Primer examen parcial

Reconocimiento de Patrones (2024-2)

Julio Waissman Vilanova

Nombre: _____

Aprendizaje no supervisado

1. ¿Qué ocurre si seleccionas un valor de K demasiado alto en el algoritmo *K–means*? ¿Cómo afecta esto a la segmentación de los datos y qué problemas podría generar en la interpretación de los resultados?

2. ¿Qué sucede si eliges un valor de ϵ demasiado pequeño en el algoritmo DBSCAN? ¿Cómo afecta esto al número de clusters y a la cantidad de puntos clasificados como ruido?

3. ¿Cuál es la importancia de estandarizar los datos antes de aplicar el Análisis de Componentes Principales, y qué podría suceder si omites este paso?

Aprendizaje PAC

4. Consideremos el modelo de aprendizaje que vamos a llamar *2–intervalos*. En este modelo $\mathcal{H}$ vamos a considerar que:

$$h : \mathbb{R} \to \{-1, +1\},$$

donde $h(x) = +1$ si el punto $x \in \mathbb{R}$ se encuentra dentro de alguno de dos intervalos (i.e. $[a, b]$ y $[c, d]$) preestablecidos. En caso contrario, $h(x) = -1$.

   - ¿Cual es el *breakpoint* de $\mathcal{H}$?
   - ¿Que valor tiene $d_{VC}(\mathcal{H})$?

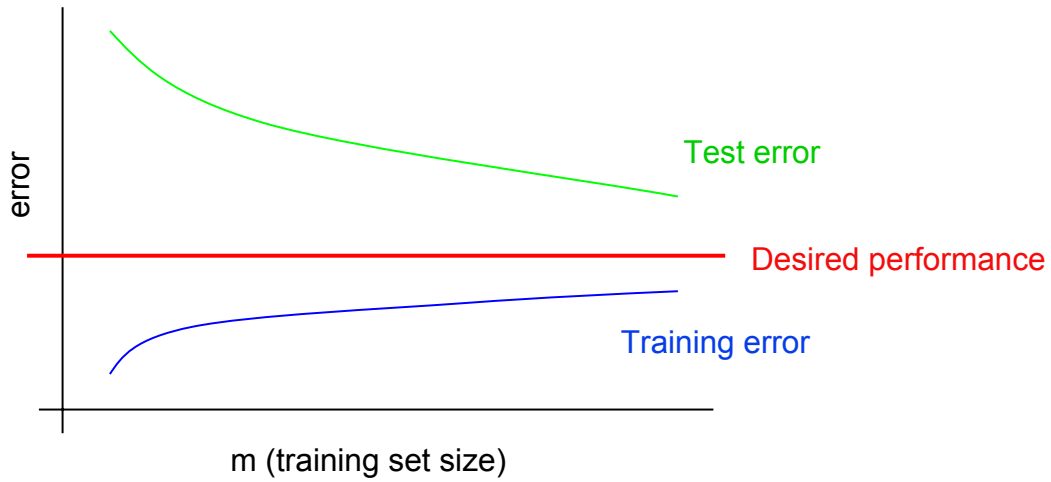
5. Ahora consideremos el caso genérico *M–intervalos*, donde $\mathcal{H}$ está definido como el conjunto de funciones $h : \mathbb{R} \to \{-1, +1\}$ tales que $h(x) = +1$ si x se encuentra en alguno de los M intervalos establecidos y $h(x) = -1$ en caso contrario.

   - ¿Que valor tiene $d_{VC}(\mathcal{H})$?
   - Si fueras a decidir utilizar un modelo 5–intervalos, ¿Cuantos datos necesitarías como mínimo en tu conjunto de aprendizaje para asegurar la generalización?

6. Define con tus propias palabras que significa *Probablemente Aproximadamente Correcto (PAC)*

Indicadores de desempeño

7. Supongamos que tenemos un problema de inspección de piezas en una línea de producción. Se cuenta con un conjunto de datos de 1000 piezas, de las cuales 950 son piezas buenas y 50 son piezas defectuosas. Se utilizó un clasificador por regresión logística con una expansión polinomial de grado 2. Al generar la curva de aprendizaje obtenemos algo similar a la curva siguiente:

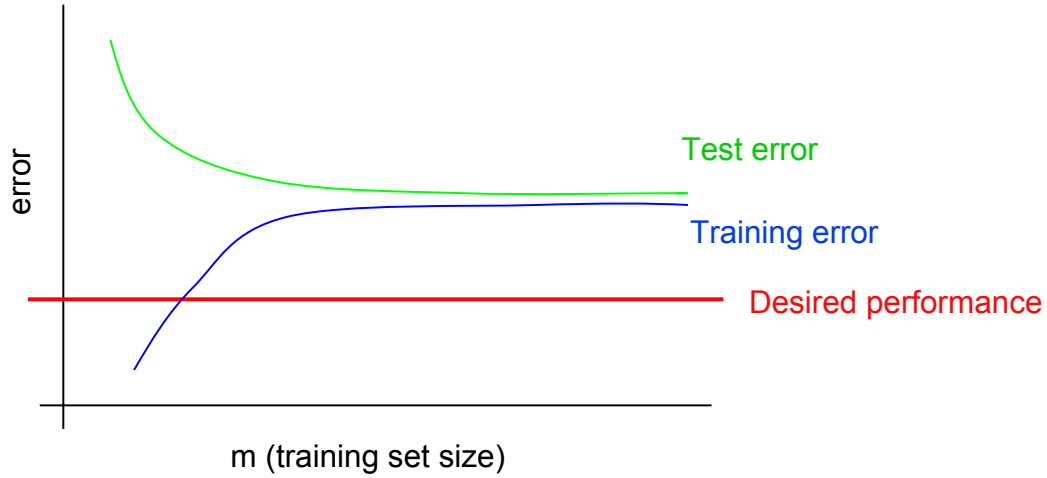


   subraya las acciones que podrían mejorar al modelo de aprendizaje.

   - Solicitar la clasificación de más piezas.
   - Probar sin expansión polinomial.
   - Probar con una expansión polinomial de orden 3.
   - Usar PCA primero y quedarse con solo los componentes que expliquen el 95 % de la varianza.
   - Aumentar el valor de λ (parámetro de regularización).
   - Disminuir el valor de λ (parámetro de regularización).
   - Utilizar una SVM con kernel gaussiano.

8. Supongamos que estamos estimando la demanda de energia eléctrica doméstica en la Cd. de Hermosillo para el próximo día, utilizando como información el consumo de energía eléctrica de los 30 días anteriores, la temperatura máxima en Hermosillo de los 30 días anteriores, la temperatura mínima en Hermosillo de los 30 días anteriores, el día de la semana, una variable que indica si el día es festivo o no y una variable que indica la estación del año (invierno, primavera, verano y otoño). Se aplica un método de regresión lineal con la información de los últimos 5 años.

   Para analizar el desempeño del algoritmo de regresión lineal, se realiza una curva de aprendizaje la cual resulta ser de la forma siguiente:



subraya las acciones que podrían mejorar al modelo de aprendizaje.

- Solicitarle a CFE información de otros 5 años anteriores.
- Disminuir el valor de λ (parámetro de regularización).
- Aumentar el valor de λ (parámetro de regularización).
- Utilizar solo la información histórica de los últimos 15 días y no de los 30 días anteriores.
- Utilizar una red neuronal en lugar de la regresión lineal.
- Agregar como atributos la raiz cuadrada de la demanda de energía eléctrica de los 30 días anteriores y la raiz cuadrada de los valores máximos y mínimos de temperatura de los 30 días anteriores.
- Agregar la humedad relativa de los 30 días anteriores.

9. Sea la siguiente matriz de confusión, resuelta después de utilizar un método de aprendizaje para clasificar datos de un problema real:

		y	
		0	1
$h_\theta(x)$	0	300	5
	1	10	30

Responde a las siguientes preguntas:

a) ¿Cual es el error de clasificación?

b) ¿Cual es la precisión del clasificador?

c) ¿Cual es el *recall* del clasificador?

d) ¿Cual es el F_1–score del clasificador?