Compresión de datos sin pérdida
Práctico 6

Ejercicio 1 *Universalidad puntual y en promedio*
Probar que un código puntualmente universal es también universal en promedio.

Ejercicio 2 *Estimadores de Laplace y de Krichevsky–Trofimov (KT)*
Calcular la probabilidad que asignan a la secuencia binaria $x^n = 001010$ los estimadores de Laplace y de Krichevsky—Trofimov (KT). ¿Existe alguna secuencia de largo finito, x^n, tal que la probabilidad asignada por el estimador KT es mayor que su probabilidad empírica de orden 0, $\hat{P}(x^n)$? Probar que no o dar un ejemplo.

Ejercicio 3 *Minimax regret*
Probar que la asignación de probabilidad NML es la **única** que alcanza exactamente el mínimo arrepentimiento de peor caso para una familia de modelos.

Ejercicio 4 *Codificación universal óptima para procesos de Markov*
Considerar la familia de modelos de Markov de orden k sobre un alfabeto finito.

1. Probar que la asignación de probabilidad que se obtiene aplicando el estimador KT en cada estado de la cadena es puntualmente universal en esta familia de modelos, con una tasa de convergencia óptima (en el sentido minimax y de la cota inferior de Rissanen).

2. Para orden de Markov $k = 1$ y alfabeto binario, $\mathcal{A} = \{0, 1\}$, calcular la probabilidad que se le asigna a la secuencia $x^n = 00101110011$.

Ejercicio 5 *Asignaciones secuenciales*
Sea $\mathcal{C} = \{P_\theta\}_{\theta\in\Theta}$ una familia paramétrica de modelos tal que se cumple

$$\sum_{a\in\mathcal{A}} P_\theta(x^{n-1}a) = P_\theta(x^{n-1})\,, \quad \forall\theta \in \Theta\,.$$

1. Mostrar que la sucesión $\{Q_n\}_{n\in\mathbb{N}}$, donde cada distribución Q_n sobre $\mathcal{A}^n$ es un mezcla de las distribuciones de $\mathcal{C}$ con una ponderación ω,

$$Q_n(x^n) = \int_{\theta\in\Theta} P_\theta(x^n)\omega(\theta)d\theta\,,$$

   define una asignación secuencial de probabilidades. Esto es, para $n > 1$, se cumple que

$$\sum_{a\in\mathcal{A}} Q_n(x^{n-1}a) = Q_{n-1}(x^{n-1})\,.$$

2. Considerar la familia de procesos de Bernoulli y calcular $Q_2^{\mathrm{NML}}(11)$ y $Q_3^{\mathrm{NML}}(11)$. Concluir que la asignación NML no es secuencial.

Ejercicio 6 *Hipótesis para cota de Rissanen*
Probar que la familia de modelos de Bernoulli satisface las hipótesis del enunciado visto en el curso para la cota inferior de Rissanen.
Sugerencia: Usar la desigualdad de Chebyshev: Si X^n es una secuencia de variables i.i.d., con media μ y varianza σ^2, entonces

$$P\left\{\left|\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) - \mu\right| > \delta\right\} < \frac{\sigma^2}{n\delta^2}\,.$$

Ejercicio 7 *Tesis de la cota de Rissanen*
Consideramos una familia paramétrica de modelos, $\mathcal{C} = \{P_\theta\}_{\theta\in\Theta}$, para secuencias sobre un alfabeto finito $\mathcal{A}$, donde Θ es un subconjunto acotado de $\mathbb{R}^k$. Para $\theta \in \Theta$ definimos $E_n(\theta)$ como la bola de radio $\frac{\log n}{\sqrt{n}}$ y centro θ, y definimos $\mathcal{X}_n(\theta)$ como el conjunto de secuencias $x^n \in \mathcal{A}^n$ cuyo estimador de máxima verosimilitud, $\hat{\theta}(x^n)$, pertenece a $E_n(\theta)$. Supongamos que $P_\theta\Big(\mathcal{X}_n(\theta)\Big) \geq 1 - \delta_n$ para todo $\theta \in \Theta$, donde δ_n tiende a 0 con n.

Las hipótesis enunciadas hasta aquí corresponden a la versión vista en clase de la cota de Rissanen. Considere la siguiente afirmación: para todo ϵ positivo existe N_0 tal que

$$D(P_\theta||Q_n) \geq (1-\epsilon)\frac{k}{2}\log n\,, \quad \forall\theta \in \Theta\,, \forall n > N_0\,.$$

1. Mostrar que este enunciado es incorrecto (por ejemplo mediante un contraejemplo).
2. Corregir el enunciado.

Ejercicio 8 *Secuencias con restricciones*
Cuando se graba información digital en algunos medio físicos, por ejemplo en discos o cintas magnéticas, puede ser necesario imponer restricciones en las secuencias de bits que se pueden almacenar, debido a las características físicas del dispositivo y el medio de almacenamiento. Por ejemplo, para evitar problemas de sincronismo, puede ser necesario limitar la cantidad de ceros consecutivos y, para evitar interferencias entre símbolos, puede ser necesario imponer que haya al menos un cero entre dos unos consecutivos. Supongamos que la secuencia de bits que se va a almacenar en un dispositivo cumple con las siguientes restricciones:

- Hay al menos un cero entre todo par de unos.
- No puede haber más de tres ceros consecutivos.

Modelizar la secuencia como un proceso aleatorio generado por una FSM con probabilidades condicionales desconocidas (salvo las impuestas por las restricciones anteriores). Evalúe el costo de modelo que establece la cota de Rissanen para este caso.