Tarea II: Temas Selectos de Estadística

Por entregar el 6 de abril;

1. **Este ejercicio es sobre el algoritmo backward-forward para cadenas ocultas de Markov.**
   Para resolver el problema de visitantes perdidos en la hermosa ciudad de Otagato se implementó un sistema de localización basado en sensores.

   En intervalos regulares ($t_i = t_0 + i\Delta$) una pulsera electrónica de cada visitante transmite una señal (que le identifica de manera única).
   Para simplificar el problema, vamos a suponer que en cada momento t_i el está siempre en uno de los 9 lugares indicados en el mapa.

   Por otro lado, hay tres receptores en la ciudad: cerca del Pipila (R1), en Tepetapa (R2) y por la Presa (R3). Desafortunadamente, un receptor no siempre recibe la señal transmitida. La siguiente tabla muestra la probabilidad de recibir la señal en función de donde está la persona.

lugar:	1	2	3	4	5	6	7	8	9
Prob. recibir señal por el Pipila:	0,1	0,4	0,9	0,9	0,9	0,9	0,9	0,6	0,1
Prob. recibir señal por Tepetapa:	0,9	0,9	0,9	0,5	0,5	0,3	0,3	0,3	0,9
Prob. recibir señal por la Presa:	0,1	0,1	0,5	0,5	0,5	0,9	0,9	0,9	0,1

   Vamos a suponer que la manera de transladarse en la ciudad sigue una cadena de Markov; la probabilidad de transición de que uno queda en el mismo estado (lugar) es 0,5; la probabilidad que uno se va a un estado *vecino* es

$$\frac{0{,}5}{\#\text{ de vecinos}}$$

donde la vecindad es definida por la linea negra en el mapa. Por ejemplo lugar 2 tiene un vecino; lugar 5 tiene 2 vecinos y lugar 3 4.

Supongamos que la señal recibida en 7 momentos consecutivos fue

$$(1,1,1); (1,1,1); (1,1,0); (1,1,1); (1,0,1); (1,1,1); (1,0,1),$$

donde en (x_1, x_2, x_3) x_i indica si se recibió en receptor R_i la señal (1:si; 0 : no) Si su caminata empezó en lugar 3: ¿ Cuál es la probabilidad que la personas se encuentra en Emergencia del IMSS (lugar 2) en momento t_7?
¿ Cuál es el lugar más probable donde el visitante se encuentra en momento t_7?

2. **Este ejercicio es sobre el algoritmo EM.**

   a) Supongamos que elegimos un punto al azar del área formada por una letra C. En la Figura 1 se ve una muestra de esta distribución.

   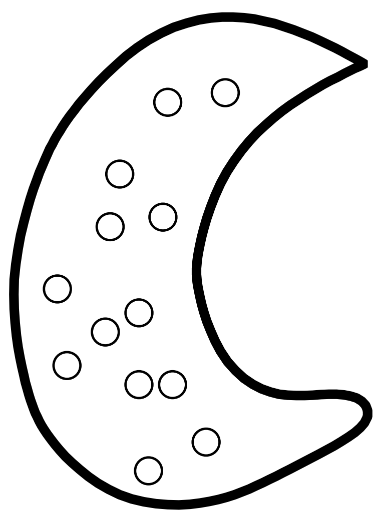

   Figura 1

   Usa el algoritmo EM para aproximar esta distribución con una mezcla de K distribuciones Gausianas usando una muestra de tamaño 100 (no está permitido usar la función `me` o `em` de R). Toma $K = 4, 8, 12$. Visualiza los resultados.

   b) En `datos.txt` se encuentran los datos de dígitos escritos a mano. Cada reglón contiene 9 mediciones: 8 posiciones $(= X)$ en la curva que se trazó al escribir el dígito; en la último posición se indica el valor numérico (=Y) del número que se dibujó. Nos limitamos a los digitos 3 y 4.

Supongamos que podemos representar cada distribución de $X|Y = i$ como una mezcla de Gausianas. Usa el algoritmo EM para ajustar estas mezclas a estos datos (busca un valor adecuado de K; transforma eventualmente tus datos).

Calcula después

$$\arg\max_i P(Y = i|X = x)$$

como el clasificador Bayesiano óptimo.

3. **Este ejercicio es sobre el algoritmo EM.**
Considera la figura 2 que representa una situación que ocurre comúnmente en tomografía SPECT y PET; en un intervalo dado, partículas son emetidas desde posiciones $i, 1 \leq i \leq n$ y algunas son detectadas en los sensores $j, 1 \leq j \leq m$. A través del conteo del número de particulas detectadas en los sensores se quiere estimar la intensidad de la emisión de partículas en cada posición i.

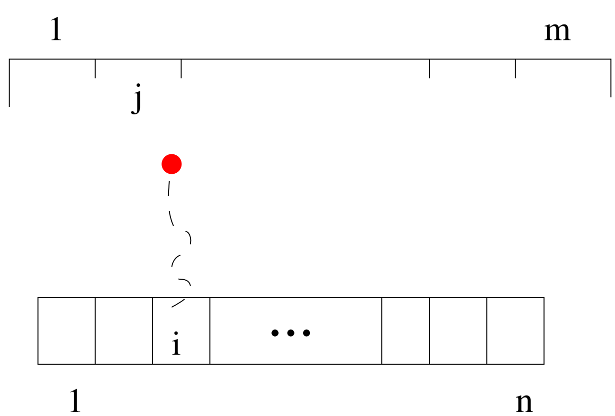

Figura 2

Hacemos los siguientes supuestos:

a) El número de partículas emetidas en un intervalo fijo en posición i sigue una distribución Poisson con parámetro λ_i.
Recuérdense:

$$Z \sim Pois(\lambda) \text{ ssi } P(Z = k) = \exp(-\lambda)\frac{\lambda^k}{k!}.$$

b) La probabilidad de que una partícula proveniente de i es detectada por sensor j es p_{ij}. Vamos a suponer que $\{p_{ij}\}$ son valores conocidos.
Llamamos N_{ij} el número de partículas de i detectadas por j. Suponiendo independencia se puede mostrar que lo anterior implica que

$$N_{ij} \sim Pois(p_{ij}\lambda_i).$$

Se supone que $\{N_{i,j}\}$ son independientes.

c) Llama Y_j lo que se observa en sensor j; por construcción

$$Y_j = \sum_i N_{ij}. \tag{1}$$

El problema es estimar $\{\lambda_i\}$ en base de $\{Y_j\}$ usando el algoritmo EM. Define $X = (Y, N)$ como el vector con los datos *completos*; Y el vector con todos los Y_j observados y N el vector con todos los N_{ij} (no observados).

a) Verifica que bajo (1) la verosimilitud de X es

$$\sum_i \sum_j \{-\lambda_i p_{ij} + n_{ij} \log(\lambda_i p_{ij}) - \log(n_{ij}!)\}$$

b) Verifica que

$$E_\lambda(N_{i,j}|Y) = \frac{y_j \lambda_i p_{i,j}}{\sum_h \lambda_h p_{h,j}}$$

Usa el hecho que vimos en la parte de tablas de contingencias que si $\{Z_i\}$ son v.a. $Pois(\lambda_i)$ independientes, entonces

$$Z_i | \sum_j Z_j = n \sim Binom(\frac{\lambda_i}{\sum_j \lambda_j}, n)$$

c) Escribe en pseudo código el algoritmo EM correspondiente.

4. **Este ejercicio es sobre el algoritmo bootstrap.**
Usa el método del Bootstrap para obtener intervalos de confianza para el *log-odds ratio* en una tabla de contingencia de 2x2.