Lista Ejercicios Tópico IA - Ciencia de la Computación -UCSP

1. Considere el problema de predecir cuán bien le irá a un estudiante en su segundo año de universidad, dado los datos de su primer año. Sea x el número de notas A que un estudiante recibe en su primer año. Nos gustaría predecir el valor y , el número de As en su segundo año. Usando el conjunto de datos performance de los estudiantes a seguir. Cada fila es un conjunto de entrenamiento. Usando regresión lineal por lo tanto $h_\theta(x) = \theta_0 + \theta_1 x$.

x	y
5	4
3	4
0	1
4	3

a)Use la función de costo ($J(\theta_0,\theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$) para determinar J(0,1).

m=4 θ_0 =0, θ_1= 1

$J(\theta_0,\theta_1) = \frac{1}{2(4)} [(5-4)^2 + (3-4)^2 + (0-1)^2 + (4-3)^2]$

$J(\theta_0,\theta_1)$ =4/8 = 0.5

b) Suponga que $\theta_0 = -1, \theta_1 = 2$, Cual es el valor de $h_\theta(6)$?

θ_0 =-1, θ_1= 2

$h_\theta(6)$= -1+2(6)=11

2. Sea f una función tal que $f(\theta_0,\theta_1)$ retorna un número. Para este problema, f es una función arbitraria (no necesariamente la función de costo de regresión lineal,osea puede haber óptimo local). Suponga que se usa gradiente descendiente para intentar minimizar (θ_0,θ_1). Cuales de las siguientes afirmaciones son verdaderas:

a) Si θ_0 y θ_1 son inicializadas a $\theta_0 = \theta_1$, entonces por simetría (se actualizan los parámetros simultaneamente), después de una iteración aún se tendría $\theta_0 = \theta_1$

Falso. porque al efectuar gradiente descendiente dependera de la funcion f y no asegura que su derivada parcial sea la misma.

b) Si el radio de aprendizaje es muy pequeño, gradiente descendiente puede demorar en convergir

Verdadero. Entre mas pequeña sea el radio de aprendizaje, mas demorara.

c) si θ_0 y θ_1 son inicializados a un mínimo local, una iteración no cambiará sus valores

Verdadero. al estar en el mismo minimo local el valor de la gradiente es 0 y no afectara a los valores de θs.

d) no importa como θ_0 y θ_1 sean inicializados, dado un α suficientemente pequeño, se espera que gradiente descendiente pueda convergir a la misma solución.

Falso. depende de la funcion, ya que puede que tenga varios minimos locales diferentes.

3. Suponga que para algún problema de regresión lineal (predicción de casas por ejemplo), dado un conjunto de entrenamiento se encuentren valores para θ_0 ,θ_1 tales que J (θ_0,θ_1) = 0. Cuales de las afirmaciones son verdaderas?

a) Para estos valores θ_0 ,θ_1 que satisfacen J(θ_0,θ_1) = 0, se tiene que h_θ ($x^{(i)}$) =$y^{(i)}$ para cada ejemplo de entrenamiento ($x^{(i)}$,$y^{(i)}$)
Verdadero. J(θ0,θ1) = 0 significa la linea definida por la ecuacion y=θ_0 + θ_1x encaja perfectamente en nuestros datos.

b) Para que esto se verdadero se debe tener que $y^{(i)}$ = 0 para cada valor i= 1 ,2,...,m
Falso. Siempre que todos nuestros ejemplos de entrenamiento estén en línea recta, podremos encontrar θ0 y θ1 de modo que J (θ0, θ1) = 0. No es necesario que y (i) = 0 para todos nuestros ejemplos.

c) Esto no es posible: por la definición de J (θ_0 ,θ_1), no es posible que existan θ_0 ,θ_1 tales que J (θ_0,θ_1) = 0
Falso. Si todos nuestros ejemplos de entrenamiento se encuentran perfectamente en una línea, entonces J (θ0, θ1) = 0 es posible

d)Gradiente descendiente es probable que caiga en un mínimo local y que no encuentre el mínimo global.
Falso….., Esto no es cierto, dependiendo de la condición inicial, el descenso del gradiente puede terminar en diferentes óptimos locales.

4. Cuales de las siguientes son razones para usar escala de características?

a) previene la matriz $X^T X$ (usada en la ecuación normal) de ser no invertible (singular)
Falso.X^TX puede ser singular cuando las funciones son redundantes o hay muy pocos ejemplos. La escala de características no resuelve estos problemas

b) es necesario para prevenir que la ecuación normal caiga en un óptimo local
Falso.
Si optas por la ecuación normal, no necesitas de la escala de valores, si optas por la gradiente si.
La función de costo J (θ) para la regresión lineal no tiene óptimos locales.

c) Acelera el algoritmo gradiente descendiente haciendo que requiera menos iteraciones para llegar a una buena solución
Verdadero. El escalado de características acelera el gradiente descendiente al evitar muchas iteraciones adicionales que se requieren cuando una o más características adquieren valores mucho más grandes que el resto.

d) Acelera el proceso de encontrar θ al usar la ecuación normal.
Falso. La magnitud de los valores de las características es insignificante en términos de costo computacional.

https://github.com/ngavrish/coursera-machine-learning-1/blob/master/quiz/week2-quiz4.md
https://github.com/DragonflyStats/Coursera-ML/blob/master/QuizNotebook/Week02Quiz.tex

5. Cuales de las siguientes afirmaciones son verdaderas

a) Si los datos son linealmente separables, un SVM usando kernel lineal retornará los mismos parámetros sin importar el valor de C (osea el valor resultante de θ no depende de C)
falso.

If the data are linearly separable, an SVM using a linear kernel will return the same parameters θ regardless of the chosen value of C (i.e., the resulting value of θ does not depend on C).	✔	0.25	A linearly separable dataset can usually be separated by many different lines. Varying the parameter C will cause the SVM's decision boundary to vary among these possibilities. For example, for a very large value of C, it might learn larger values of θ in order to increase the margin on certain examples.

b) si se está entrenando un SVM multiclase con el método one-vs all, no es posible usar un kernel
Falso., cada SVM que se entrena en ONE vs All es un SVM estandar, asi que uno es libre de usar un kernel

c) el valor máximo de un kernel Gaussiano (sim $(x,l^{(1)})$) es 1.
Verdadero.

The maximum value of the Gaussian kernel (i.e., $sim(x, l^{(1)})$ is 1.	✔	0.25	When $x = l^{(1)}$, the Gaussian kernel has value $\exp(0) = 1$, and it is less than 1 otherwise.

d) Suponga que se tiene ejemplos de entrada en 2D ($x^{(i)} \in R^2$). La frontera de decisión en SVM (con el kernel lineal) es una línea recta
Verdadero. SVM sin kernel predice la salida basada solo en (theta.T.x) asi que da una linea recta de desicion, exactamente como la regresión logística la hace.

6. Considere los siguientes modelos de regresión logística para un problema de clasificación binaria con función sigmoidea g (z) = $\frac{1}{1+e^{-z}}$

Modelo 1: $P(Y= 1|X,\theta_1,\theta_2) = g(\theta_1X_1+\theta_2X_2)$
Modelo 2: $P(Y= 1|X,\theta_1,\theta_2) = g(\theta_0+\theta_1X_1+\theta_2X_2$
Se tiene los tres ejemplos de entrenamiento:

$X^{(1)} = [11]^T, X^{(2)} = [10]^T, X^{(3)} = [00]^T$
$Y^{(1)} = 1, Y^{(2)} = -1, Y^{(3)} = 1$

a)¿Tiene importancia que el tercer ejemplo sea etiquetado en el modelo 1?
No por el modelo 1, siempre resultara g(0)=0.5 sin importar cuanto cambie los Θs no variara la respuesta a la funcion sigmoidea.

b) ¿el valor aprendido $\Theta = (\theta_1, \theta_2)$ sería diferente si cambiará la etiqueta a -1?,
Si, debido que al ser diferente a lo obtenido por g(z) hara variar los Θs.

c) y en el modelo 2?, justifique su respuesta
Si tiene importancia, al existir Θ_0 independiente de los valores de x, si afectara al encontrar los Θs, también el valor aprendido sera diferente.

7. Considere un dataset con 4 ejemplos, donde cada ejemplo es un punto en el espacio de características 2D, y la variable clase es dada por:
$X_1 = (X_{11} = -1, X_{12} = -), Y_1 = -1$
$X_2 = (X_{21} = -1, X_{12} = +1), Y_1 = +1$
$X_3 = (X_{31} = +1, X_{12} = -1), Y_1 = +1$
$X_4 = (X_{41} = +1, X_{12} = +1), Y_1 = -1$

a) Puede un SVM lineal clasificar perfectamente este dataset?. Si es verdad, muestre los puntos y frontera de decisión calculada por SVM. Si no, explique porque no.
Falso. Esto es XOR, que no es separable linealmente..

b) Suponga que se re-expresa los datos de entrada de $X_i = (X_{i1}, X_{i2})$ para $\varphi(X_i) = (X_{i1}, X_{i1} * X_{i2})$. Puede un SVM lineal clasificar perfectamente estos datos 2D. Explique como o porque no.
Verdad. La línea de decisión es el eje(axis) horizontal x1….

8. Suponga que se entrena un clasificador con regresión logística $h\theta = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$. Suponga que $\theta_0 = 6$, $\theta_1 = 0$, $\theta_2 = -1$. Cual de las figuras siguientes representa la frontera de decisión encontrada por el clasificador y por que?

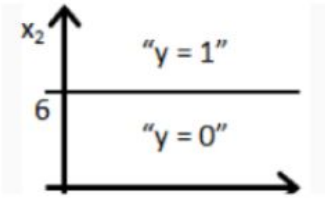


Figura 1: a)

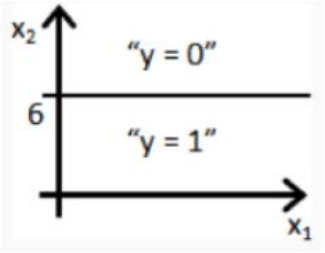


Figura 2: b)

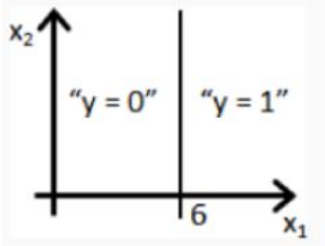


Figura 3: c)

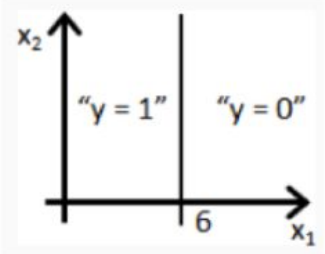


Figura 4: d)

RPTA: figura 2

En esta figura pasamos de negativo a positivo cuando x2 va de arriba 6 a abajo 6, lo que es cierto para los valores dados de 0.

9. En una frase caracterice las diferencias entre clasificación y regresión.

Clasificación aproximar una función (f) de la entrada (X) a la salida discreta (y).

Regresión aproximar una función (f) de la entrada (X) a la salida continua(y).

Con la información sobre el tamaño de las casas en el mercado inmobiliario, para predecir su precio en función del tamaño es un resultado continuo, por lo que es un problema de regresión.

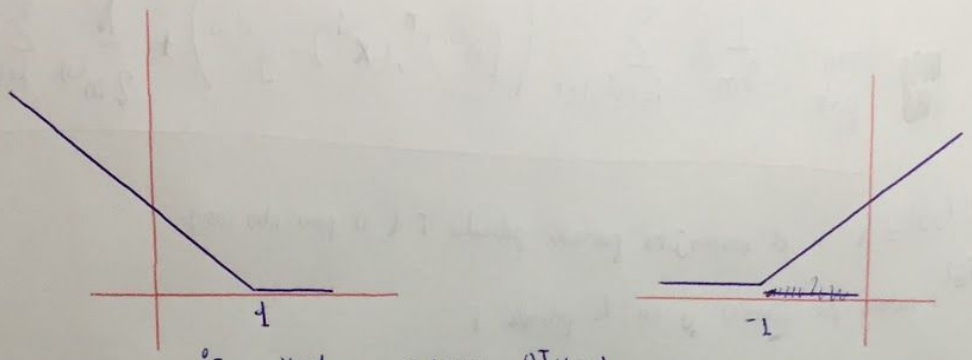
Podríamos convertir este ejemplo en un problema de clasificación al hacer nuestra salida sobre si la casa "se vende por más o menos que el precio solicitado". Aquí estamos clasificando las casas según el precio en dos categorías discretas.

10. Explique la intuición matemática de Large Margin de SVM, use fórmulas.

(10) Explique la intuicion matematica de Large Margin de SVM, use formulas

SVM hipotesis

$$h_\theta x \begin{cases} 1 & \text{if } \theta^T . x \geq 0 \\ 0 & \text{otro caso} \end{cases}$$



if $y=1$ queremos $\theta^T . x \geq 1$

if $y=0$ queremos $\theta^T . x \leq -1$

(I) Large margins

p = largo de proyección de v sobre u

producto Interno $u^T . v = p . \|u\| = u_1 v_1 + u_2 v_2$

$\|u\|$ = largo del vector $= \sqrt{u_1^2 + u_2^2} \in \mathbb{R}$

(II)

$$\min_\theta \frac{1}{2} \sum_{j=1}^{n} \theta_j^2 = \frac{1}{2}(\theta_1^2 + \theta_2^2) = \frac{1}{2}\left(\sqrt{\theta_1^2 + \theta_2^2}\right)^2 = \boxed{\frac{1}{2}\|\theta\|^2}$$

$\theta_0 = 0$ y $n = 2$

s.t. $\theta^T . x^{(i)} \geq 1$ if $y^{(i)} = 1$

$\theta^T . x^{(i)} \leq -1$ if $y^{(i)} = 0$

(III) $\omega = (\sqrt{\omega})^2$

$\theta^T . x^{(i)} = ?$ ($u^T . v$)

$\theta^T . x^{(i)} = \boxed{p^{(i)} . \|\theta\|}$

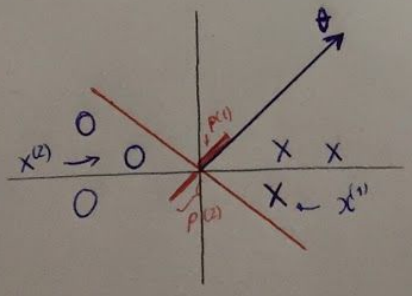
$\theta^T . x^{(i)} = \theta_1 x_1^{(i)} + \theta_2 . x_2^{(i)}$

(IV)

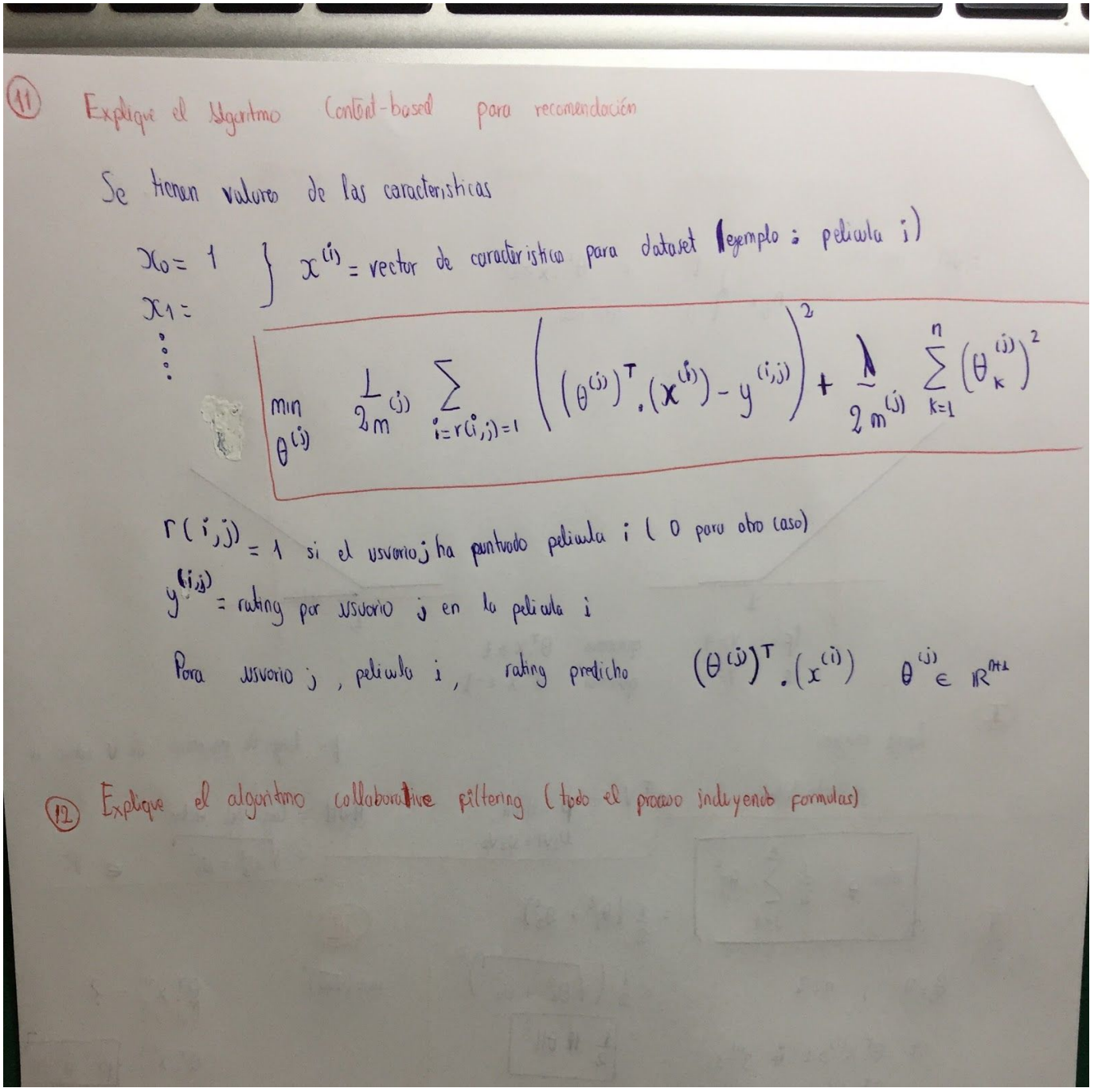
$p^{(i)} \|\theta\| \geq 1$ if $y^{(i)} = 1$

$p^{(i)} \|\theta\| \leq -1$ if $y^{(i)} = 0$

$p^{(i)}$ es la proyección de $x^{(i)}$ en el vector θ



11. Explique el algoritmo Content-based para recomendación.



(11) Explique el algoritmo Content-based para recomendación

Se tienen valores de las características

$x_0 = 1$
$x_1 =$
$\vdots$
$\Big\}\ x^{(j)}$ = vector de características para dataset (ejemplo : película j)

$$\min_{\theta^{(j)}} \frac{1}{2m^{(j)}} \sum_{i:r(i,j)=1} \left((\theta^{(j)})^T \cdot (x^{(i)}) - y^{(i,j)} \right)^2 + \frac{\lambda}{2m^{(j)}} \sum_{k=1}^{n} \left(\theta_k^{(j)} \right)^2$$

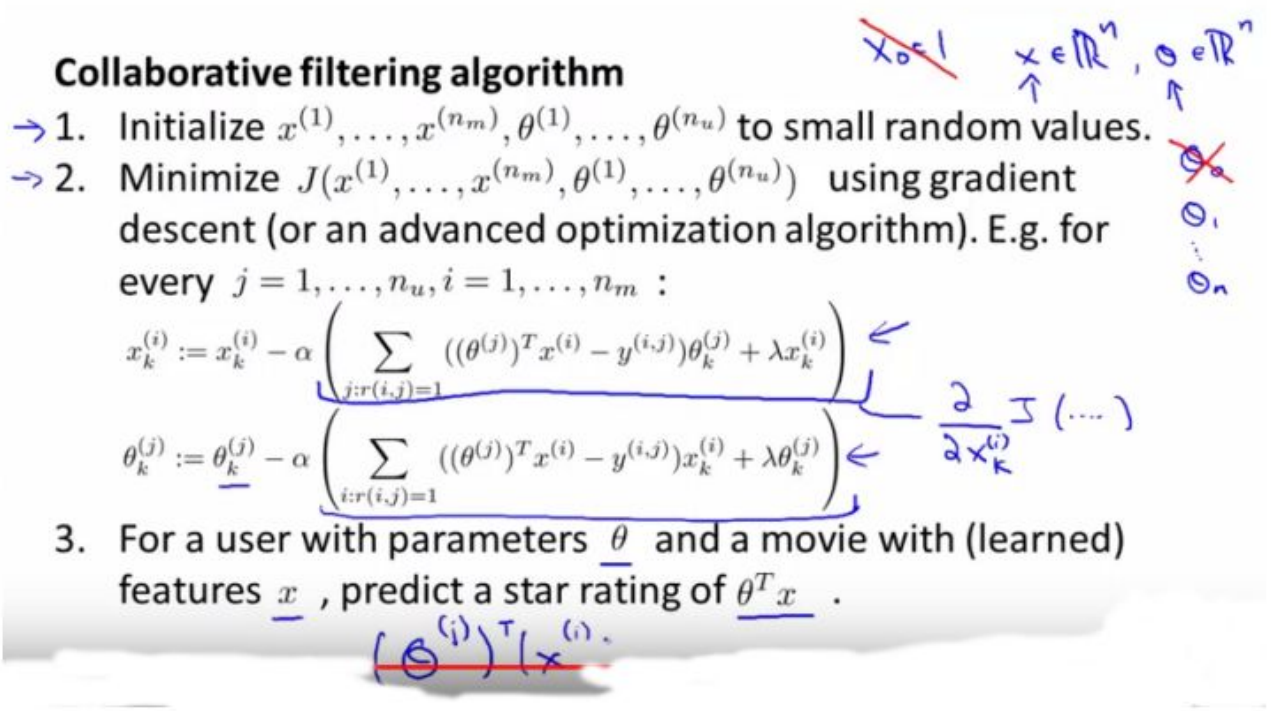
$r(i,j) = 1$ si el usuario j ha puntuado película i (0 para otro caso)

$y^{(i,j)}$ = rating por usuario j en la película i

Para usuario j , película i , rating predicho $(\theta^{(j)})^T \cdot (x^{(i)})$ $\theta^{(j)} \in \mathbb{R}^{n+1}$

(12) Explique el algoritmo collaborative filtering (todo el proceso incluyendo formulas)

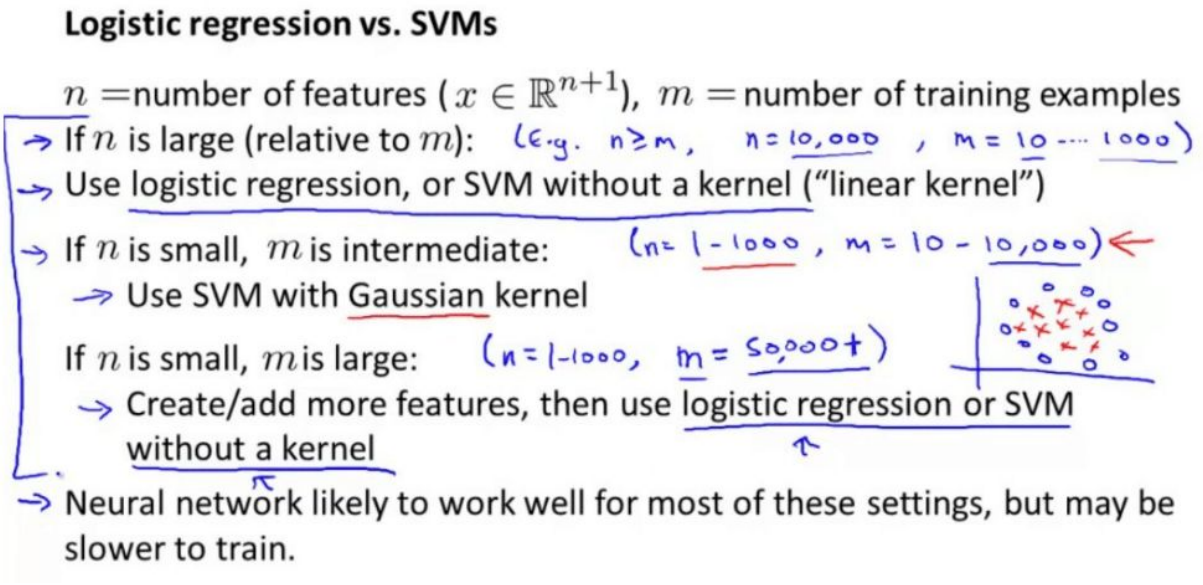
12. Explique el algoritmo collaborative filtering (todo el proceso incluyendo fórmulas).



1. Primero inicializamos X y theta a peque;os valores aleatorios, similar a una red neuronal
2. Después vamos a minimizar la función de costos usando gradiente descendiente o uno de los algoritmos de optimización avanzados. Por lo tanto, si toman las derivadas encontrarán las actualizaciones de gradiente descendiente.
3. Finalmente, si un usuario tiene algunos parámetros, theta y si hay una pelicula con algun tipo de variables aprendidas X, entonces podríamos predecir que a esa película de darian una calificación buena por parte de ese usuario.

Entonces diríamos que si el usuario j aún no ha calificado la película i, entonces lo q hacemos es predecir que el usuario j va a calificar la película i de acuerdo con (θ ^(j))^T (X^i)

13. En que situaciones debería usar regresión logística o SVM, justifique.



14. Supongamos que se quiere construir un modelo para predecir la deserción de estudiantes en la UCSP. Diseñe un clasificador o predictor (justifique su elección), defina las características, defina la hipótesis, asuma que los parámetros se aprendieron y demuestre su uso con un ejemplo en particular.

Clasificador: por que es discreto, deserta=1, no deserta=0
Características para un individuo x.
x1: promedio de las notas de los cursos del último semestre
x2: número de cursos en las que superó el límite de faltas dividido con el número de cursos que lleva.
x3: porcentaje de deserción de la carrera
x4: último semestre que cursó.

$h\theta = g(z) = \frac{1}{1+e^{-z}}$; $z= \theta_0+\theta_1X_1+\theta_2X_2+\theta_3X^2_3+\theta_4X^2_4$

ejemplo: nro de cursos= 5

	c1	c2	c3	c4	c5
notas	0	0	12	13	10

hallando x: $\theta = [0\ 1\ -1\ 1\ ]$
x1= 6.5
x2=0.4
x3=60
x4=2

$Z= 6.5 +0.4 -60^2+ 2^2 <0$
para $z < 0$,
$g(z) > 0.5$
Por lo tanto. g(z) = 1 -> el alumno deserta.

15. Explique el procedimiento de k-fold cross validation, para regresión y clasificación.

- El método K-Fold Cross-Validation es también un proceso iterativo. Consiste en dividir los datos de forma aleatoria en k grupos de aproximadamente el mismo tamaño. k-1 grupos se emplean para entrenar el modelo y uno de los grupos se emplea como test, este proceso se repite k veces utilizando un grupo distinto como test en cada iteración. El proceso genera k estimaciones del test error cu loyo promedio se emplea como estimación final. El resultado es una matriz que contiene los k puntajes de evaluación.

Creo que el procedimiento es el mismo para el método con el que se quiera predecir, cuando haces la validación cruzada K-fold, estás probando qué tan bien tu modelo puede ser entrenado por algunos datos y luego predecir datos que no ha visto. El propósito de la validación cruzada es la verificación del modelo, no la construcción del modelo.

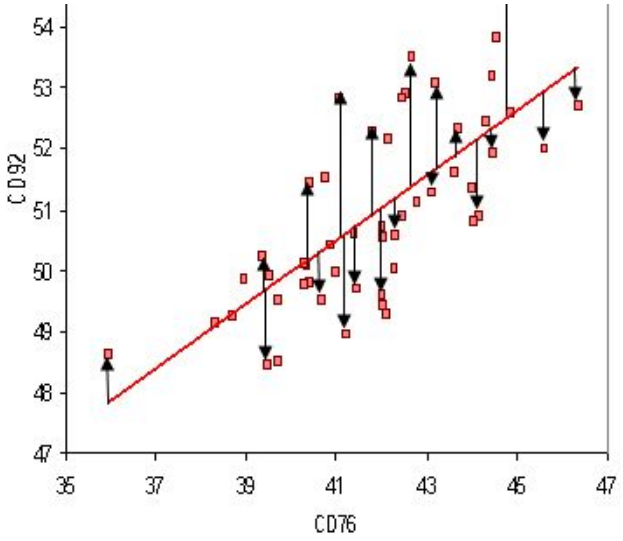
16. Por qué es importante definir un conjunto de test, explique cómo se define este conjunto. (Angel)

El conjunto de test nos sirve para probar nuestro modelo una vez haya terminado su etapa de entrenamiento usando el conjunto de entrenamiento. Usando métricas de evaluación como el RMSE podemos puntuar nuestro modelo gracias a los resultados obtenidos de la predicción del conjunto de prueba.

De un conjunto total de datos separamos el 80% de los datos para el conjunto de entrenamiento y 20% para el conjunto de prueba. El conjunto de test se puede definir utilizando Muestreo Estratificado, la población (conjunto total) se divide en subgrupos homogéneos llamados estratos, y se toma el número correcto de instancias de cada estrato para garantizar que el conjunto de prueba sea representativo de la población general.

17. Explique la métrica RMSE. (Angel)

Root Mean Square Error: En regresión lineal la línea de regresión predice el valor promedio de “y” asociado con un valor de x dado. RMSE nos ayuda a obtener una medida de la propagación de los valores “y” en torno a ese promedio.



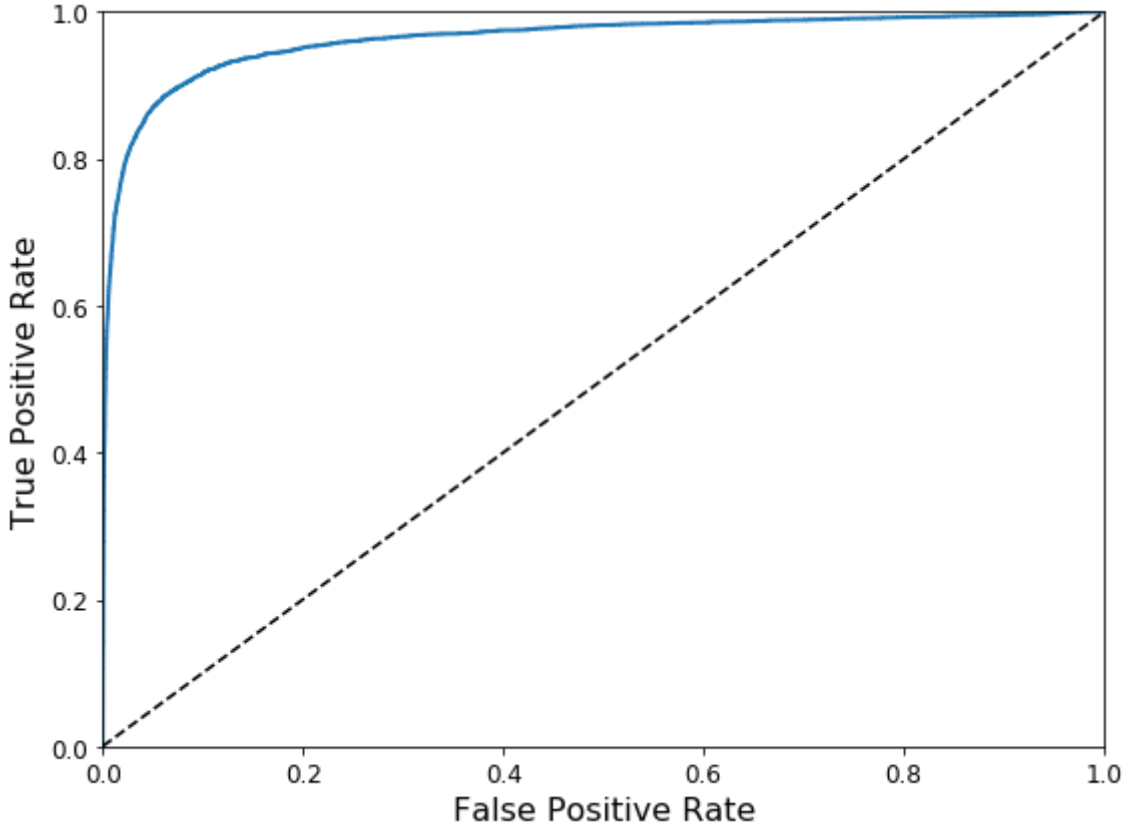
Para construirlo necesitas obtener los “residuos” que son la diferencia entre los valores reales y los valores predichos.Pueden ser positivos o negativos ya que el valor predicho debajo o sobre el valor real. Elevando al cuadrado los residuos, promediando los cuadrados, y tomando la raíz cuadrada nos da el error r.m.s.

$$RMSErrors = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}$$

Donde $\hat{y}_i$ es el valor predicho

18. Explique las curvas ROC y la métrica AUC ._dani
The receiver operating characteristic (ROC) es otra herramienta común utilizada con los clasificadores binarios.
Es muy similar a la curva de precisión / recuperación, pero en lugar de trazar la precisión vs a la recuperación.

la curva ROC traza **true positive rate (another name for recall)** vs **the false positive rate**.
En resumen dibuja recall (true positive rate) vs **the false positive rate**.
El FPR es igual a uno menos **the true negative rate**, que es la proporción de instancias negativas que se clasifican correctamente como negativas.
The TNR is also called specificity.
Por lo tanto, la curva ROC traza la sensibilidad (**recall**) versus 1 - **specificity**.



AUC. ES el área debajo de la curva.

Una forma de comparar clasificadores es medir el área bajo la curva (AUC). Un clasificador perfecto tendrá un AUC ROC igual a 1, mientras que un clasificador puramente aleatorio tendrá un AUC ROC igual a 0.5.
/*Es decir, mientras nuestra curva ROC, más se parezca a un rectángulo o cuadrado, tendrá un área más cercana a 1*/

Creo que esto podría ayudar: http://users.dsic.upv.es/~jorallo/Albacete/CharlaROC-1.5.pdf

19. Explique las métricas precision, recall y F1 measure.

Precision - Recall

- Estas métricas responden a las siguientes preguntas.
  - De todas las muestras clasificadas como positivas, ¿que porcentaje es correctamente clasificado?

$$Precision = \frac{TP}{TP + FP} \quad (8)$$

    - También llamada PPV (positive predictive value)
  - De todas las muestras positivas, ¿que porcentaje es correctamente clasificado?

$$Recall = \frac{TP}{TP + FN} \quad (9)$$

    - Muchos otros nombres: sensitivy, hit rate y TPR (true positive rate)

F-Measure

- Es la media armónica de *precision* y *recall*:

$$F_1 = 2 * \frac{precision * recall}{precision + recall} \quad (10)$$

- Se busca que esta métrica de un valor alto solo cuando la precision y el recall sean altos.

Metricas de ranking

- Estas métricas son muy utlizadas en sitemas de ranking. Por ejemplo, el buscador de google.
  - Un motor de búsqueda puede ser visto como un clasificador binario: **dado un término de búsqueda el documento Dx es relevante para la búsqueda**
- Las métricas anteriores pueden ser interpretadas como probabilidades en vez de como proporciones:
  - **Precision**: probabilidad de que un documento devuelto en la búsqueda sea relevante.
  - **Recall**: probabilidad de un documento relevante sea devuelto en la búsqueda.

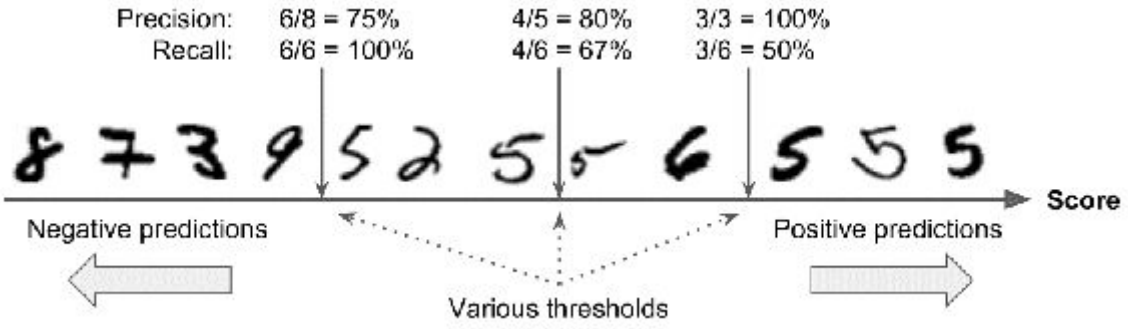
20. Si en un problema de clasificación, tuviera un conjunto de entrenamiento con clases desbalanceadas, que métrica sería la más adecuada, justifique.

For binary classifiers ROC is also a good metric. ROC is not affected by class imbalance, and provides the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative example.

https://www.quora.com/What-error-metric-would-you-use-to-evaluate-how-good-a-binary-classifier-is-What-if-the-classes-are-imbalanced-What-if-there-are-more-than-2-groups

21. En relación a Precision/Recall tradeoff en clasificación, si se tiene un threshold, por ejemplo de 0.5, como consigo más precision y más recall, explique.

Otra forma de verlo _dani



Si mueves el trshold mas a la derecha, vas a tener mas presicion pero menos recall.
Lo mismo pasa a la invesa.