Universidad Técnica Federico Santa María
Departamento de Informática

INF-477 Introducción a las Redes Neuronales Artificiales
Cuestionario I. Abril 2018.

1. Dada una tarea $f_0 : X \to Y$ que se desea aprender, hemos visto que una red neuronal construye una hipótesis $f : X \to Y$ como composición de funciones más simples $f_1, f_2, \ldots, f_L$ denominadas *capas*. Discuta cuál es la ventaja de este "approach" con respecto a modelos clásicos de aprendizaje automático (regresión logística ordinaria, SVM, árboles, etc). ¿Cuando recomendaría el uso de una red neuronal en vez de alguno de aquellos modelos?

2. Suponga que un conjunto de datos viene separado en tres subcojuntos: conjunto de entrenamiento (*training set*) de 50.000 ejemplos, conjunto de validación (*validation set*) de 10.000 ejemplos y conjunto de pruebas (*test set*) de 100.000 ejemplos. Si tuviese que elegir entre una red neuronal (feedforward) de arquitectura $1000 \times 100 \times 10$ y otra, de arquitectura $1000 \times 100 \times 100 \times 10$[1], ¿En cuál de los conjuntos evaluaría los modelos? Justifique.

3. ¿Qué sucede si se entrena un perceptron[2] para un problema que no es linealmente separable?

4. Suponga que debe entrenar una red neuronal (feedforward) para clasificar opiniones (textos breves) entre opiniones *positivas* y opiniones *negativas*. Para ello, dispone de un conjunto de n "textos" etiquetados $D = \{(d^{(\ell)}, y^{(\ell)})\}_{\ell=1}^{n}$ donde $d^{(\ell)}$ corresponde a una secuencia/lista de palabras, pertenecientes a un vocabulario $\mathcal{V}$, e $y^{(\ell)} \in \{+1, -1\}$ es la categoría asignada por un comité de humanos. Explique: (i) ¿Qué tipo de capa de salida utilizaría?, (ii) ¿Qué función de entrenamiento seleccionaría para el entrenamiento? (iii) ¿Cómo pre-procesaría antes de entrenar la red?, (iv) ¿Cuáles serían los hiper-parámetros más importantes a seleccionar? y (v) ¿Cómo seleccionaría estos parámetros?

5. ★ La función de pérdida denominada *cross-entropy* es una de las elecciones más comunes en el entrenamiento de redes neuronales artificiales para clasificación. Demuestre que usar esta función resulta equivalente a estimar los pesos de la red usando el método de *máxima verosimilitud*, paradigma clásico en el diseño de modelos estadísticos.

6. Explique la ventaja más importante de utilizar funciones de activación *ReLu* $s(\xi) = \max(0, \xi)$ en el diseño de redes neuronales. ¿Por qué el empleo de esta función de activación suele verse como un modo de dotar a la red de representaciones latentes (ocultas) de largo variable?

7. Considere una red con funciones de activación clásicas (*tanh* y salida *softmax*). ¿Por qué los pesos suelen inicializarse con valores pequeños y aleatorios?

8. Explique cómo se obtiene la inicialización de pesos denominada *Glorot Uniform*, también conocida como *Inicialización de Xavier*, que consiste en asignar un peso aleatorio con distribución

$$U\left(-\sqrt{\frac{6}{n+m}}, \sqrt{\frac{6}{n+m}}\right),$$

   a cada capa con n neuronas de entrada y m neuronas de salida. ¿En qué difiere dicha inicialización de la distribución propuesta por He?.

9. Explique conceptualmente el problema del "desvanecimiento de los gradientes" en redes neuronales profundas.

[1] Como es usual, se está empleando acá la notación $S^{(1)} \times S^{(2)} \times \cdots \times S^{(L)}$ donde $S^{(i)}$ es el número de neuronas de la capa i. El número de neuronas de la capa de entrada es $S^{(1)}$ y el número de neuronas de la capa de salida es $S^{(L)}$.

[2] Entiéndase: una red neuronal feedforward sin capa oculta, con 1 neurona de salida.

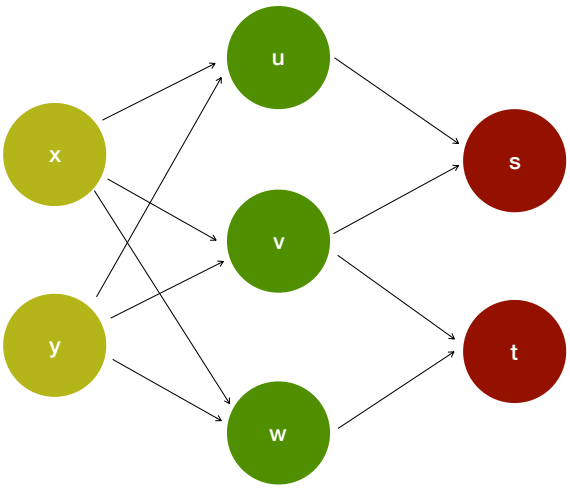

Figure 1: Grafo para la pregunta 12.

10. ★ Explique formalmente el problema del "desvanecimiento de los gradientes" en redes neuronales profundas. Concretamente, derive una cota superior para la magnitud de los gradientes y muestre que ésta decae exponencialmente rápido en la profundidad de la red. Puede hacer las simplificaciones que estime pertinentes para obtener un resultado legible.

11. Explique el significado del *tamaño de batch* (*batch size*) en la implementación moderna del algoritmo BP (back-propagation). ¿Que valor recomendaría si su conjunto de entrenamiento es de 10.000 ejemplos? ¿Que valor recomendaría si su conjunto de entrenamiento es de 1.000.000 de ejemplos?

12. Considere el grafo de computación definido en la figura 1, donde un arco de v_1 a v_2 indican una relación de dependencia directa, es decir, "v_2 requiere v_1". Derive una expresión para la derivada parcial de s con respecto a x como función de las derivadas parciales de s con respecto a u, v y w. ¿Por qué este ejercicio es importante en el contexto de laqs redes neuronales artificiales?

13. ★ Sea $G(\theta)$ una función (real) que depende de cierto vector de parámetros $\theta \in \mathbb{R}^d$ y suponga que deseamos encontrar un mínimo de G. Considere un algoritmo que, partiendo de cierta solución inicial $\theta^{(0)}$, ejecuta la siguiente regla de actualización

$$\theta^{(t+1)} = \theta^{(t)} - \eta_t d_t \,,$$

donde η_t satisface

$$\sum_t \eta_t = \infty, \quad \sum_t \eta_t^2 < \infty \,,$$

y d_t es un vector aleatorio tal que

$$\mathbb{E}(d_t|\theta_t) = \frac{\partial G}{\partial \theta}(\theta^{(t)}), \quad \mathbb{E}(\|d_t\|^2|\theta_t) < \infty \,.$$

Demuestre que este algoritmo converge a un punto estacionario de G en valor esperado. Concretamente, demuestre que

$$\lim_{T\to\infty} \left(\min_{t=1,\ldots,T} \mathbb{E}\left(\frac{\partial G}{\partial \theta}(\theta^{(t)}) \right) = 0 \,.$$

¿Por qué este resultado es importante en el contexto de las redes neuronales artificiales?

14. Explique en qué consiste la técnica denominada *progressive decay* para el entrenamiento de redes neuronales artificiales.

15. Explique la diferencia y el objetivo de las técnicas que hemos denominado *Momentum* y *Momentum de Nesterov* para entrenamiento de redes neuronales artificiales.

16. Explique dos ventajas y una diferencia de las técnicas denominadas *Ada-Delta* y *RMS-Prop* utilizadas en el entrenamiento de redes neuronales artificiales.

17. ¿Qué problema es más relevante en el entrenamiento de redes neuronales: (a) la existencia de múltiples óptimos locales en el espacio de parámetros; o (ii) la existencia múltiples puntos silla en el espacio de parámetros? Explique.

18. Considere un problema en que se desea construir un programa para clasificar automáticamente *reviews*, escritos por personas de una red social, como *positivos* o *negativos* en función de las palabras que contienen. Explique, cuáles son los pasos que habría que considerar para abordar este problema mediante una red neuronal artificial. Indique además, cuántas neuronas de entrada y salida tendría la red, qué funciones activación consideraría para la salida, y qué función de pérdida usaría para entrenar la red.

19. ★ Sea $f(\mathbf{x}) \in \mathbb{R}$ la respuesta de una red neuronal feedforward ante un patrón de entrada $\mathbf{x} \in \mathbb{R}^d$ e y la respuesta deseada. Suponga que la neurona de salida usa una función de activación *sigmoidal* y todas las neuronas escondidas usan funciones de activación *ReLu* $s(\xi) = \max(0, \xi)$. Suponga además que se entrena esta red utilizando la función de pérdida denominada *cross-entropy*, es decir $L(f(x)) = -y \ln(f(x)) - (1-y)\ln(1-f(x))$, utilizando gradiente descendente.

    (a) Demuestre que $L(\cdot)$ es una función convexa de $f(x)$.

    (b) Demuestre que $s(\cdot)$ es una función convexa de ξ.

    (c) Explique por qué, aún L y s son funciones convexas, la red neuronal "puede caer en óptimos locales" durante el entrenamiento. Discuta las consencuencias prácticas de este fenómeno.

20. Considere una red neuronal convolucional alimentada con imágenes RGB de 32×32 pixeles. Suponga que a la capa de entrada sigue un capa convolucional con 16 filtros de 4×4 (stride 1), una capa de pooling con filtros de 2×2 (stride 2), otra capa convolucional con 32 filtros de 4×4 (stride 1), otra capa de pooling con filtros de 2×2 (stride 2) y finalmente un MLP con 256 neuronas ocultas y 10 neuronas de salida. Ilustre las transformaciones que sufre un patrón de entrada al pasar por cada capa y determine el número total de parámetros de la red. Suponga que en ninguna de las capas se implementa padding y que todas las convoluciones involucradas son válidas.

21. ★ ¿Es cierto o falso que una capa convolucional de la forma

$$a^{(i)} = \sigma_i(W^{(i)} * a^{(i-1)} + b^{(i)}),$$

    puede implementarse usando una capa feedforward clásica de la forma

$$a^{(i)} = \sigma_i(W^{(i)} \cdot a^{(i-1)} + b^{(i)}),$$

    mediante un re-dimensionamiento de la entrada y la salida? Si su respuesta es afirmativa, ¿Qué forma tiene la matriz $W^{(i)}$? ¿Cuántos parámetros libres tiene con respecto a una matriz de pesos tradicional? Hint: recuerde la definición de una matriz de *Toeplitz*.

22. ¿Es cierto o falso que una convolución "válida" se puede obtener a partir de una convolución "completa" usando padding y/o recortando parte del resultado obtenido? Explique.

23. ¿Es cierto o falso que una convolución "completa" se puede obtener a partir de una convolución "válida" usando padding y/o recortando parte del resultado obtenido? Explique.

RNA+CVV LaTeX