Ejercicio 9 : CC50Q - Teoría de la Información y Redes Neuronales

Prof.: Pedro Ortega <peortega@dcc.uchile.cl>
Aux.: Francisco Claude <fclaude@dcc.uchile.cl>

11 de octubre de 2005

- Entrega: Lunes 17 de Octubre, 12:00 horas, en la cátedra. -

Enunciado

El objetivo del presente ejercicio es familiarizarlo con el teorema de codificación de una fuente con ruido. Para lograr este objetivo, Ud. deberá completar la demostración del teorema en las partes solicitadas. Apóyese en la demostración del teorema disponible en el libro "Information Theory, Inference, and Learning Algorithms", capítulo 10.

Teorema

1. Para cada canal discreto binario sin memoria, la capacidad C dada por

   $$C = \max_{\mathcal{P}_X} I(X;Y)$$

   posee la siguiente propiedad. Para todo $\varepsilon > 0$, $R < C$ y N suficientemente grande, existe un código de largo N y una tasa $\geq R$ y un algoritmo de decodificación, tal que la probabilidad máxima de error de bloque es $< \varepsilon$.

2. Si se acepta una probabilidad de error de bit p_b, entonces pueden alcanzarse tasas hasta $R(p_b)$, donde

   $$R(p_b) = \frac{C}{1 - H(p_b, 1-p_b)}.$$

3. Para toda probabilidad de error de bit p_b, las tasas de transferencia mayores a $R(p_b)$ no son alcanzables.

Definiciones preliminares

La idea de la demostración consiste en calcular la probabilidad de error de bloque promediada sobre *todos* los códigos de bloques. Al realizar este cálculo, se puede demostrar que este promedio es bajo. Por lo tanto uno de ellos debe tener una probabilidad de error de bloque inferior al promedio.

Antes de calcular la probabilidad de error de bloque, hay que definir la familia de códigos de bloque. Aquí usaremos códigos que son fáciles de construir y analizar, y que se basan en la extensión de los canales para disminuir los errores.

Secuencias típicas:

Para formalizar la idea, hay que definir antes la noción de *secuencias típicas*. Intuitivamente, una secuencia típica es una secuencia que puede codificarse en un número de bits aproximadamente igual a aquel dictado por su entropía. El conjunto de las secuencias típicas posee la propiedad de ser aquel que "se llevan casi toda la probabilidad" de la distribución. O, equivalentemente, el conjunto de las secuencias no-típicas prácticamente no ocurre.

Definición: Una secuencia $\mathbf{x} \in X^N$ es una secuencia típica de $P(x)$ a una tolerancia β ssi

$$\left| \frac{1}{N} \log_2 \frac{1}{P(\mathbf{x})} - H(X) \right| < \beta$$

donde N es el largo de $\mathbf{x}$.

Problema: Considere un arreglo de probabilidad $\mathcal{X}$ con $\mathcal{A}_X = \{a, b\}$ y $\mathcal{P}_X = \{0{,}8, 0{,}2\}$. Determine las

secuencias típicas del arreglo extendido $\mathcal{X}^3$ hasta una tolerancia 0,1.

La noción de secuencias típicas tiene su contraparte a nivel de canales. Para un canal dado (simple o extendido), podemos definir un conjunto de secuencias *típicas conjuntas*. Intuitivamente, las secuencias típicas conjuntas son aquellas que suelen aparecer en conjunto: una en cada extremo del canal.

Definición: Un par de secuencias $\mathbf{x}, \mathbf{y}$ de largo N cada una son típicas conjuntas a una tolerancia β con respecto a la distribución $P(x, y)$ ssi

$$\mathbf{x} \text{ es típica de } P(x), \text{i.e. } \left| \frac{1}{N} \log_2 \frac{1}{P(\mathbf{x})} - H(X) \right| < \beta,$$

$$\mathbf{y} \text{ es típica de } P(y), \text{i.e. } \left| \frac{1}{N} \log_2 \frac{1}{P(\mathbf{y})} - H(Y) \right| < \beta,$$

$$\mathbf{x}, \mathbf{y} \text{ es típica de } P(x, y), \text{i.e. } \left| \frac{1}{N} \log_2 \frac{1}{P(\mathbf{x}, \mathbf{y})} - H(X, Y) \right| < \beta.$$

Problema: Dé un ejemplo en donde las secuencias $\mathbf{x}, \mathbf{y}$ son típicas de $P(x, y)$ (a una tolerancia β), pero en que ni $\mathbf{x}$ ni $\mathbf{y}$ lo son de $P(x)$ y $P(y)$ respectivamente.

Problema: Para un canal Z con probabilidad de error $f = 0{,}2$, determine si

1. $\mathbf{x} = 0000011111$ y $\mathbf{y} = 0000011101$
2. $\mathbf{x} = 0101010101$ y $\mathbf{y} = 1101010101$
3. $\mathbf{x} = 0001110000$ y $\mathbf{y} = 0001100000$

son típicamente conjuntas a una tolerancia $\beta = 0{,}2$, donde $P(x = 0) = 0{,}5$ y $P(x = 1) = 0{,}5$.

Ahora que hemos definido a las secuencias típicas conjuntas, definiremos a $J_{N\beta}$ al *conjunto de todas las secuencias típicas conjuntas* de largo N para una distribución $P(x, y)$ (a una tolerancia β).

En clases hemos observado que al extender un canal (i.e. transmitiendo bloques de símbolos en vez de bit a bit) las probabilidades tienden a acumularse en regiones localizadas de la matriz de probabilidad. El siguiente teorema explica este fenómeno.

Teorema de secuencias típicas conjuntas: Sean $\mathbf{x}, \mathbf{y}$ secuencias muestreadas al azar del arreglo de probabilidad conjunto $(X, Y)^N$ definido por

$$P(\mathbf{x}, \mathbf{y}) = \prod_{n=1}^{N} P(x_n, y_n)$$

Entonces,

1. la probabilidad de que $\mathbf{x}$ e $\mathbf{y}$ sean típicamente conjuntas (a una tolerancia β) tiende a 1 cuando $N \to \infty$.
2. el número de secuencias típicas conjuntas $|J_{N\beta}|$ tiende a $2^{NH(X,Y)}$. Más precisamente,
$$|J_{N\beta}| \leq 2^{N(H(X,Y)+\beta)}.$$
3. si $\mathbf{x}'$ e $\mathbf{y}'$ son muestreados en forma *independiente* (i.e. no en forma conjunta) de las distribuciones marginales $P(\mathbf{x})$ e $P(\mathbf{y})$, entonces la probabilidad de que el par $(\mathbf{x}, \mathbf{y})$ caiga (por casualidad) dentro del conjunto típico conjunto $J_{N\beta}$ es cercano a $2^{-NI(X;Y)}$. Para ser más preciso,
$$P((\mathbf{x}', \mathbf{y}') \in J_{N\beta}) \leq 2^{-N(I(X;Y)-3\beta)}.$$

Problema Dado un canal Z con probabilidad de error $f = 0{,}1$ y las probabilidades del arreglo de emisión $\mathcal{P}_X = \{0{,}5, 0{,}5\}$, determine (para una tolerancia $\beta = 0{,}1$)

1. el número de secuencias $\mathbf{x} \in X^3$ típicas,
2. el número de secuencias $\mathbf{y} \in Y^3$ típicas,
3. el número de secuencias $\mathbf{x}, \mathbf{y} \in X^3, Y^3$ típicas, donde $\mathbf{x}$ e $\mathbf{y}$ han sido muestreadas en forma *independiente*,
4. el número de secuencias $\mathbf{x}, \mathbf{y} \in X^3$ típicas conjuntas. Haga una matriz para ilustrar estos conjuntos: una fila por secuencia $\mathbf{y}$ posible y una columna por secuencia $\mathbf{x}$ posible. Inspírese en el cuadro 10.2 del libro.

Problema Demuestre el teorema de secuencias típicas conjuntas (en el libro está incompleto).

Códigos de bloque aleatorios:

Ahora que entendemos lo que son las secuencias típicas conjuntas, el siguiente esquema de codificación-decodificación nos hará sentido. Inventaremos al azar un código para transmitir números naturales, sin perder generalidad.

1. *Escoger tamaño y tasa:* Dado un canal binario y sin memoria, escogemos un tamaño de bloque N para nuestros códigos y una tasa de transferencia R'.

2. *Inventar el código en forma aleatoria:* Acorde con nuestra distribución de probabilidad de la fuente, $P(x)$, generamos $S = 2^{NR'}$ códigos *al azar*, muestreados de la distribución de probabilidad
$$P(\mathbf{x}) = \prod_{n=1}^{N} P(x_n).$$
Nótese que de esta manera, por cada bloque de N bits que transmitimos a través del canal, el receptor estará decodificando una entre S alternativas, i.e. NR' bits (*entienda esta última frase*). A los códigos resultantes los llamaremos $\mathbf{x}^{(1)}$, $\mathbf{x}^{(2)}$, $\ldots$, $\mathbf{x}^{(S)}$.

3. *Establecer el protocolo de transmisión:* Una vez generados los códigos, el emisor y el receptor se ponen de acuerdo para utilizar los códigos anteriores como medio de comunicación.

4. *Codificación y transmisión:* Supongamos ahora que el emisor quiere enviar el número/símbolo s. Entonces, s lo codifica como $\mathbf{x}^{(s)}$ y envía este código a través del canal.

5. *Recepción:* El receptor recibirá un mensaje $\mathbf{y}$ acorde con las propiedades de transmisión del canal. Es decir,
$$P(\mathbf{y}|\mathbf{x}^{(s)}) = \prod_{n=1}^{N} P(y_n|x_n^{(s)}).$$

6. *Decodificación del mensaje:* El mensaje recibido $\mathbf{y}$ se decodifica utilizando la *decodificación de conjuntos de secuencias típicas conjuntas*. Esta decodificación es muy sencilla: $\mathbf{y}$ se decodifica como el número $\hat{s}$ si el par $(\mathbf{x}^{(}\hat{s}), \mathbf{y})$ es típicamente conjunto. Si existe más de una decodificación posible, entonces la decodificación se declara como *fallida*.

Problema: Invente un código de bloque aleatorio de largo $N = 12$ para un canal binario simétrico con $f = 1/6$ para transmitir a una tasa de $1/3$. Use un dado (de seis caras) para este propósito.

Notemos que para este sistema de codificación/decodificación existen dos posibles fuentes de error de transmisión. El primero es el fallo en la decodificación, dado por el propio protocolo. El segundo es cuando el número decodificado $\hat{s}$ no coincide con el número s emitido.

Ahora que hemos entendido cómo funciona nuestro sistema de comunicación basado en secuencias típicas conjuntas, estamos listos para partir la demostración del teorema.

Demostración

Parte 1

La demostración de la primera parte del teorema la lograremos tomando los datos (máxima probabilidad de error de bloque y mínima tasa de transferencia R) y generando todos los códigos de bloque al azar lo suficientemente buenos. Veremos que para un N lo suficientemente grande, este error promedio (sobre todas las codificaciones) es muy pequeño, y que esto implica que dentro de este conjunto debe haber uno particular que cumple nuestras restricciones.

Partamos calculando el error promedio. Por simetría, basta con calcular la probabilidad de transmitir mal el número $s = 1$. Tenemos dos fuentes de error:

1. *secuencia emitida y recibida no son típicas conjuntas:* La probabilidad de que $\mathbf{x}^{(1)}$ e $\mathbf{y}$ no sean típicas conjuntas decrece a cero para N suficientemente grande, debido a la parte 1 del teorema de secuencias típicas conjuntas. En particular, podemos denotar δ a esta probabilidad máxima.

2. *número enviado no coincide con número decodificado:* Esto es equivalente a suponer que en vez de decodificar $\hat{s} = 1$, decodificamos cualquiera de los demás $S-1$ $(= 2^{NR'} - 1)$ valores alternativos. Además, para que esto ocurra, la secuencia recibida debe ser típica conjunta con la secuencia enviada. Vimos que existe una probabilidad máxima para que esto ocurra, dada por la parte 3 del teorema de secuencias típicas conjuntas.

Juntando ambas posibilidades, obtenemos la siguiente cota para el error promedio de bloque $\langle p_B \rangle$:

$$\begin{aligned} \langle p_B \rangle &\leq \delta + \sum_{s'=2}^{2^{NR'}} 2^{-N(I(X;Y)-3\beta)} \\ &\leq \delta + (2^{NR'} - 1) \cdot 2^{-N(I(X;Y)-3\beta)} \\ &\leq \delta + 2^{-N(I(X;Y)-R'-3\beta)} \end{aligned}$$

Si escogimos una tasa $R' < I(X;Y) - 3\beta$ entonces el segundo sumando es muy pequeño:

$$\langle p_B \rangle \leq \delta + 2^{-N\alpha}$$

y aumentando N podemos hacerlo tan pequeño como queremos. En particular, podemos disminuir su tamaño para dejarlo en una cantidad inferior a δ. Así obtenemos que

$$\langle p_B \rangle \leq 2\delta.$$

Ahora ya estamos casi listos. Hemos escogido a R' para que sea inferior a $I(X;Y)-3\beta$, pero quisiéramos maximizar nuestra tasa de transferencia. Entonces, cambiamos la distribución de probabilidad de la entrada, para maximizar la información mutua. Así, R' termina siendo inferior a $R' < C-3\beta$. Es decir, nuestra tasa de transferencia es una cantidad despreciable inferior a la *capacidad del canal*.

Notemos que el cálculo anterior de la probabilidad de error de bloque $\langle p_B \rangle$ lo hemos realizado en forma estadística, i.e. no sobre una codificación $\mathcal{C}$ particular, sino que sobre el conjunto total de códigos de un largo N dado. Entonces, dentro de este conjunto, debe existir al menos una codificación con probabilidad de error de bloque promedio inferior a 2δ. Seleccionemos esta codificación.

La codificación escogida posee una probabilidad *promedio* muy baja de cometer errores. Sin embargo, nada nos garantiza que dentro de sus S códigos no existan algunos con altísima probabilidad de error. Ahora aplicaremos un truco para borrar aquellos códigos conflictivos, técnica conocida como *expurgación*. Con esto, lograremos que la máxima probabilidad de error de bloque sea inferior a 4δ, disminuyendo nuestra tasa de transferencia en una cantidad negligible.

Problema: Para esto, ordenemos los S códigos según su probabilidad de error, de menor a mayor. Dibuje un gráfico de barras/histograma esquemático. Sabemos que la suma de estas barras es inferior a $2\delta S$. Entonces, muestre que borrando la mitad mayor sobrevive la mitad menor de los códigos, y que el máximo de ellos no puede poseer una probabilidad de error mayor a 4δ. Además, muestre que la tasa de transferencia resultante se reduce a $R' - 1/N$.

Problema: En base al resultado anterior, finalice la demostración del teorema.

Parte 2

Primero, notemos que, debido a la parte 1, somos capaces de construir códigos de bloque con error p_B despreciable y tasa de transferencia C dado un canal de comunicación base. Sobre este nuevo canal (parchado) montaremos otro código de bloque para demostrar la parte 2.

Hasta ahora, la idea de comunicación ha sido transmitir K bits en bloques de $N > K$ bits que contienen redundancia (en los $N - K$ restantes). Esta redundancia le permite al decodificador corregir los errores producidos en el canal ruidoso. En particular, pensemos en un código de bloque que fue diseñado para un canal binario simétrico. Si el código de bloque casi perfecto, entonces obtiene tasas de transferencias cercanas a la capacidad del canal, i.e. $K/N \approx 1 - H(f, 1-f)$, donde f es su probabilidad de distorsión. Ahora, démosle otro uso a este sistema de codificación/decodificación.

Si intercambiamos los roles, entonces el decodificador antiguo se convierte en nuestro nuevo codifi-

cador. Análogamente, el codificador antiguo se convierte en nuestro decodificador. Notemos que de esta manera, hemos convertido el código de bloque original $K \to N \to K$ en otro con $N \to K \to N$, i.e. hemos construido un *compresor con pérdida*. Ahora debemos preguntarnos, ¿cuánta información se ha perdido en la compresión de $N \to K$?

La respuesta es sencilla. Como el código de bloque original tomaba bloques de K bits de la fuente y le sumaba $N - K$ bits de redundancia para *corregir* los fN bits que se distorsionaban durante la transmisión, entonces nuestro nuevo compresor está *corrompiendo* fN bits en la compresión $N \to K$. En resúmen, el compresor reconstruye un bit con una probabilidad de error igual a $p_b = f$ a una tasa

$$N/K = \frac{1}{1 - H(p_b, 1 - pb)}.$$

El procedimiento anterior nos indica una forma de construir un compresor para cualquier probabilidad de error de bit p_b que nos demos.

Problema: Tomando

1. el canal de comunicación parchado de error despreciable y tasa de transferencia C,
2. el compresor con pérdida de tolerancia p_b y tasa de compresión $1/(1 - H(p_b, 1 - p_b))$,

indique cómo construir un canal de comunicación de error de bit p_b y tasa de transferencia

$$R(p_b) = \frac{C}{1 - H(p_b, 1 - p_b)}.$$

Además, haga un esquemático de la arquitectura resultante.

Parte 3

Problema: Esta parte es la más sencilla de todas. Basado en la información disponible en el libro, demuestre la desigualdad de procesamiento de datos (*data processing inequality*) y luego demuestre la tercera parte del teorema *detallando los pasos con sus propias palabras.*