Trabajo Práctico Parte 2: Modelado Bayesiano y aplicaciones

Estadística e Inferencia II

Objetivo general:

En esta segunda parte aplicaremos los distintos modelos vistos para el tratamiento de conjuntos de datos, enfocándonos en la interpretación y en la elección del modelo más adecuado para realizar predicciones. El objetivo es profundizar en el flujo de trabajo bayesiano, extendido ahora a modelos más complejos.

Cada grupo deberá elegir uno o más conjuntos de datos (pueden utilizar alguno de los analizados en los módulos de la materia o seleccionar otros de fuentes públicas) y desarrollar un análisis bayesiano completo: formular hipótesis, ajustar modelos y comparar resultados.

Nota: recuerden incluir en la entrega el archivo con el código en Python y el conjunto de datos utilizado (o, en su defecto, indicar claramente la fuente de donde fue descargado).

Finalmente, como requisito obligatorio, deberán presentar un **informe integrador** en el que comenten los resultados obtenidos tanto en el TP Parte 1 como en el TP Parte 2. Los puntos clave para desarrollar dicho informe se encuentran al final de este documento.

Parte 0: EDA

Presenten brevemente el conjunto de datos seleccionado, indicando qué variables van a utilizar en el análisis y mostrando un resumen descriptivo (*summary*) de cada una. Si lo desean, pueden usar un dataset diferente en cada parte del TP. En todos los casos, incluyan al menos un *summary* inicial que sirva para conocer mejor los datos antes de modelar.

Parte 1: Modelos Lineales Generalizados (GLMs)

1. para el Dataset elegido, elegir variables para aplicar un modelo GLM Bayesiano (pueden considerar una variable independiente y una variable respuesta, o trabajar con más de una para cada categoría) y presentar un brvee resumen de cada una y la relación entre variables (si ya lo hicieron en la Parte 0, no es necesario repetirlo).
2. Seleccionar un modelo GLM adecuado al tipo de variable respuesta:
   - Continua: regresión lineal o con varianza variable.
   - Binaria: regresión logística.
   - Modelo con estructura jerárquica
   - (Opcional) De conteo: regresión Poisson o negativa binomial.
3. Formular las ecuaciones del modelo y justificar las distribuciones elegidas.
4. Interpretar los parámetros de la distribución posterior: medias, intervalos de credibilidad, relación con los predictores.
5. Realizar predicciones a posteriori e interpretarlas en el contexto del problema.
6. Si el modelo lo permite, incluir predictores categóricos, interacciones o estructura jerárquica.

(Opcional)

- Implementar el modelo con `Bambi` y comparar los resultados con una versión escrita en `PyMC`.
- Explorar selección de variables mediante inferencia predictiva por proyección.
- Analizar la varianza residual o la heterocedasticidad mediante un modelo de varianza variable.

Parte 2: Modelos de Mezcla

1. Plantear un modelo de mezcla finita (por ejemplo, mezcla de dos o más normales) y justificar la elección.
2. Analizar la identificabilidad del modelo y discutir posibles problemas de simetría entre componentes.

3. Comparar distintos valores de K (número de componentes) y evaluar cuál describe mejor los datos.

4. Representar gráficamente los componentes y la mezcla ajustada.

5. Discutir los resultados y su interpretación en el contexto del problema.

(Opcional)

- Ajustar un modelo con exceso de ceros (Zero-Inflated o Hurdle).
- Explorar la idea de un modelo de mezcla no finita, mencionando el proceso de Dirichlet.

Parte 3: Procesos Gaussianos (GPs) - (Opcional)

Realizar una pequeña exploración utilizando un proceso gaussiano para modelar una relación continua simple. Pueden emplear un ejemplo de demostración o un subconjunto reducido del dataset elegido, visualizando cómo el modelo captura la tendencia y la incertidumbre de las predicciones.

Indicaciones para el Informe Integrador

Como cierre del Trabajo Práctico Integrador (Partes 1 y 2), cada grupo deberá presentar un informe final que no exceda las **10 páginas**.

El informe debe estar redactado en formato claro y coherente, con foco en la interpretación de resultados y no en el código. No se debe incluir código Python en el cuerpo del documento (solo las salidas o gráficos relevantes para la discusión).

Estructura sugerida

1. **Introducción:** breve presentación de los objetivos generales del trabajo y la relación entre las Partes 1 y 2.

2. **Descripción de los conjuntos de datos:** indicar qué dataset o datasets se utilizaron en cada ejercicio, las variables seleccionadas y un breve resumen descriptivo de cada una.

3. **Modelos y análisis:** para cada parte del TP, describir de forma breve y concisa el modelo ajustado o el estudio realizado. Explicar las decisiones tomadas (por ejemplo, elección de priors, hiperpriors, familia de distribución o estructura jerárquica) y los principales resultados obtenidos.

4. **Resultados y conclusiones:** presentar las inferencias más relevantes, las comparaciones entre modelos si es posible y la interpretación de los parámetros desde una perspectiva bayesiana. Incluir las conclusiones generales del trabajo y posibles líneas de mejora o extensión.

5. **Referencias y fuentes:** si se utilizó bibliografía, documentación externa o elementos no vistos en clase (funciones, modelos, kernels, etc.), deben mencionarse claramente al final del informe.

El objetivo del informe es integrar los conocimientos teóricos y prácticos del curso, comunicando el proceso de modelado y las conclusiones de manera clara, ordenada y reflexiva para luego hacer una puesta en común.