Universidad Técnica Federico Santa María
Departamento de Informática

INF-393 Aprendizaje Automático

Cuestionario I 2018-II Campus San Joaquín

Los ejercicios marcados con (★) o (★★) requieren un nivel mayor de profundización y por lo tanto tendrían mayor puntaje en un eventual examen.

1. Explique qué se entiende por sobreajuste (*overfitting*), porqué se produce y cómo puede prevenirse.

2. Explique el rol del conjunto de entrenamiento (*training set*), conjunto de validación (*validation set*) y conjunto de pruebas (*test set*) en el diseño de una solución basada en aprendizaje automático.

3. Explique qué se entiende por *aprendizaje supervisado*, *aprendizaje no-supervisado* y *aprendizaje semi-supervisado*.

4. Explique en qué consiste la técnica de *validación cruzada* (k-fold cross validation). Escriba un pseudo-código que implemente esta técnica para la selección de un hiper-parámetro genérico.

5. Considere un algoritmo de aprendizaje que utiliza un espacio de hipótesis A y otro que utiliza un espacio de hipótesis B. Suponga que A es un subconjunto estricto de B ¿Es cierto o es falso que el primero tendrá mejor capacidad de generalización? Justifique.

6. Considere un modelo lineal de la forma $y = f(x) + \epsilon$, con $f(x) = w^T x + b$, $x \in \mathbb{R}^d$ y $d = 1000000$. Explique porqué, pese a su simplicidad, el aprendizaje de este modelo desde un conjunto de ejemplos $S = \{(x^{(\ell)}, y^{(\ell)})\}_{\ell=1}^{n}$, $n = 100$, es susceptible de sobre-ajuste (*overfitting*). ¿Cómo lo podría prevenir?

7. (★) Considere dos variables aleatorias continuas $x \in \mathbb{R}^d$ e $y \in \mathbb{R}$. Demuestre que si se minimiza el error de entrenamiento usando la función de pérdida cuadrática $L(f(x), y) = (f(x) - y)^2$ sobre un espacio de hipótesis y un conjunto de ejemplos arbitrariamente grandes, la solución es exactamente la función de regresión $f(x) = E[y|x]$.

8. (★) Considere dos variables aleatorias $x \in \mathbb{R}^d$ e $y \in \{1, 2, \ldots, K\}$. Demuestre que si se minimiza el error de entrenamiento usando la función de pérdida $L(f(x), y) = (I(f(x) \neq y))$ sobre un espacio de hipótesis y un conjunto de ejemplos arbitrariamente grandes, la solución es exactamente el clasificador de Bayes $f(x) = \arg\max_y P(y|x)$.

9. (★★) Considere dos variables aleatorias $x \in \mathbb{R}^d$ e $y \in \mathbb{R}$. Demuestre que si se minimiza el error de entrenamiento usando la función de pérdida (denominada tilted loss)

$$L_\tau(f(x), y) = \begin{array}{ll} \tau\,(y - f(x)) & y - f(x) \geq 0 \\ (\tau - 1)\,(y - f(x)) & y - f(x) < 0 \end{array},$$

   $\tau \in (0, 1)$, sobre un espacio de hipótesis y un conjunto de ejemplos arbitrariamente grandes, la solución es exactamente el τ-ésimo percentil de la distribución $y|x$, $f(x) = F^{-1}_{y|x}(\tau)$.

10. (★) Demuestre la convergencia del algoritmo de clasificación que hemos denominado *perceptrón*.

11. Discuta al menos 1 ventaja y 1 desventaja de regularizar el aprendizaje de un modelo lineal usando la norma ℓ_2 (e.g. "ridge regression") versus la norma ℓ_1 (e.g. "lasso").

12. (★) Considere problema de aprendizaje donde la función de entrenamiento es de segundo orden (e.g. regresión lineal o logística) como función de los parámetros θ. Demuestre que en el caso de una matriz Hesiana diagonal $H = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_d^2)$, la relación entre la solución no regularizada θ y la solución obtenida regularizando con la norma ℓ_1 (e.g. "lasso") θ_λ es la siguiente $\theta_\lambda = S_\gamma(\theta)$, donde $S_\gamma(\theta) = \text{sign}(\theta)(|\theta| - \gamma)$, con $\gamma = \lambda/\sigma_i^2$. ¿Cómo se generaliza esta conclusión al caso de matrices Hesianas arbitrarias?

13. Explique brevemente en qué consiste el supuesto denominado "Bayesiano ingenuo" (*naive Bayes*) en el diseño de clasificadores generativos.

14. Explique brevemente qué elecciones o simplificaciones distinguen a LDA (*linear discriminant analysis*) de QDA (*quadratic discriminant analysis*). Muestre que las primeras implican fronteras de clasificación lineales, mientras que las segundas implican fronteras de clasificación de segundo orden. Discuta también las ventajas y desventajas esta diferencia.

15. ¿Es cierto o es falso que el modelo logístico obtiene fronteras de clasificación lineales? Explique.

16. ¿Es cierto o es falso que el modelo logístico es intrínsecamente binario? Si no es así, escriba las ecuaciones que definen el modelo en el caso de múltiples categorías, indicando cuántos parámetros libres (aprendibles) existen.

17. Proponga un método para regularizar el aprendizaje del clasificador logístico que permita obtener matrices de covarianza dispersas.

18. Explique brevemente la ventaja de enfoque generativo versus uno discriminativo en el diseño de un clasificador.

19. Suponga que se desea entrenar un clasificador Bayesiano ingenuo a partir de un conjunto de entrenamiento $S_E = \{(x^{(\ell)}, y^{(\ell)})\}_{\ell=1}^n$, con $y^{(\ell)} \in \{0, 1\}$ y $x_i^{(\ell)} \in \{0, 1\}$, $\forall i$. Encuentre valores para los parámetros que maximicen la función de log-verosimilitud correspondiente al modelo.

20. Considere un conjunto de entrenamiento $S_E = \{(x^{(\ell)}, y^{(\ell)})\}_{\ell=1}^n$ recolectado para el entrenamiento de un clasificador ¿Qué significa que S_E sea linealmente separable?

21. ¿En qué consiste la estrategia ECOC en el diseño de clasificadores?

22. (★) Suponga que se está evaluando la aplicación de la estrategia *one-versus-one* (OVO) frente a la estrategia *one-versus-the-rest* (OVR) para extender un clasificador binario al caso de múltiples clases. ¿Es cierto que OVO construirá más clasificadores que OVR? ¿Es cierto que OVO resultará siempre más lento que OVR? Justifique considerando la complejidad de entrenar el clasificador como función del número de datos m.

23. ¿Es cierto o es falso que si resolvemos el problema de regresión lineal mediante gradiente descendente (GD) podríamos encontrar óptimos locales? Asuma primero que se utiliza como función de pérdida (*loss function*) la función $L(f(x), y) = (f(x) - y)^2$. ¿Cambia su respuesta si se utiliza como función de pérdida la función $L(f(x), y) = \big((f(x) - y)^2 - \epsilon\big)_+$ con $\epsilon > 0$? ¿Cambia su respuesta si el problema se resuelve usando gradiente descendente estocástico (SGD)? Justifique.

24. Suponga que tenemos dos algoritmos iterativos para encontrar la solución θ^* de un cierto problema de aprendizaje. Si el primer algoritmo es dos veces más rápido que el otro en realizar una iteración, ¿Es cierto que éste terminará antes que el segundo? Asuma que el criterio de parada consiste en encontrar una solución aproximada θ_ϵ de la misma calidad, $\theta_\epsilon \in \mathcal{B}(\theta^*, \epsilon)$?

25. (★) Sea $I(X, Y)$ la información mutua correspondiente a dos variables X e Y. Demuestre que $I(X, Y) = H(X) + H(Y) - H(X, Y)$ donde $H(Z)$ es la entropía de Z.

26. ¿Cuál es la ventaja de la información mutua con respecto al $Z - score$ como criterio de filtrado para selección de atributos?

27. ¿Cuál es la diferencia entre un método de filtrado y un método tipo wrapper para selección de atributos? ¿Cuál es la diferencia de un método embedido con respecto a estos dos enfoques? De un ejemplo de cada caso.

28. (★★) Considere un problema de aprendizaje donde la función objetivo de entrenamiento $J(\theta)$ es convexa. Demuestre que, si se elige la constante de entrenamiento apropiadamente, el algoritmo de gradiente descendente converge al mínimo de $J(\theta)$. ¿Qué significa elegir la constante de entrenamiento apropiadamente?

29. Considere un problema de aprendizaje donde la función objetivo de entrenamiento $J(\theta)$ es convexa. ¿Qué significa que el algoritmo de gradiente descendente estocástico aplicado a este problema resulte no monótono? ¿Cómo se puede manejar este problema en la práctica?

30. Explique la diferencia entre reducción de dimensionalidad y selección de atributos. ¿Cuál sería una ventaja del primer enfoque con respecto al segundo? ¿Una desventaja importante?

31. (★) Considere la utilización del $Z-score$ y del $F-score$ para hacer un *ranking* de los atributos asociados a un conjunto de datos de entrenamiento. Demuestre que ambos criterios de filtrado producirán exactamente el mismo resultado. Construya además un ejemplo que muestre que si tenemos una variable X_1 estadísticamente dependiente de la respuesta Y y otra variable X_2 estadísticamente independiente de la respuesta, cualquiera de los dos criterios podría ordenar incorrectamente los atributos.

32. (★) Considere la divergencia de Kulback-Leibler $KL(p||q)$ y la divergencia de Jensen-Shannon $JS(p||q)$. Demuestre que $KL(p||q) = 0$ si y sólo si $p(x) = q(x)$, excepto en puntos tales que $p(x) = 0$, mientras que $JS(p||q) = 0$ si y sólo si $p(x) = q(x)$, excepto en puntos tales que tanto $p(x) = 0$, como $q(x) = 0$. Utilice esta observación para argumentar porqué sería inconveniente aproximar una distribución q a partir de un modelo p, minimizando $KL(p||q)$ en vez de $KL(q||p)$.

33. (★★) Considere un modelo estándar de regresión lineal $y = f(x) + \epsilon$, con $f(x) = w^T x + b$, $x \in \mathbb{R}^d$, entrenado mediante mínimos ordinarios a partir de un conjunto de entrenamiento $S_E = \{(x^{(\ell)}, y^{(\ell)})\}_{\ell=1}^n$, y denotemos por $R_{tr}(\hat{w}, \hat{b})$ el error de entrenamiento obtenido, es decir, $R_{tr}(\hat{w}, \hat{b}) = \frac{1}{n}\sum_\ell (f(x^{(\ell)}) - y^{(\ell)})^2$. Suponga que luego se recolecta un conjunto de test $S_T = \{(\tilde{x}^{(\ell)}, \tilde{y}^{(\ell)})\}_{\ell=1}^m$, de manera tal que $(\tilde{x}^{(\ell)}, \tilde{y}^{(\ell)})$ tiene distribución idéntica a la de $(x^{(\ell)}, y^{(\ell)})$, y medidos el error del modelo en este conjunto,$R_{ts}(\hat{w}, \hat{b}) = \frac{1}{m}\sum_\ell (f(\tilde{\mathbf{x}}^{(\ell)}) - \tilde{y}^{(\ell)})^2$.

    Indicando los supuestos necesarios, demuestre que: $\mathbb{E}[R_{tr}(\hat{w}, \hat{b})] \leq \mathbb{E}[R_{ts}(\hat{w}, \hat{b})]$.

34. (★★) Considere un modelo de regresión logística de la forma $p(y = 1) = \sigma(f(x))$ con $\sigma(\xi) = (1 + \exp(-\xi))^{-1}$, $f(x) = wx + b$, $x, w, b \in \mathbb{R}^1$, $y \in \{0, 1\}$, que se desea entrenar partir de un conjunto de entrenamiento $S_E = \{(x^{(\ell)}, y^{(\ell)})\}_{\ell=1}^n$. Demuestre que si los ejemplos en S_E son linealmente separables, la función de log-verosimilitud $\ell(w, b) = \sum_\ell y^{(\ell)}\sigma(f(x^{(\ell)})) + (1 - y^{(\ell)})(1 - \sigma(f(x^{(\ell)})))$ no es acotada y por lo tanto el eventual entrenamiento del modelo basado en este criterio no converge.

35. (★★) Considere un modelo estándar de regresión lineal $y = f(x) + \epsilon$, con $f(x) = w^T x + b$, $x \in \mathbb{R}^d$, que se desea ajustar a partir de un conjunto de entrenamiento $S_E = \{(x^{(\ell)}, y^{(\ell)})\}_{\ell=1}^n$. Suponga que, en vez de entrenar el modelo directamente con los ejemplos en S_E, primero se inyecta ruido a cada patrón de entrada, es decir, se construye el conjunto $S_E = \{(\tilde{x}^{(\ell)}, y^{(\ell)})\}_{\ell=1}^n$, con $\tilde{x}^{(\ell)} = x^{(\ell)} + \epsilon^{(\ell)}$, $\epsilon^{(\ell)} \sim \mathcal{N}(0, \gamma I)$, $\gamma > 0$ [1], y luego se minimiza el error cuadrático medio

$$E_S = \frac{1}{2}\sum_{n=1}^{n}(f(\tilde{x}^{(\ell)}) - y^{(\ell)})^2.$$

    Demuestre que, en valor esperado, el resultado obtenido es equivalente a entrenar el modelo usando Ridge Regression para un valor del parámetro de regularización que es inversamente proporcional a γ^2.

RNA+CVV LaTeX

[1] $I \in \mathbb{R}^{d\times d}$ es la matriz identidad.

[2] Asuma que $\epsilon^{(\ell_1)}$ es independiente $\epsilon^{(\ell_2)}$, $\forall \ell_1 \neq \ell_2$.