Pregunta 1

(a) Veamos que π_k es un estimador consistente de η_k. Consideremos tomar la esperanza condicionada a ambos lados de la igualdad, esto es

$$E\left[\sum_{i=1}^{n}\pi_k(X_i)X_{ij}|X_1,X_2,\ldots,X_n\right]=E\left[\sum_{i=1}^{n}\mathbb{1}(k=Y_i)X_{ij}|X_1,X_2,\ldots,X_n\right]\ \ \forall k\in[K]\,\forall j\in[d]$$

Por linealidad de la esperanza (y quitamos de la condicional lo que no dependa)

$$\sum_{i=1}^{n}E\left[\pi_k(X_i)X_{ij}|X_i\right]=\sum_{i=1}^{n}E\left[\mathbb{1}(k=Y_i)X_{ij}|X_i\right]\quad\forall k\in[K]\ \forall j\in[d]$$

Sacando los valores que se vuelven constantes al estar condicionando se obtiene

$$\sum_{i=1}^{n}\pi_k(X_i)X_{ij}=\sum_{i=1}^{n}E\left[\mathbb{1}(k=Y_i)|X_i\right]X_{ij}\ \ \forall k\in[K]\ \forall j\in[d]$$

Pero la esperanza de una función indicatriz es simplemente la probabilidad, esto es

$$\sum_{i=1}^{n}\pi_k(X_i)X_{ij}=\sum_{i=1}^{n}P\left(k=Y_i\right)X_{ij}\ \ \forall k\in[K]\ \forall j\in[d]$$

Reemplazando por η_k

$$\sum_{i=1}^{n}\pi_k(X_i)X_{ij}=\sum_{i=1}^{n}\eta_k(X_i)X_{ij}\ \ \forall k\in[K]\ \forall j\in[d]$$

Por lo cual el predictor de bayes es un estimador consistente, justificando la heurística

(b) Primero reescribamos el problema de optimización completo

$$\begin{aligned}\max_{\pi:\mathbb{R}^d\to\Delta_k}\quad & -\sum_{i=1}^{n}\sum_{k=1}^{K}\pi_k(X_i)\ln\pi_k(X_i)\\ s.t.\quad & \pi_k(X_i)\geq 0\quad\forall i\in[n],\forall k\in[K],\\ & \sum_{k=1}^{K}\pi_k(X_i)=1\quad\forall i\in[n],\\ & \sum_{i=1}^{n}\pi_k(X_i)X_{ij}=\sum_{i=1}^{n}\mathbb{1}(Y_i=k)X_{ij}\quad\forall k\in[K],\forall j\in[d].\end{aligned}$$

Ahora vamos a formular el Lagrangiano, los multiplicadores de Lagrange serán $\mu_{ki}\leq 0,\gamma_i,w_{kj}$ en el mismo orden que están las restricciones anteriores, de lo que resulta

$$\mathcal{L}(\pi,\mu,\gamma,\lambda) = -\sum_{i=1}^{n}\sum_{k=1}^{K}\pi_k(X_i)\ln\pi_k(X_i) + \sum_{i=1}^{n}\sum_{k=1}^{K}\mu_{ki}\pi_k(X_i) + \sum_{i=1}^{n}\gamma_i\left(\sum_{k=1}^{K}\pi_k(X_i)-1\right) + \sum_{k=1}^{K}\sum_{j=1}^{d}w_{kj}\left(\sum_{i=1}^{n}\pi_k(X_i)X_{ij} - \sum_{i=1}^{n}\mathbb{1}(Y_i=k)X_{ij}\right)$$

Ahora consideremos encontrar derivadas parciales. Para encontrar el óptimo $\pi_k^*(X_i)$, tomamos la derivada parcial de $\mathcal{L}$ respecto a $\pi_k(X_i)$ y la igualamos a cero, esto es

$$\frac{\partial\mathcal{L}}{\partial\pi_k(X_i)} = -\ln\pi_k(X_i) - 1 + \mu_{ki} + \gamma_i + \sum_{j=1}^{d}w_{kj}X_{ij} = 0$$

Reordenamos

$$\ln\pi_k^*(X_i) = \sum_{j=1}^{d}w_{kj}^*X_{ij} + \gamma_i + \mu_{ki} - 1$$

Y exponenciamos

$$\pi_k^*(X_i) = \exp\left\{\sum_{j=1}^{d}w_{kj}^*X_{ij} + \gamma_i + \mu_{ki} - 1\right\}$$

Notemos que en la función objetivo original se tiene el término $\ln\pi_k(X_i)$, el cual no está definido cuando $\pi_k(X_i) = 0$, y además se tiene que en el óptimo vimos que es función de una exponencial, por lo que nunca será $\pi_k^*(X_i) = 0$. Con lo anterior, podemos concluir que entonces $\pi_k^*(X_i) > 0$. Pero entonces, por ser holguras complementarias se tiene que $\mu_{ki} = 0$ (ya que corresponde a la variable de la restricción $\pi_k(X_i) \geq 0$) la que dijimos que se cumple de forma estricta) por lo que se tiene que

$$\pi_k^*(X_i) = \exp\left\{\sum_{j=1}^{d}w_{kj}^*X_{ij} + \gamma_i - 1\right\} = \exp\left\{\langle w_k^*, X_i\rangle\right\}\cdot e^{\gamma_i - 1}$$

De la restricción $\sum_{k=1}^{K}\pi_k(X_i) = 1$ se tiene que en el óptimo

$$\sum_{k=1}^{K}\exp\{\langle w_k^*, X_i\rangle\}\cdot e^{\gamma_i-1} = 1$$

Factorizamos

$$e^{\gamma_i-1}\sum_{k=1}^{K}\exp\{\langle w_k^*, X_i\rangle\} = 1$$

Por lo tanto

$$e^{\gamma_i - 1} = \frac{1}{\sum_{k=1}^{K} \exp\{\langle w_k^*, X_i \rangle\}}$$

Reemplazando en $\pi_k^*(X_i)$ se obtiene

$$\pi_k^*(X_i) = \frac{\exp\{\langle w_k^*, X_i \rangle\}}{\sum_{k=1}^{K} \exp\{\langle w_k^*, X_i \rangle\}}$$

Demostrando lo pedido

Pregunta 2

(a) Se tiene que la función de log-verosimilitud de los datos corresponde por definición a

$$L(S|\pi) = \sum_{i=1}^{n} \ln \pi_{Y_i}(X_i)$$

Sustituyendo $\pi_{Y_i}(X_i)$ por la parametrización encontrada se obtiene que

$$L(S|\pi) = \sum_{i=1}^{n} \ln \left(\frac{\exp(\langle w_{Y_i}, X_i \rangle)}{\sum_{k=1}^{K} \exp(\langle w_k, X_i \rangle)} \right) = \sum_{i=1}^{n} \left(\langle w_{Y_i}, X_i \rangle - \ln \left(\sum_{k=1}^{K} \exp(\langle w_k, X_i \rangle) \right) \right)$$

Esto es

$$-\frac{1}{n} L(S|\pi) = -\mathcal{R}_S(W) = \frac{1}{n} \sum_{i=1}^{n} \left(\langle w_{Y_i}, X_i \rangle - \ln \left(\sum_{k=1}^{K} \exp(\langle w_k, X_i \rangle) \right) \right)$$

Demostrando que la log-verosimilitud dada una hipótesis $\pi \in F$ es proporcional a $-\mathcal{R}_S(W)$, tal como queríamos demostrar. Ahora veamos que $\mathcal{R}_S$ es convexa

Primero veamos que $g(x) = \ln \left(\sum_{k=1}^{K} e^{x_k} \right)$ es convexa. Sea $x, y \in \mathbb{R}^k$, entonces

$$g(\lambda x + (1-\lambda)y) = \ln \left(\sum_{k=1}^{K} e^{\lambda x_k + (1-\lambda) y_k} \right) = \ln \left(\sum_{k=1}^{K} e^{\lambda x_k} e^{(1-\lambda) y_k} \right)$$

Aplicando la desigualdad de Holder y teniendo en cuenta que la función ln es creciente, resulta en

$$\ln \left(\sum_{k=1}^{K} e^{\lambda x_k} e^{(1-\lambda) y_k} \right) \leq \ln \left(\left(\sum_{k=1}^{K} e^{\lambda x_k \cdot \frac{1}{\lambda}} \right)^{\lambda} \left(\sum_{k=1}^{K} e^{(1-\lambda) y_k \cdot \frac{1}{1-\lambda}} \right)^{(1-\lambda)} \right)$$

Pero simplificando y separando la multiplicación del logaritmo por sumas fuera se obtiene

$$g(\lambda x + (1-\lambda)y) \leq \lambda \ln\left(\sum_{k=1}^{K} e^{x_k}\right) + (1-\lambda)\ln\left(\sum_{k=1}^{K} e^{y_k}\right) = \lambda g(x) + (1-\lambda)g(y)$$

Demostrando así la convexidad de g por definición. Pero entonces $\ln\left(\sum_{k=1}^{K}\exp(\langle w_k, X_i\rangle)\right)$ es una composición de una función lineal con una función convexa, resultando en una función convexa. Notemos que la función $-\langle w_{Y_i}, X_i\rangle$ es lineal y por tanto convexa. Así

$$\mathcal{R}_S = \frac{1}{n}\sum_{i=1}^{n}\left(-\langle w_{Y_i}, X_i\rangle + \ln\left(\sum_{k=1}^{K}\exp(\langle w_k, X_i\rangle)\right)\right)$$

Es una suma de funciones convexas, resultando en una función convexa

Como la función objetivo es convexa entonces el problema de minimizar $\mathcal{R}_S(W)$ es un problema convexo si Ω es convexo y cerrado.

(b) Para encontrar la constante de Lipchitz de la función vamos a calcular su derivada. Consideremos la función restringida a una sola muestra

$$\mathcal{R}_S^i = -\langle w_{Y_i}, X_i\rangle + \ln\left(\sum_{k=1}^{K}\exp(\langle w_k, X_i\rangle)\right)$$

Calculemos su derivada con respecto al vector de pesos correspondiente a la clase j

$$\frac{\partial \mathcal{R}_S^i}{\partial w_j} = \begin{cases} -X_i + \frac{\exp(\langle w_j, X_i\rangle)}{\sum_{k=1}^{K}\exp(\langle w_k, X_i\rangle)}X_i & \text{si } j = Y_i, \\ \frac{\exp(\langle w_j, X_i\rangle)}{\sum_{k=1}^{K}\exp(\langle w_k, X_i\rangle)}X_i & \text{si } j \neq Y_i. \end{cases}$$

Por lo que la derivada (considerando todas las clases y muestras) corresponden a lo siguiente.

$$\frac{\partial \mathcal{R}_S}{\partial w} = \frac{1}{n}\sum_{i=1}^{n}\left(\left(\frac{\exp(\langle w_1, X_i\rangle)}{\sum_{k=1}^{K}\exp(\langle w_k, X_i\rangle)} - \mathbb{1}_{(Y_i=1)}\right)X_i, \ldots, \left(\frac{\exp(\langle w_K, X_i\rangle)}{\sum_{k=1}^{K}\exp(\langle w_k, X_i\rangle)} - \mathbb{1}_{(Y_i=K)}\right)X_i\right)$$

Para verificar que $\mathcal{R}_S(w)$ es Lipschitz basta con verificar que la norma de $\nabla\mathcal{R}_S(w)$ es acotada, ya que podemos usar el teorema del valor medio.

Consideremos

$$p_k^i = \frac{\exp(\langle w_k, X_i\rangle)}{\sum_{j=1}^{K}\exp(\langle w_j, X_i\rangle)}.$$

Notemos que $\sum_{k=1}^{K} p_k^i = 1$, y además del gradiente original nos resulta

$$\frac{\partial \mathcal{R}_S}{\partial w} = \frac{1}{n}\sum_{i=1}^{n}\left(\left(p_1^i - \mathbb{1}_{(Y_i=1)}\right)X_i, \ldots, \left(p_k^i - \mathbb{1}_{(Y_i=K)}\right)X_i\right)$$

Para encontrar la constante de Lipchitz podemos encontrar la norma de este gradiente, esto es

$$\left\|\frac{\partial \mathcal{R}_S}{\partial w}\right\|_2 = \left\|\frac{1}{n}\sum_{i=1}^{n}\left(\left(p_1^i - \mathbb{1}_{(Y_i=1)}\right)X_i, \ldots, \left(p_k^i - \mathbb{1}_{(Y_i=K)}\right)X_i\right)\right\|_2$$

Sacando constante y usando la desigualdad triangular

$$\left\|\frac{\partial \mathcal{R}_S}{\partial w}\right\|_2 \leq \frac{1}{n}\sum_{i=1}^{n}\left\|\left(\left(p_1^i - \mathbb{1}_{(Y_i=1)}\right)X_i, \ldots, \left(p_k^i - \mathbb{1}_{(Y_i=K)}\right)X_i\right)\right\|_2$$

También podemos separar los vectores en dos, en la parte con p y la parte con la función indicatriz, y nuevamente usando la desigualdad triangular resulta en (en el vector con la función indicatriz debiera tener signo negativo, pero sale de la norma con signo positivo)

$$\left\|\frac{\partial \mathcal{R}_S}{\partial w}\right\|_2 \leq \frac{1}{n}\sum_{i=1}^{n}\left(\left\|\left(p_1^i X_i, \ldots, p_k^i X_i\right)\right\|_2 + \left\|\left(\mathbb{1}_{(Y_i=1)}X_i, \ldots, \mathbb{1}_{(Y_i=K)}X_i\right)\right\|_2\right)$$

Del vector de las indicatrices hay exactamente una entrada que es 1 y el resto 0, por lo que su norma es equivalente a la de la norma del vector X_i. Para el vector del lado izquierdo lo podemos separar entre los distintos bloques y reescribimos su norma

$$\left\|\frac{\partial \mathcal{R}_S}{\partial w}\right\|_2 \leq \frac{1}{n}\sum_{i=1}^{n}\left(\sqrt{\sum_{k=1}^{K}\left\|p_k^i X_i\right\|_2^2} + \|X_i\|_2\right)$$

Podemos sacar la constante p_k^i de la norma, y factorizar lo que no depende de k de la sumatoria, resultando en

$$\left\|\frac{\partial \mathcal{R}_S}{\partial w}\right\|_2 \leq \frac{1}{n}\sum_{i=1}^{n}\left(\sqrt{\|X_i\|_2^2\sum_{k=1}^{K}p_k^i} + \|X_i\|_2\right)$$

Pero recordando que $\sum_{k=1}^{K} p_k^i = 1$ nos resulta en

$$\left\|\frac{\partial \mathcal{R}_S}{\partial w}\right\|_2 \leq \frac{1}{n}\sum_{i=1}^{n}\left(\|X_i\|_2 + \|X_i\|_2\right) = \frac{2}{n}\sum_{i=1}^{n}\|X_i\|_2$$

Lo cual no depende de w, por lo que que la derivada es acotada. Con esto, se demuestra que la función es Lipchitz y su constante en la norma indicada corresponde a la encontrada, esto es, su constante es $\frac{2}{n}\sum_{i=1}^{n}\|X_i\|_2$

Ahora para encontrar la constante de suavidad derivemos una vez más para encontrar el Hessiano. Recordemos el gradiente e ignoremos el término $-X_i$ cuando $j = Y_i$ ya que al derivar una vez más al no depender de w se perderá, entonces tenemos que

$$\frac{\partial \mathcal{R}_S^i}{\partial w_j} = \frac{\exp(\langle w_j, X_i\rangle)}{\sum_{k=1}^{K}\exp(\langle w_k, X_i\rangle)}X_i$$

Derivemos con respecto a w_l, nos resulta usando la regla del cociente que

$$\frac{\partial^2 \mathcal{R}_S^i}{\partial w_j \partial w_l} = \frac{\left(\frac{\partial}{\partial w_l}\exp(\langle w_j, X_i\rangle)\right)\left(\sum_{k=1}^{K}\exp(\langle w_k, X_i\rangle)\right) - (\exp(\langle w_j, X_i\rangle))\left(\frac{\partial}{\partial w_l}\sum_{k=1}^{K}\exp(\langle w_k, X_i\rangle)\right)}{\left(\sum_{k=1}^{K}\exp(\langle w_k, X_i\rangle)\right)^2}X_i^T$$

Primero veamos el caso si $j = l$, en este caso se tiene que

$$\frac{\partial}{\partial w_l}\exp(\langle w_j, X_i\rangle) = \exp(\langle w_j, X_i\rangle)X_i$$

Y además

$$\frac{\partial}{\partial w_l}\sum_{k=1}^{K}\exp(\langle w_k, X_i\rangle) = \exp(\langle w_l, X_i\rangle)X_i$$

Reemplazando

$$\frac{\partial^2 \mathcal{R}_S^i}{\partial w_j \partial w_l} = \frac{(\exp(\langle w_j, X_i\rangle)X_i)\left(\sum_{k=1}^{K}\exp(\langle w_k, X_i\rangle)\right) - (\exp(\langle w_j, X_i\rangle))(\exp(\langle w_l, X_i\rangle)X_i)}{\left(\sum_{k=1}^{K}\exp(\langle w_k, X_i\rangle)\right)^2}X_i^T$$

Factorizando nos resulta

$$\frac{\partial^2 \mathcal{R}_S^i}{\partial w_j \partial w_l} = \frac{(\exp(\langle w_j, X_i\rangle)X_i)\left(\sum_{k=1,k\neq l}^{K}\exp(\langle w_k, X_i\rangle)\right)}{\left(\sum_{k=1}^{K}\exp(\langle w_k, X_i\rangle)\right)^2}X_i^T$$

De lo que resulta

$$\frac{\partial^2 \mathcal{R}_S^i}{\partial w_j \partial w_l} = p_j^i(1 - p_j^i) X_i X_i^T$$

Ahora veamos el caso si $j \neq l$, en este caso se tiene que

$$\frac{\partial}{\partial w_l} \exp(\langle w_j, X_i \rangle) = 0$$

Y además

$$\frac{\partial}{\partial w_l} \sum_{k=1}^{K} \exp(\langle w_k, X_i \rangle) = \exp(\langle w_l, X_i \rangle) X_i$$

Reemplazando

$$\frac{\partial^2 \mathcal{R}_S^i}{\partial w_j \partial w_l} = \frac{-\left(\exp(\langle w_j, X_i \rangle)\right)\left(\exp(\langle w_l, X_i \rangle) X_i\right)}{\left(\sum_{k=1}^{K} \exp(\langle w_k, X_i \rangle)\right)^2} X_i^T$$

De lo que resulta

$$\frac{\partial^2 \mathcal{R}_S^i}{\partial w_j \partial w_l} = -p_j^i p_l^i X_i X_i^T$$

Entonces, en términos generales se tiene que

$$\frac{\partial^2 \mathcal{R}_S^i}{\partial w_j \partial w_l} = p_j^i(\mathbb{1}_{(j=l)} - p_l^i) X_i X_i^T$$

Sea $H \in \mathbb{R}^{K \times d}$ arbitrario, entonces se tiene que

$$\langle \nabla^2 \mathcal{R}_S(w) H, H \rangle = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{K} \sum_{l=1}^{K} p_j^i(\mathbb{1}_{(j=l)} - p_l^i)(H_j^T X_i)(H_l^T X_i)$$

De la desigualdad de Cauchy-Schwarz se tiene que

$$p_j^i(\mathbb{1}_{(j=l)} - p_l^i)(H_j^T X_i)(H_l^T X_i) \leq p_j^i \|X_i\|^2 \|H_j\| \|H_l\|$$

Notando que $\sum_{k=1}^{K} p_k^i = 1$, se tiene que

$$\langle \nabla^2 \mathcal{R}_S(w) H, H \rangle \leq \max_i \|X_i\|^2 \|H\|_F^2$$

Por lo tanto, se concluye que la función es suave con constante de suavidad

$$C = \max_i \|X_i\|^2$$

Pregunta 3

(a) Primero veamos que $P(\epsilon) \leq N(\epsilon)$. Sea $\mathcal{P} \subseteq \Omega$ un ϵ-empaquetamiento máximo y $\mathcal{C} \subseteq \Omega$ un ϵ-cubrimiento mínimo. Sea $p \in \mathcal{P}$, como $\mathcal{C}$ es un ϵ-cubrimiento entonces existe un $c \in \mathcal{C}$ tal que $\|c - p\|_2 \leq \epsilon$. Sea $p' \in \mathcal{P}$ con $p' \neq p$, se tiene por la desigualdad triangular que $\|p' - p\|_2 \leq \|p' - c\|_2 + \|c - p\|_2$. Reordenando se tiene que $\|p' - c\|_2 \geq \|p' - p\|_2 - \|c - p\|_2 > 2\epsilon - \epsilon = \epsilon$ (Recordemos que $\|p' - p\|_2 > 2\epsilon$ ya que las bolas centradas en p y p' con radio ϵ no se intersectan). Entonces, para cada $p \in \mathcal{P}$ existe un $c \in \mathcal{C}$ que lo contiene en su bola de tamaño ϵ y dicho c no incluye a ningún otro $p' \in \mathcal{P} \setminus \{p\}$ en dicha bola, por lo que podemos concluír que $|\mathcal{P}| \leq |\mathcal{C}|$ y entonces $P(\epsilon) \leq N(\epsilon)$.

Ahora veamos que $N(2\epsilon) \leq P(\epsilon)$. Sea $\mathcal{P} \subseteq \Omega$ un ϵ-empaquetamiento. Si $\mathcal{P}$ no es un 2ϵ-cubrimiento, entonces hay un punto $x \in \Omega$ tal que $\|x - p\|_2 \geq 2\epsilon$ para todo $p \in \mathcal{P}$, por lo que el conjunto $\mathcal{P} \cup \{x\}$ también es un ϵ-empaquetamiento. Por lo tanto, cualquier ϵ-empaquetamiento máximo debe también ser un 2ϵ-cubrimiento. Por lo tanto, se concluye que un 2ϵ-cubrimiento mínimo debe tener menor o igual cardinalidad que el ϵ-empaquetamiento máximo.

Con esto podemos concluir que

$$N(2\epsilon) \leq P(\epsilon) \leq N(\epsilon)$$

Demostrando lo pedido

(b) Primero veamos que $\left(\frac{R}{\epsilon}\right)^m \leq N(\epsilon)$. Notemos que el hiper-volumen de la hiper-esfera $\mathcal{B}(0, R)$ es $V(\mathcal{B}(0, R)) = c_m R^m$ donde c_m es una constante que depende de la dimensión. Además tenemos que el hiper-volumen de una hiper-esfera $\mathcal{B}(x, \epsilon)$ es $V(\mathcal{B}(x, \epsilon)) = c_m \epsilon^m$. Entonces si yo tengo L bolas de tamaño ϵ, el volumen V que ocupan está acotado por $V \leq L \cdot c_m \epsilon^m$. Y como yo quiero que cubran volumen de $c_m R^m$ tenemos la expresión

$$c_m R^m \leq L \cdot c_m \epsilon^m \rightarrow \left(\frac{R}{\epsilon}\right)^m \leq L$$

Demostrando que $\left(\frac{R}{\epsilon}\right)^m \leq N(\epsilon)$, es una constante que es lo mínimo que tenemos que pedir para cubrir el espacio.

Ahora veamos que $N(\epsilon) \leq \left(1 + \frac{2R}{\epsilon}\right)^m$. Consideremos un $\frac{\epsilon}{2}$-empaquetamiento máximo $\mathcal{P}$, se tiene entonces que

$$\mathcal{B}\left(p, \frac{\epsilon}{2}\right) \subseteq \mathcal{B}\left(0, R + \frac{\epsilon}{2}\right) \quad \forall p \in \mathcal{P}$$

Por lo tanto, como cada uno es subconjunto entonces la unión también debe serlo

$$\bigcup_{p\in\mathcal{P}} \mathcal{B}\left(p, \frac{\epsilon}{2}\right) \subseteq \mathcal{B}\left(0, R+\frac{\epsilon}{2}\right)$$

Por lo que al ser un subconjunto el volumen debe ser menor o igual

$$V\left(\bigcup_{p\in\mathcal{P}} \mathcal{B}\left(p, \frac{\epsilon}{2}\right)\right) \leq V\left(\mathcal{B}\left(0, R+\frac{\epsilon}{2}\right)\right)$$

Por ser un empaquetamiento, entonces todas las bolas son disjuntas, por lo que

$$\sum_{p\in\mathcal{P}} V\left(\mathcal{B}\left(p, \frac{\epsilon}{2}\right)\right) \leq V\left(\mathcal{B}\left(0, R+\frac{\epsilon}{2}\right)\right)$$

Usando la fórmula de volumen encontrada anteriormente

$$\sum_{p\in\mathcal{P}} c_m \frac{\epsilon}{2}^m = |\mathcal{P}| c_m \frac{\epsilon}{2}^m = P\left(\frac{\epsilon}{2}\right) c_m \frac{\epsilon}{2}^m \leq c_m \left(R+\frac{\epsilon}{2}\right)^m$$

Entonces

$$P\left(\frac{\epsilon}{2}\right) \leq \left(\frac{R+\frac{\epsilon}{2}}{\frac{\epsilon}{2}}\right)^m = \left(1+\frac{2R}{\epsilon}\right)^m$$

Y por lo demostrado en la parte a se tiene que

$$N(\epsilon) \leq P\left(\frac{\epsilon}{2}\right) \leq \left(1+\frac{2R}{\epsilon}\right)$$

Concluyendo lo pedido. En resumen, se tiene que

$$\left(\frac{R}{\epsilon}\right)^m \leq N(\epsilon) \leq \left(1+\frac{2R}{\epsilon}\right)^m$$

(c) Por la desigualdad estudiada en clases

$$N_{LB}(2L\epsilon, \mathcal{G}, L_1(D)) \leq N(L\epsilon, \mathcal{G}, L_\infty(D))$$

Pero como $\phi(W,(X,Y))$ es Lipchitz con constante $\|X\|_2$, entonces usando la norma ℓ_∞ dicha constante resulta en $\sup_{x\in\mathcal{X}}(\|x\|_2) = 2L$. Por lo tanto, si existe un ϵ-cubrimiento de Ω, entonces al usar dichos puntos como parámetros de la función estamos consiguiendo un $L\epsilon$-cubrimiento de $\mathcal{G}$, esto es, $\mathcal{N}$ es un $L\epsilon$-cubrimiento de $\mathcal{G}$. Por lo que se concluye que

$$N_{LB}(2L\epsilon, \mathcal{G}, L_1(D)) \leq N(L\epsilon, \mathcal{G}, L_\infty(D)) \leq |\mathcal{N}|$$

Tal como queríamos demostrar

Así, si consideramos un cubrimiento ϵ de Ω, por lo demostrado en el punto anterior el número N de funciones a considerar es de a lo más

$$N \leq \left(1 + \frac{2R}{\epsilon}\right)^{Kd}$$

Y por tanto la complejidad muestral dado el δ corresponde a

$$O\left(\left(1 + \frac{2R}{\epsilon}\right)^{Kd} log\left(\frac{1}{\delta}\right)\right)$$

Mostrando que la regresión logística es PAC-aprendible