IMT2220, Cálculo para ciencia de datos, 2023-2
Tarea 4

Fecha entrega: 16 de Noviembre de 2023

Instrucciones

Pueden discutir los problemas con sus compañeros. Sin embargo, sus entregas, tanto la parte escrita como los códigos, deben ser individuales. **No está permitido** copiar las respuestas de alguien más ni dejar que otros copien sus respuestas. El hacerlo se reflejará con la nota mínima en la evaluación (1.0).
Su entrega debe estar compuesta por un único archivo .pdf **escrito en** LaTeX que incluya todas las respuestas y todos los gráficos que se están pidiendo, junto con un archivo .zip que incluya todos los códigos que produjeron y utilizaron durante esta tarea. Estos dos archivos deben ser subidos en Canvas. Si hay más de una línea en un gráfico, utilice distintos estilos de línea o distintos colores para diferenciarlas e incluya leyendas. No se olvide de nombrar todos sus ejes y líneas en las leyendas. **Cada gráfico debe ser comentado** (un gráfico sin un análisis del mismo no es una respuesta apropiada).

Integrales Impropias

Hasta ahora, siempre hemos trabajado con integrales en regiones finitas bien definidas (piense rectángulos, círculos, triángulos, etc.), pero ocurre que a veces es necesario integrar funciones en dominios no acotados, o bien, computar integrales de funciones que se indefinen en un dominio finito. De aquí nacen las integrales impropias. Algunos ejemplos de esto son:

$$\int_1^\infty \frac{1}{x^2}\,dx, \quad \int_0^1 \frac{1}{x^{1/2}}\,dx$$

La pregunta en cuestión ahora yace en: ¿Cómo las calculamos? La idea es simple: si tenemos un dominio no acotado, reemplazar el límite no acotado por una costante y hacerla tender a infinito (positivo o negativo). Es decir:

$$\int_1^\infty \frac{1}{x^2}\,dx = \lim_{b\to\infty} \int_1^b \frac{1}{x^2}\,dx = \lim_{b\to\infty} -\frac{1}{x}\bigg|_1^b = 1 - \lim_{b\to\infty} \frac{1}{b} = 1$$

Por otro lado, para dominios finitos, solo basta con calcular la integral y evaluar para determinar si esta converge o no a algún valor en particular:

$$\int_0^1 \frac{1}{x^{1/2}}\,dx = 2\sqrt{x}\bigg|_{x=0}^1 = 2$$

Este tipo de integrales es de especial utilidad en probabilidades, donde usualmente encontramos funciones continuas que integrar en dominios no acotados. Algunos criterios conocidos para determinar si una integral converge o no corresponden a los siguientes:

$$\int_1^\infty \frac{1}{x^p}\,dx < \infty \text{ si } p > 1$$
$$\int_0^1 \frac{1}{x^p}\,dx < \infty \text{ si } p < 1$$
$$\int f(x)\,dx < \infty \text{ si existe una función } g \text{ tal que } f(x) \leq g(x) \text{ y } \int g(x)\,dx < \infty$$

donde usamos $\int f < \infty$ para decir que la integral en cuestión converge. Para más detalles al respecto, refierase al capítulo 7.8 del libro de Stewart, Cálculo en una variable.

Problemas

1. (10 puntos) Verifique que el volumen encerrado por la superficie de revolución generada al rotar la función $f(x) = 1/x$ alrededor del eje X para $x \in [1, \infty)$ es finito, mientras que el área de esta superficie es infinita.

2. (10 puntos) Determine los valores λ tales que la integral

$$\int\int_D \frac{1}{(x^2+y^2)^\lambda}\,dA$$

es convergente, cuando D es el disco unitario en $\mathbb{R}^2$.

3. (10 puntos) Un cuerpo esférico de radio 5 tiene densidad de masa

$$\rho(x,y,z) = 1 - \frac{x^2+y^2+z^2}{100}$$

donde (x, y, z) son los puntos medidos de forma que el centro del cuerpo esférico es el origen. Calcule el centro de masa de este cuerpo.

4. (15 puntos) Calcule la integral:

$$I = \int_{-\infty}^{\infty} e^{-x^2}\,dx$$

Hint: Exprese I^2 como una integral doble y realize un cambio de variables apropiado.

5. En este problema estudiaremos el problema de encontrar clusters en los datos. Dado un set de datos **no etiquetados** $\{(x_i, y_i)\}_{i=1,\ldots,m} \subset \mathbb{R}^2$, queremos agruparlos de forma de poder identificar los distintos grupos (o cluters) dentro de si mismo. Una forma de encontrar dichos clusters de información es mediante algoritmos como K-means, o KNN. Sin embargo, también existen métodos en base a probabilidades, que resuelven este problema usando el algoritmo de maximización de esperanza.

La esperanza de una variable aleatoria X se define como:

$$\mathbb{E}[X] = \int_{\Omega} (x \cdot f_X(x))\, dx$$

donde $f_X(x)$ es la función densidad de X y Ω es el espacio muestral. Por ejemplo, si X fuera una variable normal con media μ y varianza σ^2 ($X \sim \mathcal{N}(\mu, \sigma^2)$), entonces $\Omega = \mathbb{R}$ y:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

La idea general de este método iterativo consiste en dos pasos: Calcular la esperanza de un modelo probabilítico en base a los parámetros estimados en una iteración anterior, y maximizar los parámetros de este modelo una vez sabida la esperanza del modelo mixto previsto (más detalles aquí). Los modelos mixtos en cuestión se ven como sigue:

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k p_k(\boldsymbol{x})$$

donde p_k corresponde a una función densidad de una familia de distribuciones asignado al cluster k, K es el total de clusters a estimar y π_k es el peso asignado a dicho cluster. Así, se debe cumplir:

$$\sum_{k=1}^{K} \pi_k = 1, \quad 0 \leq \pi_k \leq 1$$

El cómo se hace para aprender los distintos parámetros de las distribuciones p_k a partir de los datos provistos no es un tema de estudio ahora, y puede leer sobre ello en el siguiente post.

El problema en cuestión es el siguiente. Considere el set de datos `data_HW4.csv` y:

(a) (6 puntos) Entrene un modelo Gaussiano mixto con 2, 3 y 4 clusters con todo el set de datos. Para esto, revise el siguiente link a scikit learn. Explique cual de estos 3 modelos le hace más sentido viendo el scatter plot de los datos.

(b) (9 puntos) Grafique las curvas de nivel asociadas a la log-verosimilitud negativa (el score con signo negativo) para cada uno de los casos anteriores (ejemplo). Explique porqué estos son distintos a pesar de tener los mismos datos. Use en su respuesta la información disponible sobre el modelo entrenado.

Si tiene problemas usando localmente sklearn (por temas de compatibilidad entre numpy y sklearn, existe la posibilidad no nula de que no pueda importar el modelo pedido), ocupe Google Colab.

La razón de porqué se grafica la log-verosimilitud negativa es porque para maximizar la función de score, el algoritmo minimiza esta con el signo contrario. La log-verosimilitud es la función en cuestión.