Tarea 3

Profesor: Felipe Tobar
Auxiliares: Mauricio Araneda, Alejandro Cuevas y Mauricio Romero
Consultas: Todos el cuerpo docente.
Fecha entrega: 30/05/2019

Formato entrega: Entregue un informe en formato PDF con una extensión de a lo más **3** páginas presentando y analizando sus resultados, detalle la metodología utilizada y adicionalmente debe entregar un jupyter notebook con los códigos que creó para resolver la tarea.

P1. Regresión logística, clasificación y Metropolis-Hastings

Genere datos con forma de medialuna a partir de la librería `sklearn` utilizando el módulo `sklearn.datasets.make_moons` (vea la documentación para más detalle). Genere $N = 1000$ datos etiquetados y cree un conjunto de entrenamiento y de prueba con $N_{test}/N = 0.2$. Utilice $\sigma_{\text{moons}} = 0.2$ como desviación estándar del ruido gaussiano añadido a las muestras. El objetivo de este inciso es implementar un clasificador lineal utilizando como modelo una regresión logística, y tanto máximo a posteriori como Markov chain Monte Carlo MCMC para estimar sus parámetros.

a) (1.5 Puntos) Muestre en 2D con un color las muestras de las clase 0 y otro distinto las muestras de la clase 1 en base a las etiquetas disponibles. En el mismo gráfico muestre el conjunto de entrenamiento y prueba. Luego, implemente un modelo de regresión logística para clasificar los datos disponibles, para esto construya la función de log-posterior y estime los parámetros de la regresión logística utilizando L-BFGS asumiendo el prior que estime conveniente: o bien Gaussiano de matriz de covarianza a su elección (no necesariamente diagonal) o plano (Uniforme sobre un intervalo). Sobre las muestras presentadas, grafique la recta que separa ambas clases. ¿Pudo clasificar correctamente todas las muestras? Justifique utilizando métricas de clasificación.

   (**Nota:** No puede utilizar la implementación de sklearn de regresión logística.)

b) (1.5 Puntos) Implemente el algoritmo de Metropolis-Hastings MH para estimar la densidad *a posteriori* del modelo encontrado en a), donde la cadena de Markov puede inicializarse con los parámetros encontrados en esa parte, usando el mismo prior que en la parte anterior.

   Genere del orden de 1000 muestras para los parámetros y grafique algunas de las rectas que corresponden a estos parámetros. Discuta su elección del prior sobre los parámetros y muestre cuán sensible son sus soluciones para distintos *proposal*. Utilice el conjunto de entrenamiento para obtener las muestras de MCMC.

c) (1 Punto) Con sus muestras generadas, aproxime la distribución posterior de cada parámetro usando histogramas y muéstrelos ¿Cómo se compara la moda de cada parámetro con lo obtenido con máximo a posteriori? discuta.

d) (1 Punto) Con sus muestras generadas, realice predicciones usando un modelo de regresión logística Bayesiana, para esto marginalice sobre los parámetros de su regresión logística, en este caso su predicción será de la forma,

$$p(y_*|x_*, X, Y) = \int p(y_*|x_*, w)p(w|X, Y)dw$$

   Donde X, Y es su dataset de entrenamiento, x_*, y_* son los puntos de test y clases de test respectivamente, y w son los parámetros de su regresión logística. Aproxime dicha integral usando integración de Monte Carlo.

P2. Proyecto curso

Esta parte debe ser respondida por cada integrante del grupo, donde debe indicar los demas integrantes del grupo.

a) (0.5 Puntos) Encasille su proyecto en algún tipo de aprendizaje (clasificación, regresión, clustering, aprendizaje reforzado, etc...). Identifique según el caso el rol de sus variables, por ejemplo en el caso de clasificación cuales son sus características y las clases o en el caso de regresión la variable a estimar y los inputs.

b) (0.5 Puntos) Proponga al menos 2 métodos para resolver su problema, fundamentando brevemente su elección.