



CC58 Tópicos en Ciencia de la Computación  
Carrera de Ciencias de la Computación  
Tiempo Límite: 110 Minutos

Examen Parcial  
Mayo, 2021  
MSc. Pedro Shiguihara

- **Objetivo.** Aplicar las técnicas de estimación de factores e inferencia para realizar comparaciones en modelos de grafos probabilísticos.

- **Lenguaje de programación.** El lenguaje a utilizar es Python en el entorno Google Colab. En caso fuese necesario, adjuntar: imágenes, videos o texto en Google Colab para documentar y reforzar tu respuesta.

Problema	Puntos	Resultado
1	6	
2	14	
Total:	20	

- **Entrega de la evidencia.** Cada una de sus respuestas debe estar respaldada por el código fuente respectivo. Utilizar el lenguaje *markdown* dentro de Google Colab para documentar correctamente la respuesta a cada pregunta. Trata de que **tu programa sea genérico y útil para cualquier dataset**. El alumno debe enviar como evidencia el archivo en extensión **ipynb** dentro de un archivo comprimido (ZIP) y además el link compartido con permisos de edición comentado en la primera línea del archivo **ipynb** a fin de poder visualizarlo también *online*.
- **Envío es por el Aula Virtual en Examen Parcial.** En caso no se cumpla, se califica con CERO. El archivo debe ser en formato ZIP: **EA\_codigoalumno.zip**. Similitud con más del 30 % en código con otros grupos, conlleva a CERO de nota para todos.
- **Respuestas misteriosas o sin justificación no recibirán puntaje.** No se aceptan soluciones que no especifiquen el proceso para conseguir dicha respuesta.

# 1. Descripción

El análisis guiado por datos es muy importante ya que abre paso a un análisis exhaustivo del dataset. Procesar y explorar profundamente un dataset permite llegar a conclusiones fehacientes que superan a las suposiciones y a reducir el bias. El dataset puede tener valores incompletos. Tanto las variables como sus valores se manejan como números.

Tu programa debe tener un botón **ipywidget** para cargar **cualquier** archivo TXT con el siguiente formato:

- Número de variables
- Cardinalidad de cada variable separadas por comas
- Número de aristas del grafo a ingresar
- Siguen las aristas, separando cada par de variables por:  $\rightarrow$
- Hiperparámetro de dirichlet  $\alpha$  (es flotante)
- Número de instancias en el dataset
- Finalmente, se describe la data con cada valor separado por comas

Se asume que las variables son descritas desde 0 a  $N - 1$ , donde  $N$  es el número de variables. De una manera similar, también ocurre con las cardinalidades.

Ejemplo:

```
3
2,2,3
2
0->1
1->2
1.0
10
0,0,0
0,0,1
0,1,0
1,0,0
0,1,2
0,1,2
1,1,0
0,1,0
1,0,1
1,1,1
```

## 2. Construcción Automática de Factores

1. (6 puntos) A partir del dataset descrito, tu programa debe estimar:
  - Todos los factores estimados correspondientes a partir del DAG brindado en la data, siguiendo la propiedad de Markov.
    - Cada factor debe ser descrito en una tabla en ipywidget de una manera similar a los factores de la TB1.

## 3. Inferencia

2. (14 puntos) Dada un conjunto de valores, tu programa debe recuperar sus probabilidades correspondientes y reportar el producto de las mismas.
  - El conjunto de valores corresponde a cada una de las variables presentes en el dataset y debe ser ingresado por cajas de textos (input text) de ipywidget.
  - Debe existir un botón para mostrar las probabilidades recuperadas de los factores generados a partir del DAG que hagan match con los valores ingresados por el usuario.
  - Debe existir un segundo botón para mostrar la multiplicación de tales probabilidades recuperadas.

Lima, mayo del 2021.