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

IIC2613 – Inteligencia Artificial (II/2017)

Solución Interrogación 3

Pregunta 1

a) ¿Como se podría realizar clasificación con un árbol de decisión, si falta el valor de alguna de las dimensiones del vector de entrada? **(1 pto.)**
Solución: existen al menos dos opciones: i) rellenar el valor faltante en base a los existente en el set de entrenamiento y ii) calculando la moda de las predicciones, si se toman todos los caminos del test que no se puede realizar.

b) Explique por qué la cantidad de parámetros activos (distintos de cero) en una red neuronal, puede dar una noción de la complejidad del modelo. **(1 pto.)**
Solución: los parámetros iguales a cero implican menos parámetros en uso (menos conexiones entre neuronas), luego, un modelo con más parámetros iguales a cero, es menos complejo.

c) Indique como utilizaría validación cruzada para estimar la mejor profundidad de un árbol de decisión. **(1 pto.)**
Solución: para cada ronda de validación, se limita la profundidad máxima del árbol. Luego, después de K rondas, se puede elegir la mejor profundidad (hasta K) en base al rendimiento promedio.

d) ¿De qué manera el método del momentum disminuye el riesgo de caer en mínimos locales en redes neuronales con capas ocultas? **(1 pto.)**
Solución: al combinar linealmente el valor del nuevo gradiente con la dirección de descenso anterior (una combinación de gradientes previos), los puntos con derivada cero pueden ser evitados, ya que no disminuye la *velocidad* del descenso.

e) ¿En qué situaciones es preferible utilizar el radio de ganancia por sobre la ganancia de información? **(1 pto.)**
Solución: cuando la cantidad de valores de la variable a testear es muy grande (ganancia de información tiende a elegir siempre a estas variables).

f) ¿Como podría utilizarse una red neuronal para aumentar la resolución de imágenes (manteniendo una buena calidad)? **Hint:** enfóquese en el entrenamiento de una red para esta tarea. **(1 pto.)**
Solución: utilizando un set de entrenamiento donde se tiene cada imagen en dos versiones (pequeña y grande) y una red que toma la versión pequeña y genera como salida una imagen de tamaño grande. La función de pérdida sería la suma de las diferencias cuadrática entre los valores de cada pixel predicho y el real (versión grande de la imagen).

Pregunta 2

a) Considere un problema de clasificación sobre variables categóricas. Extienda el algoritmo de construcción de los árboles de decisión basado en la ganancia de información, para que se puedan realizar tests sobre dos variables de manera simultánea **(2 ptos.)**.
Solución: Basta con modificar la métrica de ganancia de información, tomando ahora las proporciones de ejemplos (probabilidades) en el producto cartesiano de las variables (atributos) elegidas.

b) En general, al momento de decidir el valor a testear en un nodo de un árbol de decisión, se toma la ganancia de información como métrica. Una de las desventajas de esta, es que no considera la nueva estructura del árbol en el cálculo (la resultante de seleccionar ese test, con distinta profundidad y número de nodos), lo que puede derivar en problemas de sobreentrenamiento. Extienda la métrica de la ganancia de información, agregando un nuevo término aditivo, de manera que ahora, para tomar la decisión, se considere información sobre la posible nueva estructura del árbol. **Hint:** considere la decisión en un nodo como la minimización del riesgo estructural empírico. **(2 ptos.)**
Solución: La idea es agregar un término que penalice tanto la nueva profundidad del árbol y/o la nueva cantidad de nodos. Es importante notar además que un incremento en la profundidad es más importante desde el punto de vista del sobreentrenamiento, cuando se hace a mayor profundidad, *i.e.*, a mayor profundidad, mayor riesgo de sobreentrenamiento. Una posible solución podría ser la siguiente:

$$F(S, A) = Gain(S, A) - \alpha \cdot 2^d \cdot |S_A|$$

donde α es una constante predefinida, d es la nueva profundidad del árbol y $|S_A|$ es la cantidad de valores que toma el atributo A (cantidad de nodos que se agregarán).

c) Considere una competencia, donde se debe resolver un problema de regresión en base a variables categóricas. Dado que en este problema el riesgo de sobreentrenamiento es alto, sólo se permite utilizar árboles de regresión sobre una (1) de las variables disponibles, con el fin de limitar la profundidad del árbol. Utilizando múltiples árboles de **manera secuencial** (cada árbol puede utilizar la variable que quiera), indique como es posible construir un sistema de regresión que estime de mejor manera la función buscada. **(2 ptos.)**
Solución: Una posible solución es estimar de manera secuencial el valor de la regresión, corrigiendo las estimación en base al residuo de esta última, *i.e.*, si $R(x)$ es el valor a estimar para un ejemplo x, la regresión se realizaría usando $K+1$ árboles de regresión F_k de profundidad 1: $R(x) = F_0(x) + F_1(x) + \cdots + F_K(x)$. En esta modalidad, el árbol F_0 realizaría la regresión tradicional sobre x, mientras que F_1 estimaría el valor $R(x) - F_0(x)$. De manera análoga, el árbol F_k estimaría $R(x) - \sum_i = 0^k F_i(x)$.

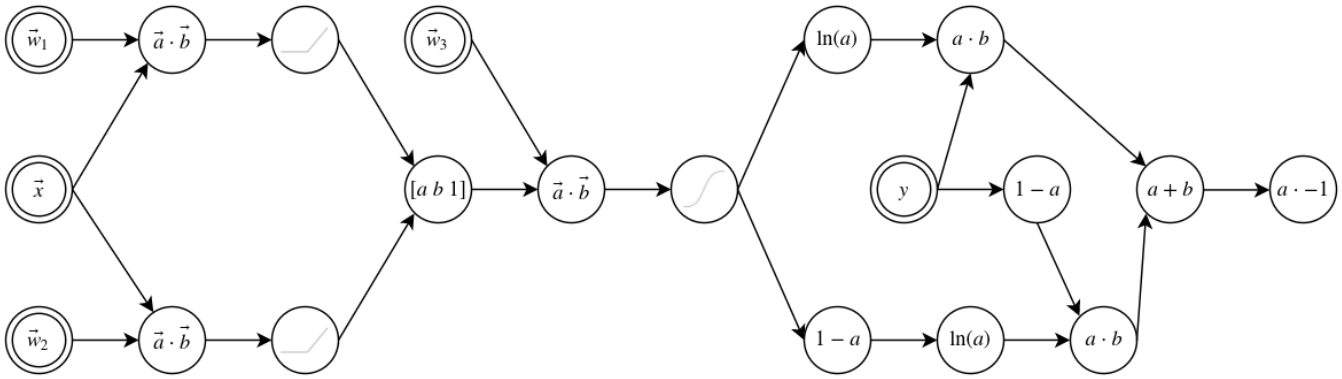
Pregunta 3

Dado un conjunto de ejemplos $x = \bigcup_i^N x_i$, con $x_i \in \Re^L$ y sus etiquetas asociadas $y = \bigcup_i^N y_i$, con $y_i \in [0, 1]$, considere la función de pérdida *cross-entropy* definida a continuación:

$$E(x, y; w) = -\sum_i y_i \ln o^i + (1 - y_i) \ln(1 - o^i)$$

donde o^i es la salida de la red (perceptrón sigmoidal) para el ejemplo i y w es el vector de parámetros de la red. En base a esto, conteste las siguientes preguntas:

a) Construya el grafo de cómputo para una red con una capa oculta de 2 neuronas con función de activación ReLU, que realice la clasificación de x, utilizando como pérdida la función $E(x, y; w)$. No combine múltiples operaciones en un sólo nodo del grafo. **(3 ptos.)**
Solución:



b) Extienda la definición de $E(x, y; w)$ para el escenario de clasificación multiclase (más de dos categorías), *i.e.*, $y_i \in [0, K]$. **(3 ptos.)**
Solución: Una posible solución es asumir que la salida de la red tiene K dimensiones (una por categoría), y que cada una de estas tiene como no linealidad una sigmoide. Luego, la siguiente función tiene como objetivo penalizar: i) valores distintos de 0 para cualquier dimensión que no sea la correspondiente a la clase correcta, y ii) cualquier valor distinto de 1 para la clase correcta:

$$\hat{E}(x, y; w) = -\sum_i \left(ln(o^i_{y_i}) + \sum_k^K ln(1 - o^i_k) \right)$$