Universidad Técnica Federico Santa María
Departamento de Informática

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

1. Explique qué distingue un modelo de aprendizaje supervisado de uno no supervisado, indicando al menos 3 aplicaciones de este último.

2. Explique qué distingue un modelo de aprendizaje generativo de uno discriminativo, indicando al menos 1 aplicación de este último.

3. Explique brevemente qué es un auto-encoder. ¿Es cierto o es falso que para propósitos prácticos sólo es relevante el encoder y que el decoder se mantiene activo sólo durante la fase entrenamiento? Justifique.

4. ¿Es cierto o es falso que el sub-espacio extraído por PCA es equivalente (hasta transformaciones lineales) al espacio latente aprendido por un auto-encoder lineal? ¿Depende esto de cómo se mida el error de reconstrucción?

5. Sea $p(x, z)$ una distribución de probabilidad válidamente definida sobre el par de variables (x, z). ¿Es cierto o falso que para cualquier distribución de probabilidad $Q(z)$ vale la siguiente desigualdad?

$$\log\left(\sum_z p(x,z)\right) \leq \sum_z Q(z)\log\left(\frac{p(x|z)p(z)}{Q(z)}\right) \tag{1}$$

   Si su respuesta es negativa, corrija la desigualdad. En cualquier caso, justifique su respuesta, indicando además en qué escenario resulta tan utilizado el resultado obtenido.

6. Explique en qué consiste el algoritmo de Gibbs y cómo viene utilizado durante el entrenamiento de RBMs. ¿Está asegurada su convergencia? ¿Cuál es su ventaja sobre un método más genérico como el algoritmo de Metropolis?

7. ¿Cuál es la simplificación fundamental que hace el algoritmo CD-k para entrenar RBMs? (la abreviación CD corresponde al término *contrastive divergence*).

8. ★ Considere el modelo $p(x, z) = \exp(-E(x, z))/Z$, con $E(x, z) = -(z^T Wx + b^T z + c^T x)$, $x \in \mathbb{R}^d$, $z \in \{0, 1\}^d$, definido por una RBM. Demuestre que

$$p(z = 1|x) = \frac{\exp(Wx + b)}{1 + \exp(Wx + b)} = \sigma(Wx + b)\,, \tag{2}$$

9. ★ Considere una RBM como aquella definida en la pregunta 8 y la función de log-verosimilitud asociada a un punto $x \in \mathbb{R}^d$, $\ell(W) = \ln p(x|W)$, donde W representa los parámetros del modelo. Demuestre que

$$\frac{\partial \ell(W)}{\partial W_{ji}} = \mathbb{E}_{h|x} h_j x_i - \mathbb{E}_{x,h} h_j x_i \tag{3}$$

   ¿Porqué el segundo término es mucho más difícil de calcular que el segundo? ¿Como se aborda este problema en el algoritmo CD-k?

10. Considere un auto-encoder donde el mecanismo generador se modela como $z \sim p_\theta(z)$ y luego $x \sim p_\theta(x|z)$. ¿Porqué resulta difícil calcular p_θ(z—x)? ¿Cómo se aborda este problema en el diseo de auto-encoders variacionales (VAEs)?

11. Explique en qué consiste una capa convolucional traspuesta (transposed convolution) y cómo es utilizada en el diseño de modelos generativos de datos visuales.

12. Considere un auto-encoder variacional donde el encoder se modela mediante la distribución $q_\phi(z|x)$, el decoder mediante la distribución $p_\theta(x|z)$ y el a-priori sobre el espacio latente mediante la distribución $p_\theta(z)$. ¿Porqué resulta conveniente elegir $q_\phi(z|x)$ y $p_\theta(z)$ gausianas?

13. Explique la diferencia entre la divergencia Kullback-Leibler $KL(p||q)$ y la divergencia de Jensen-Shannon $JS(p||q)$. ¿Cómo trata cada métrica los puntos x a los que $p(x)$ asigna probabilidad 0? Reflexione sobre las consecuencias de esta observación.

14. Considere un auto-encoder variacional como el de la pregunta 12. Como hemos visto, entrenar un auto-encoder variacional resulta equivalente a maximizar la siguiente función objetivo,

$$J(\phi, \theta) = \sum_\ell p_\theta(x^{(\ell)}) - KL(q_\phi(z|x^{(\ell)})||p_\theta(z|x^{(\ell)})) \tag{4}$$

    Mencione una consecuencia positiva y una consecuencia negativa de cambiar la divergencia KL por una divergencia de Jensen-Shannon.

15. Considere un auto-encoder variacional como el de la pregunta 12. Como hemos visto, entrenar un auto-encoder variacional resulta equivalente a maximizar

$$J(\phi, \theta) = \mathbb{E}_{x\sim p(x)} p_\theta(x^{(\ell)}) - KL(q_\theta(z|x^{(\ell)})||p_\theta(z|x^{(\ell)}))\,, \tag{5}$$

    donde $p(x)$ es la distribución real de los datos. Demuestre que esta función objetivo es equivalente a maximizar

$$J'(\phi, \theta) = \mathbb{E}_{z\sim q_\phi(z|x)\log p_\theta(x|z)} - KL(q_\theta(z|x^{(\ell)})||p_\theta(z)) \tag{6}$$

    Interprete cada miembro de esta última función objetivo.

16. Considere un auto-encoder variacional como el de la pregunta 12 y la función objetivo $J'(\phi, \theta)$ definida en la pregunta 15. Explique cómo se calculan las derivadas de $J'(\phi, \theta)$ en función de ϕ y θ en modo de obtener estimadores estables. Explique la relevancia de este "truco" (denominado en ocaciones el "reparametrization trick") en el éxito experimental de los VAE.

17. Explique en qué consiste un "juego a suma cero" y un "equilibrio de Nash" en la teoría clásica de juegos.

18. ¿Cuáles son los componentes fundamentales en el modelo GAN (generative adversarial training)? ¿Cuál es el rol de cada componente? ¿Es correcto afirmar que en este modelo, el resultado del aprendizaje representa un "equilibrio de Nash" para cierto juego continuo a suma cero? Si esto último es correcto, cuál es la función de valor correspondiente.

19. ★ Considere un juego continuo a suma cero, con dos jugadores, espacios de acciones $X = Y = \mathbb{R}$ y función de valor $U = xy$, $x \in X$, $y \in Y$. Suponga que el primer jugador (que desea maximizar U) actualiza iterativamente su estrategia (x) de la forma $x \leftarrow x + \frac{\partial U}{\partial x}$ y que el segundo jugador (que desea minimizar U) actualiza iterativamente su estrategia (y) de la forma $y \leftarrow y - \frac{\partial U}{\partial x}$. Es claro que un punto silla (equilibrio) de U es el punto $(x^*, y^*) = (0, 0)$. Muestre que si modelamos las actualizaciones ejecutadas por los jugadores de manera continua en el tiempo t jamás se converge a (x^*, y^*). ¿Qué sugiere este resultado?

    Hint: Si modelamos las actualizaciones ejecutadas por los jugadores de manera continua en el tiempo t, obtenemos las siguientes ecuaciones diferenciales:

$$\frac{\partial x}{\partial t} = \frac{\partial V}{\partial x},\ \frac{\partial y}{\partial t} = -\frac{\partial V}{\partial x}\ . \tag{7}$$

20. ★ Considere una GAN con discrminador $D(x)$ y generador $p_g(x) = G(z)$, $z \sim p_z(z)$. Demuestre que si asumimos que el discriminador puede siempre implemnentar su solución óptima, entrenar el generador es equivalente a minimizar

$$C(G) = 2 \cdot \mathrm{JS}(p(x)||p_g(x)) - \log(4)\,.$$

donde $\mathrm{JS}(p(x)||q(x))$ es la divergencia de Jensen-Shannon entre $p(x)$ y $q(x)$.

21. ★ Considere una GAN con discrminador $D(x)$ y generador $p_g(x) = G(z)$, $z \sim p_z(z)$. Demuestre que para cualquier discriminador fijo $D^*(x)$, la función objetivo del lado del generador

$$\min_G C_{D^*}(G)\,, \tag{8}$$

con

$$C_{D^*}(G) = \mathbb{E}_{x\sim p(x)}\left[\log D^*(x)\right] + \mathbb{E}_{z\sim p_z(z)}\left[\log(1 - D^*(G(z)))\right] \tag{9}$$

acepta como solución óptima cualquier función $G(\cdot)$ que mapee $z \sim p_z$ a alguna de las modas de $D^*(x)$. Reflexione sobre las implicancias de este resultado.

22. Explique cómo modelaría el problema de entrenar una red-convolucional para contar el número de personas en una imagen.