Tarea: Temas Selectos de Estadística

Por entregar el 13 de marzo;

1. **Este ejercicio es sobre la evaluación de diferentes algoritmos de clasificación binaria.**
Tenemos un conjunto de prueba y un conjunto de entrenamiento y dos algoritmos de clasificación C_1 y C_2 entrenado con el conjunto de entrenamiento. Define X_i como la variable que indica si algoritmo C_i clasifica un dato del conjunto de prueba como 0 o 1 y define Y_i como la variable que indica si algoritmo C_i lo clasifica correctamente o no. Lo anterior nos da dos tablas de contingencia (una basada en (X_1, X_2) y otra en (Y_1, Y_2)).

   a) Una manera para cuantificar si los clasificadores se comportan de manera similar es verificar si las distribuciones marginales de Y_i son iguales. Muestra que lo anterior es equivalente a verificar si $p_{0,1} = p_{1,0}$ con p las probabilidades subyacente a la tabla de contingencia de (Y_1, Y_2). Deriva el estimador de Máxima Verosimilitud para este hipótesis. Usa una estadística de prueba basada en razones de verosimilitud para calcular el valor de p bajo esta hipótesis para los siguientes datos:

	$Y_2 = 0$	$Y_2 = 1$
$Y_1 = 0$	8	7
$Y_1 = 1$	11	21

   b) Una disventaja del enfoque anterior es que no se toma en cuenta la variabilidad generada por el conjunto de entrenamiento. (*) ¿ Como *suavizar* este efecto ? Todo lo anterior fue basado en la tabla de contingencia usando las Y's.
   (*)¿Qué complica una prueba basada en los X's?

2. **Este ejercicio es sobre pruebas de hipótesis usando métodos computacionalmente intensivos.**
Hemos viste que si el tamaño de la muestra, n, es suficientemente grande, podemos suponer que nuestras estadísticas de distancia (*power-divergence statistics*) son aproximádamente chi-cuadrada. ¿Qué hacer si n es chica ? Un remedio es recurrir a pruebas exactas. Tomamos el caso de una tabla bidimensional $d = 2$.

   Si llamamos $N_{i,j}$ el número de observaciones en celda (i, j), hemos visto en la clase que si el muestreo sigue una distribución multinomial:

$$P(N_{i,j} = n_{i,j}, \forall i, j) = \frac{n!}{\prod_{i,j} n_{i,j}!} \prod_{i,j} p_{i,j}^{n_{i,j}}$$

   a) Verifica que bajo la hipótesis de independencia:

$$P(N_{i,j} = n_{i,j}, \forall i, j | N_{i,+} = n_{i,+}, N_{+,j} = n_{+,j}) = \frac{\prod_i n_{i,+}! \prod_j n_{+,j}!}{n! \prod_{i,j} n_{i,j}!}, \qquad (1)$$

   es decir ya no depende de las p's.

b) El método de prueba exacta de Fisher para una tabla de contingencia T consiste en primero elegir una medida de distancia, por ejemplo Pearson - chi-cuadrada y después sumar la probabilidad (según (1)) de todas las tablas con las misma marginales que T y que tienen en esta medida de distancia un valor igual o mayor (peor) que lo observado para T.
Escribe un algoritmo en C que implementa lo anterior para $d = 2$ y variables binarias. Compara el resultado con lo que uno obtiene usando una distribución de chi-cuadrada como aproximación.

3. Considera el siguiente conjunto de datos con el número de accidentes en Dinamarca en 1981 que involucran peatones, desglozado por día de la semana.

Día de la semana	Número de accidentes
lunes	279
martes	256
miércoles	230
jueves	304
viernes	330
sábado	210
domingo	130

a) Verifica si puedes apoyar la hipótesis (a 95 %) que el día de la semana no influye en el número de accidentes.
Verifica si puedes apoyar la hipótesis (a 95 %) que tanto durante la semana, el día no afecta el número de accidentes como durante el fin de semana el día no influye en el número de accidentes (pero entre semana y fin de semana puede existir una diferencia).

b) Usa el método delta para calcular un intervalo de confianza de 95 % para el índice de Gini de estos datos (no hay necesidad de deriver primero el caso general).

4. Considera la siguiente tabla de contingencia sobre el número de accidentes en Suecia en un lapso de 18 semanas en 1961 (sin límite de velocidad en las carreteras) y en un lapso de 18 semanas en 1962 (cuando se estableció un límite de velocidad).

límite de velocidad	en carreteras primarias	en carreteras secundarias
90 km/h	19	79
sin límite	102	175

a) (*) ¿ Puedes decir que es más peligroso viajar en carretaras secundarias que en carreteras primarias ?

b) Calcula un intervalo de confianza de 95 % para el oddsratio.

5. Vimos en la clase el algoritmo de *Iterative Proportional Fitting*. El paso crucial es definido por la transformación:

$$T_c(p)(x) = p(x)\frac{n(x_c)/n}{p(x_c)}. \tag{2}$$

a) Verifica que si $p()$ es una función de probabilidad, $T_c(p)$ también lo es.

b) Implementa una función IPF en C. Puedes suponer que la función tiene como argumentos la tabla de contingencia (como un arreglo) y una lista con los cliques del grafo subyacente.
Ajusta un modelos gráfico adecuado para la tabla de contingencia con los accesos de fuera de CIMAT a las páginas WWW del CIMAT entre enero y marzo del 2000.

		$X_3=0$	$X_3=0$	$X_3=1$	$X_3=1$
		$X_4=0$	$X_4=1$	$X_4=0$	$X_4=1$
$X_1=0$	$X_2=0$	1748	60	106	53
$X_1=0$	$X_2=1$	157	3	8	3
$X_1=1$	$X_2=0$	261	2	5	13
$X_1=1$	$X_2=1$	17	0	2	1

Los X_i's refieren a visitar o no alguna pagina WWW de la maestría i del CIMAT.

c) (*) Ya mencionamos que una desventaja de la formulación en (2) es que se tiene guardar en la memoria $\{p(x)\}$. Si supongamos que $x_i \in \{0,1\}$ y que $P()$ es de la forma

$$p(x) \sim \exp(\sum_{i=1}^{k} \mu_i \prod_{j \in A_i} x_j), \tag{3}$$

donde A_i son ciertos subconjuntos dados de $\{1, \cdots, n\}$, podemos - bajo ciertos supuestos - evitar el uso de tanta memoria.
El truco consiste en escribir (2) en función de los parámetros μ_i. Si k es pequeño y los A_i tienen muy pocos elementos, eso puede ser mucho más eficiente.

Deriva una versión en función de los parámetros (puedes tomar un caso particular; por ejemplo $d=4$, $A_i = \{i\}, 1 \leq i \leq 4$ y $A_5 = \{1,2\}$). Como primer paso escribe en (2) P en función de las μ's a través de (3) y aprovecha que no hay necesitad de correr (2) sobre todos los x.