Universidad Técnica Federico Santa María
Departamento de Informática

INF-393 Máquinas de Aprendizaje

PAUTA Quiz 1.

Se ruega al lector estudioso que comunique al profesor cualquier error que descubra en este documento.

1. *Considere dos variables aleatorias X e Y con f.d.p. conjunta $p(x, y)$ y suponga que $Y \in \{c_1, c_2\}$. Explique qué es el* Error de Bayes *en* Teoría de Decisión *y escriba una ecuación que permita calcularlo cuando $p(x, y)$ es conocida.*

   El error de Bayes es el menor error de clasificación que es posible obtener sobre datos generados de acuerdo a $p(x, y)$. Si fijamos el valor de x, este viene dado por

   $$B(x) = \min_{f(x)\in\mathbb{Y}} \sum\nolimits_y I(f(x) \neq y)p(y|x) = 1 - \max_y p(y|x)\,. \tag{1}$$

   El error de Bayes es el valor esperado de $B(x)$ en $p(x)$, es decir,

   $$B^* = \sum\nolimits_x B(x)p(x)\,. \tag{2}$$

2. *Explique brevemente la diferencia entre selección de atributos, regularización y reducción de dimensionalidad.*

   Selección de atributos es un problema combinatorial que consiste en encontrars el subconjunto de los atributos originales que maximiza alguna función objetivo de interés (típicamente el error de predicción). Este problema se puede relajar como el problema de obtener un *ranking* de los atributos originales. Con los métodos de regularización clásicos (e.g. Tikhonov) los atributos pueden ser "parcialmente ignorados" podando el valor de los coeficientes que acompañarían a tales atributos en un modelo no regularizado. Intuitivamente, se trata de un método "continuo" de selección. El nivel de poda a aplicar sobre cada atributo se determina durante el entrenamiento del modelo, no como una tarea separada, indicando el nivel global de poda deseado (parámetro de regularización). Esto último se puede especificar como un "penalty" en la f.o. de entrenamiento o como una restricción a satisfacer. En reducción de dimensionalidad, se produce una transformación de los atributos originales x, que en el caso más simple es de la forma $z = Ax$. Esto implica la síntesis de nuevos atributos a costas de la interpretabilidad del modelo resultante.

3. *¿Cuál es la ventaja (teórica) de la información mutua con respecto al Z-score como función de filtrado para selección de atributos? ¿Cuál es la desventaja más relevante?*

   Permite detectar relaciones de dependencias no lineales, pero es mucho más costosa de computar.

4. *¿Cuál es la diferencia entre PCA y LDA como métodos de reducción de dimensionalidad? Explique brevemente, refiriéndose a la función objetivo optimizada por cada método. ¿Es cierto o es falso que*

LDA es un método no-lineal de reducción de dimensionalidad?

Tanto PCA como LDA son modelos lineales de reducción de dimensionalidad en el sentido de que computan un *embedding* z de la forma $z = Ax$. PCA optimiza la proyección para maximizar la varianza (esperada) de los datos después de la proyección. LDA optimiza la proyección para maximizar el cuociente entre la varianza inter-clases y la varianza intra-clases.

5. *¿Cuál es el efecto esperado regularizar un modelo usando la norma ℓ_1 versus la norma ℓ_2? Asumiendo que la matriz Hessiana de la función objetivo es diagonal, explique la diferencia entre la "poda" de coeficientes que realiza uno u otro método sobre la solución no regularizada.*

Regularizar con la norma ℓ_1 permite obtener soluciones verdaderamente sparse (dispersas), es decir algunos parámetros del modelo realmente se anulan. En un modelo lineal, esto implica que los atributos correspondientes a esos parámetros son completamente ignorados por el modelo regularizado. Regularizar con la norma ℓ_2 sólo permite aproximar este objetivo, obteniéndose valores casi nulos, pero no exactamente nulos. Las variables correspondientes a esos parámetros siguen presentes en el modelo.

Bajo los supuestos indicados, denotando por H la matriz Hessiana de la función objetivo, y denotando por $\hat{\theta}$ la solución no regularizada, tenemos que la regularización ℓ_1 corresponde al siguiente "filtro"

$$T_\gamma^1(\hat{\theta}_i) = \text{sign}(\hat{\theta}_i)\ \max\left(|\hat{\theta}_i| - \frac{\gamma}{H_{ii}}, 0\right)\,, \tag{3}$$

mientras que la regularización ℓ_2 corresponde a aplicar

$$T_\gamma^2(\hat{\theta}_i) = \frac{H_{ii}}{H_{ii}+\gamma}\hat{\theta}_i\,, \tag{4}$$

Esto se aprecia mejor en el siguiente dibujo