Universidad Técnica Federico Santa María
Departamento de Informática

INF-393 Aprendizaje Automático

Cuestionario II 2018-II Campus San Joaquín

Los ejercicios marcados con (★) requieren un nivel mayor de profundización.

1. ¿Cuál es el rol del parámetro C en una C-SVM? Si nos interesa obtener un menor error de entrenamiento, ¿Deberíamos usar un valor mayor o menor de C?

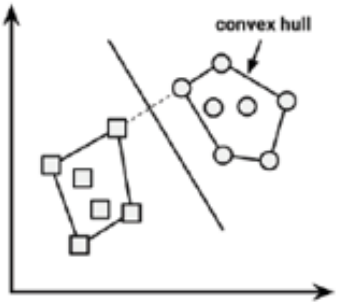
2. Verdadero o falso: "A un mayor valor de C, corresponde una mayor cantidad de vectores de soporte". Justifique.

3. ¿Cuál es la justificación teórica para considerar una extensión del concepto clásico de margen (*hard margin*) que "ignore" parte de los datos de entrenamiento (*soft margin*)?

4. Explique en qué consiste el denominado *kernel trick*.

5. ¿Cómo se puede extender una SVM a problemas de múltiples categorías? Proponga una reformulación del problema que no pase por descomponerlo en sub-problemas binarios.

6. ★ Demuestre que entrenar una SVM de margen rígido (*hard margin*) es equivalente a encontrar la distancia más corta entre las envolturas convexas correspondientes a cada una de las clases. La envoltura convexa de un conjunto de puntos $\{x^{(\ell)}\}_{\ell=1}^{n}$, es el conjunto de puntos $\{\sum_\ell \alpha_\ell x^{(\ell)} : \sum_\ell \alpha_\ell = 1, \alpha_\ell \geq 0\}$.



7. ★ Considere un problema de optimización convexo con restricciones lineales. ¿Es cierto o es falso que la solución del denominado *problema dual* (en el sentido de Lagrange) provee una cota inferior a la solución del problema original?

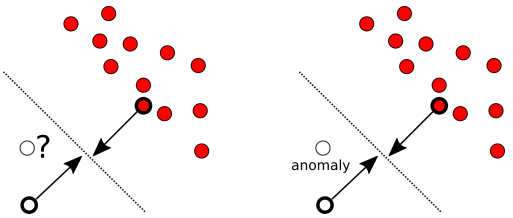
8. ★ Demuestre que el kernel gausiano $K(x,z) = e^{\frac{-||x-z||^2}{\sigma^2}}$, con $\sigma^2 > 0$ es un kernel válido.

9. Si $K_1(x,z)$ y $K_2(x,z)$ son kernels admisibles
   (a) Demuestre que $K(x,z) = K_1(x,z) + K_2(x,z)$ es kernel admisible.
   (b) Demuestre que $K(x,z) = K_1(x,z)K_2(x,z)$ es kernel admisible.

10. ★ Considere un conjunto no etiquetado de puntos $\{x^{(\ell)}\}_{\ell=1}^{n}$, $x^{(\ell)} \in \mathbb{R}^d$ y un modelo de detección anomalías de la forma $f(x) = \text{sign}(w^T x - \rho)$ donde $f(x) > 0$ se interpreta como "ausencia de anomalía". Suponga que se entrena f resolviendo el siguiente programa convexo

$$\min_{w,b} \|w\|^2 + C \sum_{\ell} \xi_\ell - \rho \tag{1}$$
$$\text{s.t. } w^T x^{(\ell)} \geq \rho - \xi_\ell,\ \xi_\ell \geq 0\ \forall \ell\,.$$

Construya una versión no-lineal del detector empleando el *kernel trick*.



11. Hemos visto que un árbol de clasificación implementa una hipótesis de la forma $f(x) = \sum_j c_j R_j(x)$ donde $R_j(x)$ es la función característica de una región (conjunto de puntos) $\mathcal{R}_j$ del espacio característico. Mencione 4 consecuencias (positivas y/o negativas) de utilizar regiones con overlap, es decir $\mathcal{R}_i \cap \mathcal{R}_j \neq \emptyset$.

12. Proponga un criterio que aproveche la estructura de un árbol de clasificación para generar un ranking de atributos. Extienda su propuesta asumiendo que dispone de un conjunto de árboles (bosques) construidos sobre diferentes muestras de datos.

13. ¿De qué forma práctica se suele controlar la tendencia a *overfitting* de un árbol de clasificación/regresión?

14. Considere un problema de clasificación con datos representados usando D atributos binarios. ¿Cuántos árboles de clasificación binarios balanceados distintos es posible construir?

15. ★ ¿Cuál es la motivación para considerar la métrica denominada *Gini impurity* en el aprendizaje de árboles de clasificación? ¿Es cierto o es falso que la distribución que minimiza esta medida es única? ¿Es cierto o es falso que la distribución que minimiza esta medida es la distribución condicional $p(y|x \in R)$? (R es la región correspondiente a la hoja donde esta medida se está calculando.

16. ★ Suponga que se desea transformar cierta hoja t de un árbol de regresión en un nodo interno con dos hijos t_R, t_L. Suponga que se desea implementar esta modificación usando una variable categórica X_j con B posibles valores (reales). ¿Cuántas formas diferentes existen de efectuar la división? ¿Como habría ordenar los valores de X_j si desea considerar sólo particiones de la forma $t_L = \{x \in t : x_j < s\}$, $t_R = \{x \in t : x_j \geq s\}$ con $s \in \mathbb{R}$? Asuma que con la modificación se desea reducir el valor de cierta función objetivo convexa como el error cuadrático.

17. Discuta brevemente los posibles problemas de usar una tasa de aprendizaje (*learning rate*) demasiado alta, o demasiado baja en el entrenamiento de una red neuronal con *back-propagation*. Mencione alguna alternativa para abordar este problema.

18. ★ ¿Es cierto o es falso que una red neuronal de tipo feed-forward con sólo 1 capa oculta puede aproximar arbitrariamente bien cualquier función continua? Vale este resultado independientemente de la función de activación utilizada.

19. Describa brevemente la diferencia entre entrenar una red neuronal usando backpropagation por lote (batch) o mini-lotes (mini-batches). ¿Cuál es la justificación teórica para esta última modalidad?

20. Diseñe una pequeña red neuronal que implemente la función XOR. Escriba explícitamente los valores de los parámetros.

21. ¿En qué problemas consideraría el uso de redes convolucionales en vez de redes feed-forward clásicas?

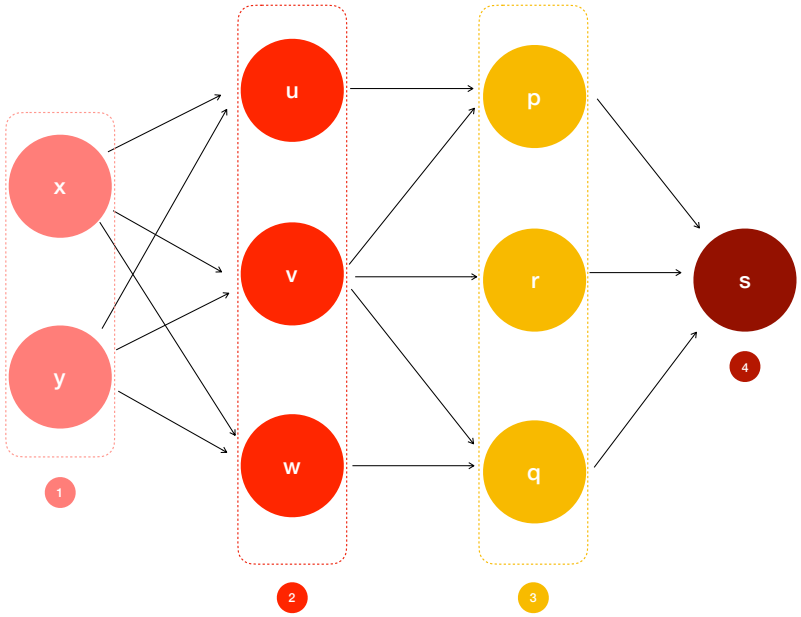


Figure 1: Grafo para la pregunta 22.

22. 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". Suponga que cada nodo multiplica los inputs que recibe. Derive una expresión explícita para la derivada parcial de s con respecto a x. Indique cómo se llevaría cabo este cálculo usando back-propagation y la ventaja computacional que representa.