Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Departamento de Ciencia de la Computación

IIC2613 – Inteligencia Artificial (I/2019)

Interrogación 2

Pregunta 1

a) ¿Cuál es la principal diferencia entre una red neuronal convolucional multicapa, y una red neuronal multicapa basada unicamente en perceptrones? **(1.5 ptos.)**
Solución: muchas posibles respuestas, basta mencionar una: i) uso de convolución en vez de producto punto, ii) uso de capas de pooling, iii) combinación de capas convolucionales con fully-connected.

b) ¿Qué diferencia hay entre el origen de las capacidades no lineales de un SVM y las de una red neuronal multicapa? **(1.5 ptos.)**
Solución: en los SVM, la no linealidad se logra a partir de *kernels*, que son fijos durante el entrenamiento. En el caso de las redes neuronales de múltiples capas, la no linealidad se da por la conjunción las funciones de activación y el aprendizaje de features.

c) En base a la solución al problema de optimización dual de los SVM, indique cómo se identifican los vectores de soporte, y que condiciones cumplen. **(1.5 ptos.)**
Solución: acá se puede contestar usando ya sea la formulación *soft-margin*, o la *hard-margin*. En ambos casos, los vectores de soporte son aquellos donde los coeficientes α son distintos de cero, lo que indica que corresponden a restricciones (ejemplos) activos, que ayudan a determinar la superficie de decisión. En el caso de la formulación *hard-margin*, los vectores de soporte siempre se encuentran sobre el margen. En el caso de la formulación *soft-margin*, los vectores de soporte pueden ubicarse también hacia el lado contrario del margen.

d) ¿Cuál es la función del *dropout* en una red neuronal? **(1.5 ptos.)**
Solución: el *dropout* permite manejar la complejidad de una red, disminuyendo la dependencia de esta a la coactivación de neuronas. De esta manera, la red se ve obligada a aprender representaciones más redundantes. Otra manera de verlo, es que el *dropout* permite entrenar una red como si fuese un ensamble de modelos, lo que disminuye su varianza.

Pregunta 2

a) Considere un problema de clasificación, donde se debe decidir el genero musical al que pertenece una canción (representada como un vector). Si se utiliza una red neuronal convolucional para resolver este problema, ¿cuál es la principal diferencia que tendría esta red, con respecto a una del mismo tipo utilizada para clasificar imágenes? **(3.0 ptos)**
Solución: dado que el input de 1D, las convoluciones deberán ser también 1D, y no 2D como en una red que procesa imágenes.

b) Construya un grafo de cómputo que permita entrenar un red neuronal con 2 capas ocultas, para un problema de categorización con Y clases, donde cada ejemplo x_i tiene Y dimensiones. **(3.0 ptos)**
Solución: Acá son posibles múltiples soluciones. Las condiciones que estas deben cumplir son ls siguiente:

- Una capa inicial, ejemplificada al menos con 1 nodo, donde cada nodo corresponde a una de las Y entradas.
- Dos capas ocultas, ejemplificadas cada una con al menos un nodo. El nodo debe indicar claramente la función que calcula (perceptrón, convolución, etc.).
- Una capa de salida, ejemplificada al menos con 1 nodo, donde cada nodo corresponde a una de las Y salidas. Cada nodo debe definir claramente la función que calcula.
- Una capa para la pérdida, ejemplificada con 1 nodo (o una secuencia lineal de nodos). El nodo debe definir claramente la función que calcula.
- Conexiones claras entre las capas (no es necesario dibujar todas las aristas del grafo).
- Diferenciación clara de cuáles son los parámetros y entradas en cada nodo.

Pregunta 3

a) Considere el siguiente problema cuadrático con restricciones lineales, asociado a los SVMs:

$$\begin{aligned} \underset{w,b,\{\xi_i\}}{\text{argmin}} \quad & \frac{1}{2}\|w\|^2 + C\sum_i^N \xi_i \\ \text{s.a} \quad & y_i(w^\top x_i + b) \geq 1 - \xi_i, \forall i \in [1, N] \\ & \xi_i \geq 0, \forall i \in [1, N] \end{aligned}$$

(De)muestre que este problema puede escribirse de la forma:

$$\underset{w,b}{\text{argmin}} \quad R(w) + C\sum_i^N \ell\left(y_i, f\left(x_i; w, b\right)\right)$$

donde $f(x_i; w, b) = w^\top x_i + b$, y $\ell(y, z) = \max(0, 1 - yz)$. ¿Cuál es la forma de la función de pérdida ℓ, si se grafica en función de yz? **(3.0 ptos)**
Solución: el primer paso, corresponde a modificar el primer conjunto de restricciones de la siguiente forma:

$$\begin{aligned} y_i(w^\top x_i + b) &\geq 1 - \xi_i \\ \xi_i &\geq 1 - y_i(w^\top x_i + b) \end{aligned}$$

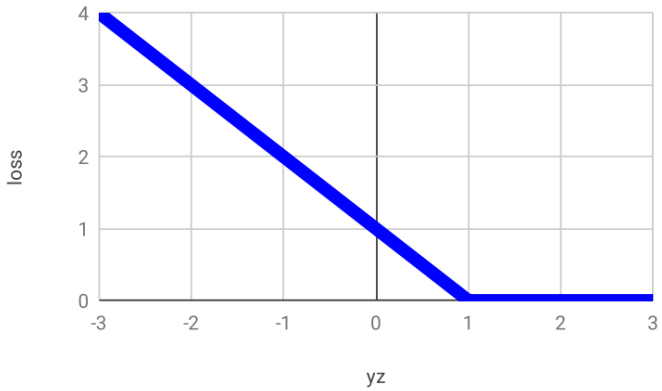
Luego, ξ_i está acotado por abajo por $1 - y_i(w^\top x_i + b)$. Dado que además se debe cumplir que $\xi_i \geq 0$, es posible combinar ambas restricciones de la siguiente manera:

$$\xi_i \geq \max\left(0, 1 - y_i(w^\top x_i + b)\right), \forall i \in [1, N]$$

A continuación, dado que estamos tratando de resolver un problema de minimización, que además incluye ξ_i de manera directa, podemos sustituir los ξ_i por sus respectivas cotas inferiores, lo que nos deja con el siguiente problema cuadrático sin restricciones:

$$\underset{w,b}{\text{argmin}} \quad \frac{1}{2}\|w\|^2 + C\sum_i^N \max\left(0, 1 - y_i(w^\top x_i + b)\right)$$

que es justamente la forma que se pide, considerando $R(w) = \frac{1}{2}\|w\|^2$. Finalmente, el gráfico de la función de pérdida queda de la siguiente forma:



b) Considere un problema de clasificación binario (dos clases), donde el espacio de características corresponde a histogramas normalizados de tamaño K. Diseñe un *kernel* para un SVM, que sea especializado para resolver un problema de este tipo. **(3.0 ptos)**
Solución: una posible solucionar es calcular la similitud entre histogramas es medir la intersección de estos. Dado

que cada casillero es un escalar, la intersección de cada uno de estos se puede calcular simplemente como el mínimo entre ambos escalares. Luego, la intersección entre dos histogramas, x e y, de tamaño K, se puede escribir de la siguiente manera:

$$\mathcal{K}(x, y) = \sum_{k=1}^{K} \min(x_k, y_k)$$

donde x_k e y_k corresponden al k-ésimo intervalo de x e y, respectivamente.