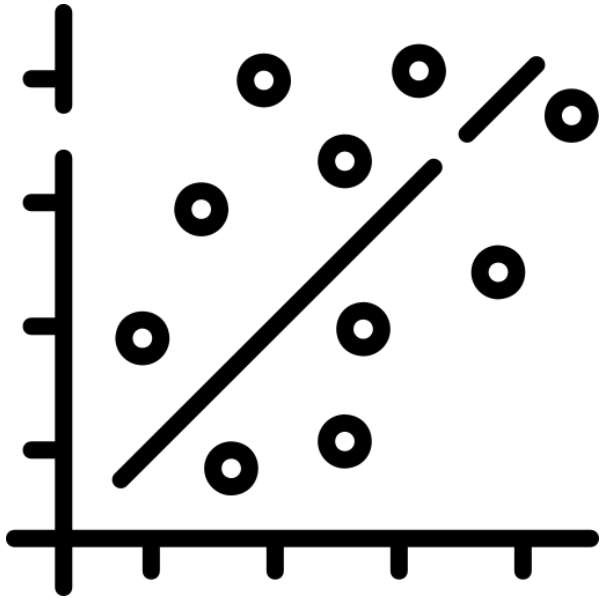
1. **Pearson:** mide la relación lineal entre variables numéricas

$$r = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

2. **Chi-Cuadrado:** mide la relación entre variables categóricas

3. **Test de Anova:** calcula si hay una diferencia en la variable numérica entre grupos de una categórica

4. **Otras:** kendall, spearman, v de cramer, test kolgomorov-smirnov, etc.



Coeficiente de Chi cuadrado

Categórica vs categórica

Prueba estadística que determinar si existe una asociación significativa entre dos variables categóricas.

1. Se calcula la tabla de contingencia.
2. Se calcula el coeficiente en base a la tabla de contingencia.
3. Se calcula los grados de libertad y se observa el p-valor con una tabla.

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Coeficiente de Chi cuadrado

Categórica vs categórica

Frecuencias observadas		
	Acción	Romance
F	5	15
M	12	8

Frecuencias esperadas		
	Acción	Romance
F	8.5	11.5
M	8.5	11.5

Tabla chi-cuadrado		
	Acción	Romance
F	$(5 - 8.5)^2/8.5 = 1.441176$	$(8 - 11.5)^2/11.5 = 1,065217391$
M	$(12 - 8.5)^2/8.5 = 1.441176$	$(8 - 11.5)^2/11.5 = 1,065217391$

$$\chi^2 \approx 5,01$$

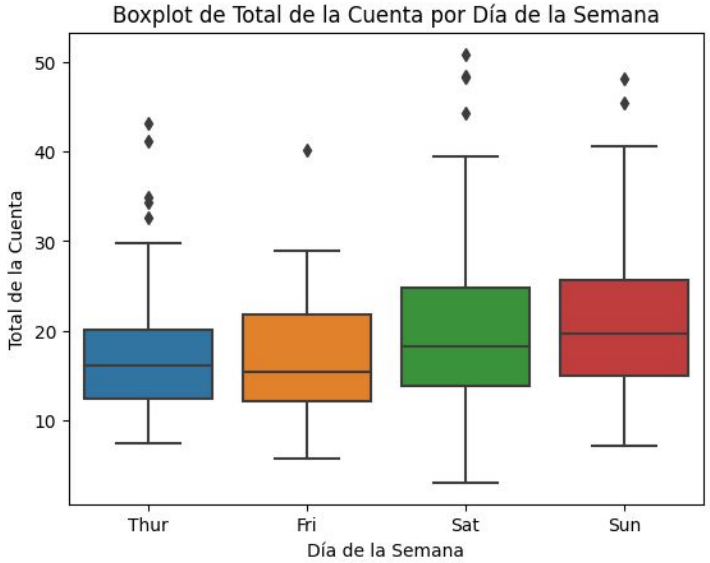
Test de ANOVA

Numérica vs categórica

Es un test estadístico para ver si hay diferencia significativa entre diferentes grupos:

H0: No hay diferencia significativa
H1: Existe al menos un grupo diferente a los demás

No dice que grupo es diferente, pero con ayuda de un análisis posterior se puede observar.



{desafío} latam_

¡Manos a la obra! Entendimiento y preparación de datos

Manos a la obra

Entendimiento y preparación de datos

Veremos ahora cómo aplicar estos procedimientos con la ayuda de Python. Para ello, sigue los pasos que te presentará tu profesor en el archivo Jupyter Notebook.

En esta ocasión veremos:

- Detección de outliers
- Correlaciones

/* Dimensionalidad y características*/

Preparación de datos

CRISP DM

Limpieza de datos	Transformación de datos	Feature Engineering	Selección de características
En base a lo aprendido se eliminan duplicados, se eliminan las variables que no sirven, se decide cómo tratar los valores nulos y outliers. Se aplican los filtros necesarios a los datos.	Normalización. Codificación de variables categóricas. Aplicación de transformaciones matemáticas a características (logaritmos, raíces cuadradas, etc.) para reducir la asimetría.	Generación de nuevas características a partir de las existentes que puedan ser más informativas para el modelo. Extracción de características relevantes de los datos originales.	Eliminar las características irrelevantes. Mantener las más importantes. Considerar el problema de dimensionalidad.

Dimensionalidad

La maldición de la dimensionalidad

La **maldición de la dimensionalidad** es un término utilizado en estadísticas y aprendizaje automático para describir los desafíos y problemas que surgen cuando trabajamos con conjuntos de datos de alta dimensionalidad, es decir, conjuntos de datos que tienen un gran número de características o variables en comparación con el número de observaciones.

Algunos de los efectos de la maldición de la dimensionalidad son los siguientes:

- **Espacio de características disperso**
- **Requerimientos computacionales**
- **Sobreajuste**

Selección de características

Métodos de selección

- **Backward Selection:** se comienza con todas las características y se eliminan iterativamente con algún criterio de evaluación.
- **Forward Selection:** se comienza con un conjunto vacío y se agregan características de forma iterativa.
- **Métodos de filtro:** se evalúan características filtrando por algún criterio de correlación.
- **Otras:** Lasso, Random Forest

¡Manos a la obra! Entendimiento y preparación de datos

Manos a la obra

Entendimiento y preparación de datos

Veremos ahora cómo aplicar estos procedimientos con la ayuda de Python. Para ello, sigue los pasos que te presentará tu profesor en el archivo Jupyter Notebook.

En esta ocasión veremos:

- Forward selection
- Filtro por correlaciones
- Selección por Lasso

Desafío - Preprocesamiento de datos

{desafío} latam_

Desafío

"Preprocesamiento de datos"

- Descarga el archivo "Preprocesamiento de datos".
- Tiempo de desarrollo asincrónico: desde 4 horas.
- Tipo de desafío: individual.

¡AHORA TE TOCA A TI! 💪

{desafío} latam_

Ideas fuerza

Las ***metodologías*** *son útiles para* ***estandarizar y utilizar las mejores*** *prácticas al momento de trabajar.*

Nos centramos en la metodología CRISP-DM principalmente en los primeros pasos de entendimiento del negocio, de la data y preparación de los datos .

Para cada etapa hay varios ***estadísticos y métodos*** *que son útiles, pero lo importante es analizar caso a caso que es lo* ***mejor para mi problema.***

¿Qué conceptos no te quedaron claros o quieres reforzar?

Recursos asincrónicos

¡No olvides revisarlos!

Para esta semana deberás revisar:

- Guía de estudio
- Desafío “Preprocesamiento de datos”

{desafío} latam_

Próxima sesión...

Continuaremos poniendo en práctica los conocimientos aprendidos y reforzando la etapa de modelamiento

{desafío} latam_

{desafío}
latam_

Academia de
talentos digitales