Examen de medio semestre de reconocimento de patrones

Julio Waissman Vilanova

noviembre 2019

1. La empresa *Necrosoft, Inc.* vende software y películas piratas por internet. Para esto tienen un sitio atractivo, una página en *facebook* y un sistema de pagos por tarjeta de crédito y *Paypal*. Dado que es una empresa de punta utiliza métodos de reconocimiento de patrones para analizar y administrar la información empresarial y como ayuda a la toma de decisiones (una cosa llamada por los gringos como *business intelligence*. Para las siguientes tareas selecciona si la tarea es de regresión (R), clasificación (C), aprendizaje no supervisado (N) u otro tipo de aprendizaje (O).

R C N O Predecir el trafico en la página web para el día de mañana, conociendo el trafico en los días pasados.

R C N O Realizar un estudio de segmentación de mercados para analizar la conveniencia que los clientes puedan pagar sus películas piratas en el Oxxo.

R C N O Determinar si un usuario es cliente potencial de películas o de software para enviarle publicidad orientada a su correo electrónico.

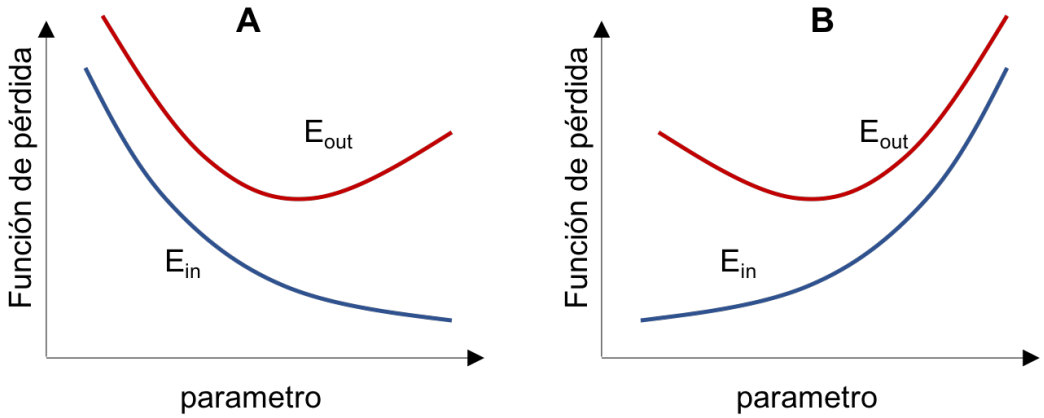
R C N O Recomendar a un usurario películas en base a la calificación que dio de otras películas y a las calificaciones de otros usuarios.

R C N O Estimar el ancho de banda necesario para el próximo año de la página del sistema.

R C N O Encontrar, a partir de la página de *facebook* grupos de personas con características comunes para seleccionar el software a piratear.

R C N O Predecir las ventas a finales de mes sobre hipotéticas ofertas, con el fin de realizar planeación estratégica en la empresa.

2. Considera las siguientes figuras A y B



Asigna cual es la curva sobre los errores en muestra E_{in} y fuera de muestra E_{out} que debería salir teóricamente para los siguientes parámetros de ajuste de métodos de aprendizaje:

A B Número de neuronas en la capa oculta en una red neuronal.

A B Parámetro λ de regularización en regresión lineal.

A B Umbral γ entre 0 y 1 que es el valor por el cual se considera que un objeto pertenece a la clase 1 en regresión logística (por default $\gamma = 0.5$.

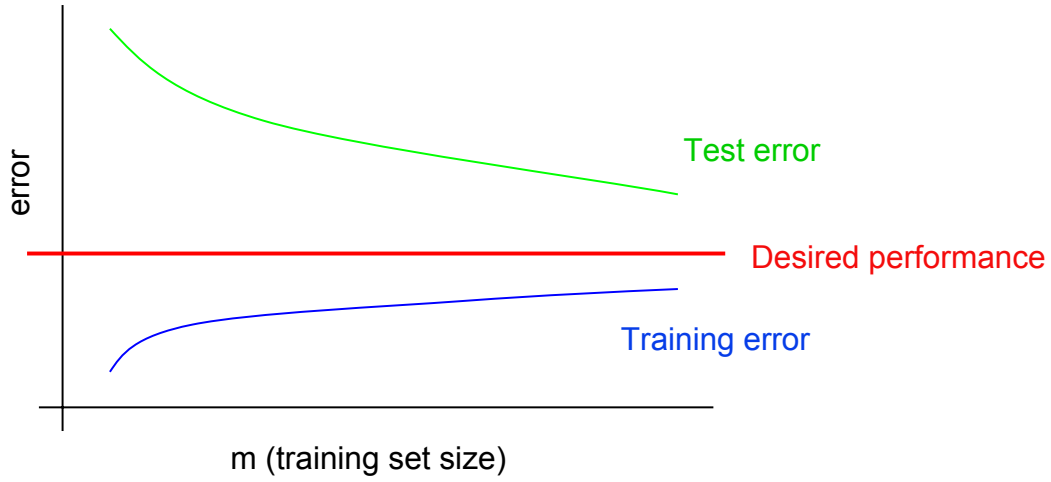
A B Valor de C es una SVM con kernel lineal.

A B Número de capas ocultas en una red neuronal.

A B Parámetro σ utilizado en el kernel gaussiano para una SVM.

3. Supongamos que se esta resolviendo un problema de análisis de sentimientos en documentos utilizando como método para clasificar los diferentes sentimientos que se pueden obtener de un *twit* utilizando un algoritmo basado en regresión logística. Si tenemos al rededor de 5000 palabras diferentes en nuestra *bolsa de palabras*, 4 sentimientos básicos («satisfecho», «molesto», «triste», «otro») y al rededor de 20,000 *twits* previamente clasificados.

   Si realizamos una curva de aprendizaje y obtenemos algo similar a la curva siguiente:

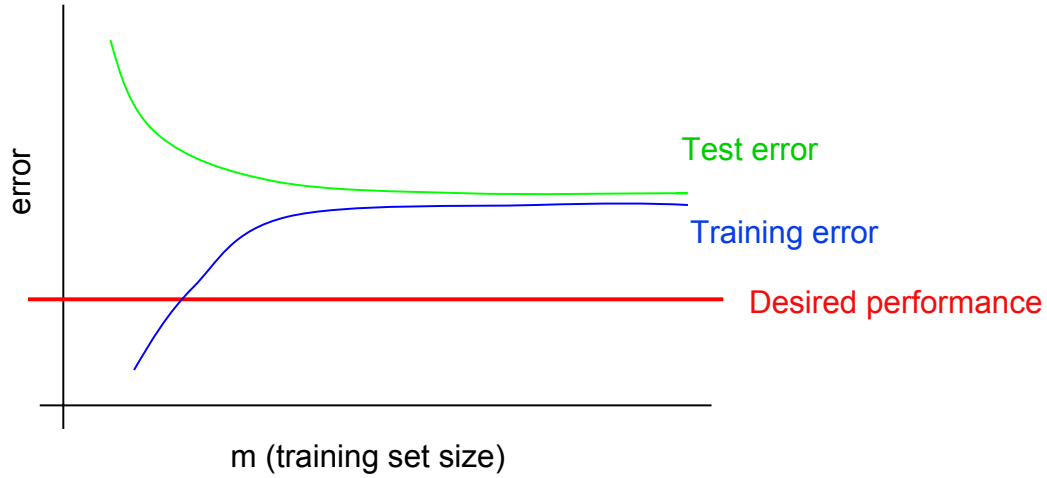


   subraya las acciones que podrían mejorar al sistema de aprendizaje.

   - Tratar de generar más *tuits* clasificados.
   - Tratar de reducir el número de palabras de la *bolsa de palabras*.
   - Tratar de aumentar las palabras de la *bolsa de palabras*.
   - Agregar otras características como la longitud del *tuit* o la hora a la que fue enviado.
   - Aumentar el número máximo de iteraciones del algoritmo de optimización.
   - Aumentar el valor de λ (parámetro de regularización).
   - Disminuir el valor de λ (parámetro de regularización).
   - Utilizar una SVM con kernel gaussiano.

4. 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 sistema 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.

5. 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? _____

6. *Chancrosoft Inc.* tiene un problema de *sentimental analysis*, esto es, la compañía tiene una cantidad importante de comentarios en un blog sobre los usuarios de su servicio de cita con sifilíticos, y tienen algunas de estas entradas ya previamente clasificadas como *usuario satisfecho*, *usuario indiferente* y *usuario molesto*. La compañía quiere hacer dos cosas: primero, a partir de esta muestra poder estimar el resto de las entradas de usuarios molestos (que pueden ser cientos de miles); por último, tratar de encontrar que es lo que hace que un usuario esté insatisfecho. Entre los datos que se clasificaron a mano previamente el 40 % está satisfecho, un 30 % está molesto y el resto es indiferente.

a) ¿Como procesas los datos para que puedan ser usados como entrada a un sistema de clasificación? ¿Como quedaría $x^{(i)}$ donde $x^{(i)}$ es un usuario? Responde aquí mismo.

b) ¿Que crees que es mejor en este caso, un clasificador binario o uno multiclase? ¿Un árbol de decisión, una SVM, una RN, una regresión logística o una regresión lineal? Justifica tu respuesta brevemente.

c) Completa las siguientes oraciones con los números de las acciones que debes de realiz ar para soluciones el problema completo que se pide. Las actividades se realizan en orden cronológico. No todas las acciones hay que realizarlas, y algunas podrías tener que hacerlas más de una vez. Las oraciones es la siguiente:

Vamos a hacer _____ y si resulta que el clasificador es de alta varianza entonces _____ pero si el clasificador es de alto bias entonces _____.

Una vez establecido en clasificador entonces hacemos _____ o también podríamos hacer _____.

Las actividades que se pueden agregar son:

1) Calcular el costo en el conjunto de aprendizaje y de validación y modificar las variables del algoritmo de aprendizaje en función de este criterio.
2) Calcular el F_1–score en el conjunto de aprendizaje y de validación y modificar las variables del algoritmo de aprendizaje en función de este criterio.
3) Calcular la curva de aprendizaje.
4) Intercambiar las posiciones de los datos en forma aleatoria.
5) Seleccionar 20 % de los datos para conjunto de prueba, 20 % de los datos para validación y 60 % de los datos para entrenamiento.
6) Seleccionar 20 % de los datos para conjunto de prueba y 80 % de los datos para entrenamiento.
7) Agregar más atributos o calcular nuevos en relación a los existentes.
8) Reducir el número de atributos.
9) Solicitar a la compañía que clasifique más entradas del blog.
10) Aplicar el método de regresión logística.
11) Aplicar el método de regresión lineal.
12) Aplicar el método de SVM.
13) Aplicar el método de Redes neuronales multicapa.
14) Aplicar el método de árboles de decisión
15) Aplicar el método de K–medias
16) Aplicar análisis en componentes principales.
17) Aplicar análisis en componentes independientes.
18) Graficar en un plano las dos primeras componentes principales y revisar como se componen.
19) Revisar los centros de cada uno de los K clusters obtenidos con las K–medias.
20) Aplicar el método de filtro colaborativo.
21) Clasificar y extraer todos los datos que pertenezcan a la clase *usuario molesto*.
22) Clasificar y extraer todos los datos que pertenezcan a la clase *usuario satisfecho*.
23) Otra actividad: _____.
24) Otra actividad: _____.
25) Otra actividad: _____.
26) Otra actividad: _____.

7. Desarrolla tu propio algoritmo de aprendizaje, utilizando Modelos Lineales Generalizados, pero asumiendo que los datos se presentan con una distribución de Poisson. En esta pregunta es más importante los procedimientos que los resultados, lo importante es ver que conocen y entienden las ideas generales para desarrollar Modelos Lineales Generalizados.

   a) Considera la distribución de Poisson, parametrizada como:

   $$\Pr(y;\lambda) = \frac{e^{-\lambda}\lambda^{y}}{y!}.$$

   Muestra que la distribución de Poisson pertenece a la familia exponencial, e indica claramente cuales son los valores de

   - $b(y)$:
   - η:
   - $T(y)$:
   - $a(\eta)$:

   b) ¿Cual es la respuesta canónica para el Modelo Lineal Generalizado con una distribución de poisson? Recurda que para la regresión lineal era $h(x) = \omega^T x + b$, y para la regresión logística era $h(x) = \frac{1}{1+\exp(-(\omega^T x+b))}$. Para la distribución de Poisson es:

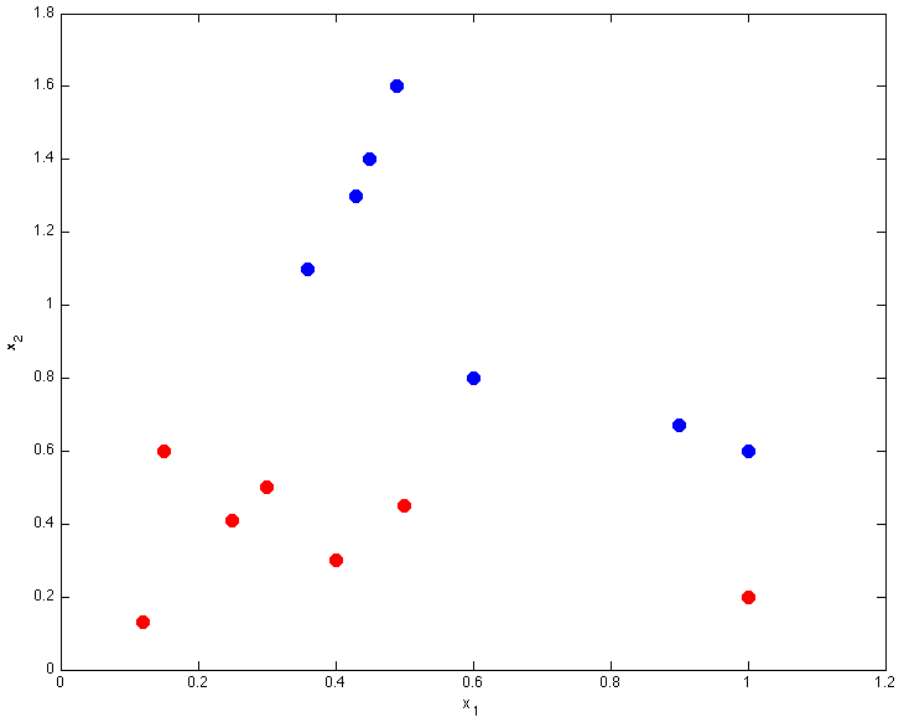
   $$h(x) =$$

   c) Si asumimos un conjunto de aprendizaje $\{(x^{(i)}, y^{(i)}; i = 1, \ldots, T\}$, deriva el algoritmo de aprendizaje por descenso de gradiente, utilizando las verosimilitudes lograrítmicas (log–likelyhood) de la misma forma que en clase las utilizamos para derivar el aprendizaje para la regresión lineal, logística y softmax. Escribe la formula final, y sobre todo todo el procedimiento que realizaste para llegar a ella.

8. Consideremos un problema de máquinas de vector de soporte:

   a) En la figura siguiente se muestran un conjunto de datos en dos dimensiones pertenecientes a dos clases.

      1) Dibuja en color negro la frontera de separación entre clases si se utiliza una SVM con kernel lineal y con $C = 0$. Dibuja tambien el margen de separación M.
      2) Dibuja en color azul la frontera de separación entre clases si se utiliza una SVM con kernel lineal y con $C = 10000$. Dibuja tambien el margen de separación M.
      3) Dibuja en color rojo la frontera de separación entre clases si se utiliza una SVM con kernel gausiano.



   b) Sea $x^{(i)} = (x_1^{(i)}, x_2^{(i)})$ y $\Phi(x^{(i)})$ la expansión de $x^{(i)}$ en polinomio de orden 2.

      1) Encuentra la operación para realizar el producto punto $x^{(i)T}x^{(j)}$.

2) Encuentra la operación para realizar el producto punto $\Phi(x^{(i)})^T\Phi(x^{(j)})$.

3) Expresa la operación $\Phi(x^{(i)})^T\Phi(x^{(j)})$ en función únicamente de $x^{(i)}$ y $x^{(j)}$ como la función $k(x^{(i)}, x^{(i)})$.