Instituto Tecnológico Autónomo de México

Aprendizaje de Maquina

Respuestas Examen Parcial

Nombre:
Pablo Barranco Soto - 151528

Profesor:
Juan Salvador Marmol Yahya

1 de Marzo del 2020

1. Clasificación Lineal

1.1. ¿Cual es el intercepto resultante θ_0?

Como sabemos los errores por los que pasa el algoritmo, no hay necesidad de saber el orden en el que fueron procesados los datos.

$$\theta_0 = \sum_{i=1}^{10} y^{(i)} * \mathbf{error}_i = -18$$

Donde $y^{(i)}$ es a i-esima etiqueta y $\mathbf{error}_i$ es el i-esimo error correspondiente.

1.2. ¿Cuál es el vector de parámetros θ resultante?

Lo mismo ocurre para conocer el valor de θ

$$\theta = \sum_{i=1}^{10} y^{(i)} * x^{(i)} * \mathbf{error}_i = [4, 4]^T$$

Con $x^{(i)}$ el i-esimo punto

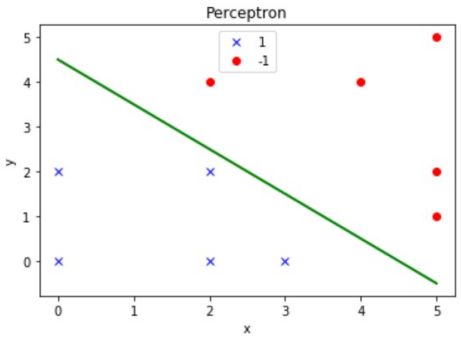

Figura 1: Perceptron Lineal

1.3. Los errores que comete el algoritmo a menudo dependen del orden en que se consideraron los puntos. ¿Podría el punto (5,2) etiquetado +1 haber sido el primero considerado para obtener los resultados de la tabla anterior? Si/No/Depende Justifique

El punto $(5, 2)$ no podría ser el primer punto en entrar en el algoritmo porque en la primer iteración está inicializada con el vector $\theta = 0$ y por eso necesariamente ocurre un error en el algoritmo.

1.4. ¿Cuáles son los parámetros θ_0 y θ correspondientes al separador de margen máximo?

Los valores son $\theta_0 = -5$ y $\theta = [1, 1]^T$ y los encontré un poco al tanteo (a ojo de buen cubero). Vi que si rotamos esa recta que está a la mitad de dos puntos con diferentes etiquetas, entonces el margen se reduciría.

1.5. ¿Cuál es el valor del margen alcanzado (distancia del margen al hiperplano)?

El valor del margen, si usamos la norma 2 para medir sería

$$\frac{1}{||\theta||} = \frac{1}{\sqrt{\theta_1^2 + \theta_2^2}} = \frac{1}{\sqrt{1+1}} = \frac{1}{\sqrt{2}}$$

1.6. Usando los parámetros θ_0 y θ correspondientes al separador de margen máximo, ¿cuál es la suma de los Hinge Losses evaluadas en cada ejemplo?

La suma de los Hinge Losses estaría dado por:

$$\sum_{i=1}^{10} \mathbf{Loss}_h(y^{(i)}(\theta \cdot x^{(i)} + \theta_0)) = \sum_{i=1}^{10} 0 = 0$$

El resultado es cero pues ningún punto está dentro de los márgenes.

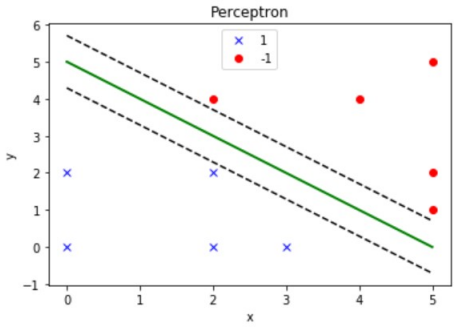

Figura 2: Separador Margen Maximo

1.7. Supongamos que modificamos un poco la solución de margen máximo y dividimos θ_0 y θ entre 2. ¿Cuál es la suma de las pérdidas de los Hinge Losses en cada ejemplo para este nuevo separador?

En el caso en que dividimos los parámetros entre dos, estaríamos haciendo el margen más grande y tres puntos estarían dentro del margen. Por lo quela suma de los Hinge Losses quedaría de la siguiente forma:

$$\sum_{i=1}^{10} \textbf{Loss}_h(y^{(i)}(\theta \cdot x^{(i)} + \theta_0)) = 1/2 + 1/2 + 1/2 = 3/2$$

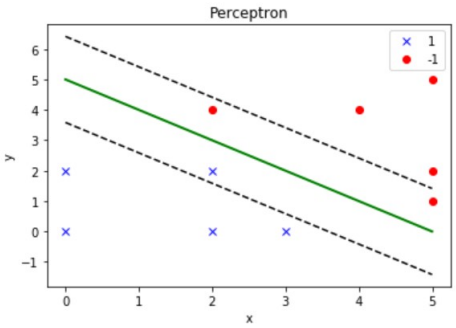

Figura 3: Separador Margen Maximo caso 2

2. Métodos de Kernel

2.1. Si nuevamente utilizamos el algoritmo de perceptrón lineal para entrenar al clasificador, ¿qué sucederá (seleccione una de las siguientes opciones y explique su respuesta)?

La respuesta es c). El algoritmo nunca convergerá porque no importa cómo sean ingresados los datos, el algoritmo continuará rotando la pendiente de la recta y cambiando la ordenada al origen en cada paso, ya que el conjunto de datos no es linealmente separable.

2.2. Según la tabla, ¿cuál es la salida de θ_0 y θ?

Dado que conocemos el numero de errores, el parametro esta dado por $\theta = \sum_{i=1}^{10} \alpha_i y^{(i)} \phi(x^{(i)})$,con α_i denotando el número de errores. Si hice bien los cálculos el resultado es el siguiente:

$$\theta = \sum_{i=1}^{10} \alpha_i y^{(i)} \phi(x^{(i)}) = \begin{pmatrix} 21 \\ -16\sqrt{2} \\ 22 \end{pmatrix}$$

Para $\theta_0 = \sum_{i=1}^{10} \alpha_i y^{(i)}$ por lo cual obtenemos:

$$\theta_0 = \sum_{i=1}^{10} \alpha_i y^{(i)} = -110$$

2.3. Basado en el cálculo de θ_0 y θ ¿La frontera de decisión $\theta^T \phi(x) + \theta_0 = 0$ clasifica correctamente todos los puntos en el conjunto de datos de entrenamiento?

Para ver si clasifica correctamente los puntos lo que hice fue calcular $sign(y^{(i)}(\theta^T \phi(x^{(i)} + \theta_0))$ de los 10 puntos del conjunto. Si el resultado es positivo quiere decir que la clasificación es correcta, de lo contrario estaría mal. Como todos los signos resultaron positivos puedo decir que la frontera de decisión $\theta^T \phi(x) + \theta_0 = 0$ clasifica correctamente.

También lo podemos ver gráficamente, debido a que

$$\theta^T \phi(x) + \theta_0 = 0$$

$$\Rightarrow \begin{pmatrix} 21 & -16\sqrt{2} & 22 \end{pmatrix} \begin{pmatrix} x_1^2 \\ \sqrt{2}x_1x_2 \\ x_2^2 \end{pmatrix} - 110 = 0$$

$$\Rightarrow 21x_1^2 - 32x_1x_2 + 22x_2^2 - 110 = 0 \qquad (2)$$

(2) describe la ecuación de una parábola con centro en (0,0). Lo podemos ver graficamente en la siguiente figura:

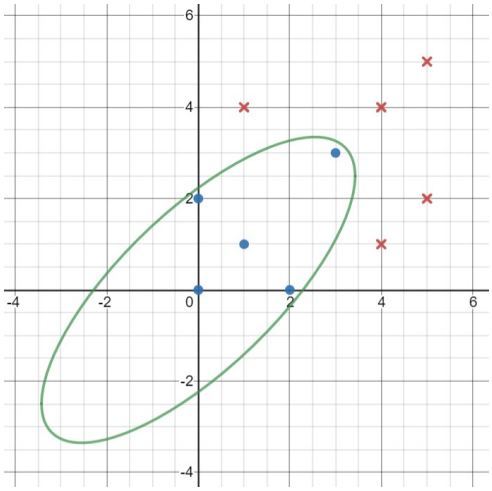

Figura 4: Kernel Perceptron

2.4. Para $x = [x_1, x_2]^T$ definimos la función del kernel como $K(x, x') = \phi(x)^T \phi(x')$. Escriba $K(x, x')$ como una función del producto punto de $x \cdot x'$. Para contestar, asuma que $z = x \cdot x'$ Defina el resultado en términos de z

$$K(x, x') = \phi(x)^T \phi(x') = \begin{pmatrix} x_1^2 & \sqrt{2}x_1x_2 & x_2^2 \end{pmatrix} \begin{pmatrix} {x'_1}^2 \\ \sqrt{2}x'_1x'_2 \\ {x'_2}^2 \end{pmatrix}$$

$$\Rightarrow \phi(x)^T \phi(x') = x_1^2 {x'_1}^2 + 2x_1x_2x'_1x'_2 + x_2^2 {x'_2}^2$$

$$\Rightarrow \phi(x)^T \phi(x') = (x_1 x_1' + x_2 x_2')^2$$

$$\Rightarrow \phi(x)^T \phi(x') = (x \cdot x')^2$$

$$\therefore K(x, x') = z^2$$

3. Descenso por gradiente

3.1. Encuentre $\nabla_\theta Loss_h(y\theta \cdot x)$ en términos de x para $y\theta \cdot x \leq 1$ y $y\theta \cdot x > 1$

La función $\mathbf{Loss}_h(y\theta \cdot x)$ no es diferenciable, pero es convexa por lo cual podemos encontrar su subgradiente $\nabla_\theta Loss_h(y\theta \cdot x)$.

Si llamamos $z(\theta) = y^{(i)}(\theta \cdot x^{(i)})$ entonces tenemos

$$\mathbf{Loss}_h(z) = \begin{cases} 0 & si \quad z \geq 1 \\ \\ 1 - z & si \quad z < 1 \end{cases}$$

por otra parte, calculamos $\frac{dz}{d\theta}$ obtenemos $z'(\theta) = y^{(i)} \cdot x^{(i)}$ y

$$\frac{Loss_h}{dz} = \begin{cases} 0 & si \quad z \geq 1 \\ \\ -1 & si \quad z < 1 \end{cases}$$

Por lo que, juntando ambas y re-expresando en términos de $y^{(i)}(\theta \cdot x^{(i)})$

$$\frac{Loss_h}{dz}\frac{dz}{d\theta} = \begin{cases} 0 & si \quad y^{(i)}(\theta \cdot x^{(i)}) \geq 1 \\ \\ -y^{(i)} \cdot x^{(i)} & si \quad y^{(i)}(\theta \cdot x^{(i)}) < 1 \end{cases}$$

$$\therefore \nabla_\theta Loss_h(y^{(i)}\theta \cdot x^{(i)}) = \begin{cases} -y^{(i)} \cdot x^{(i)} & si \quad y^{(i)}(\theta \cdot x^{(i)}) < 1 \\ \\ 0 & o.c \end{cases}$$

Que también se puede escribir como $I_{y^{(i)}(\theta \cdot x^{(i)})<1}(-y^{(i)} \cdot x^{(i)})$ con I como la función indicadora.

3.2. Si θ contiene los parámetros actuales. ¿Cuál es la regla de actualización de gradiente estocástico, donde $\eta > 0$ es la tasa de aprendizaje? (Elija todas las opciones que correspondan)

Las respuestas correctas serían b) y d).

$$\theta \to \theta - \eta \nabla_\theta [\mathbf{Loss}_h(y^{(i)}\theta \cdot x^{(i)})] - \eta\lambda\theta$$

$$\theta \to \theta - \eta \nabla_\theta [\mathbf{Loss}_h(y^{(i)}\theta \cdot x^{(i)})] - \eta \nabla_\theta [\frac{\lambda}{2}||\theta||^2]$$

para $x^{(i)}$ aleatorias con etiqueta $y^{(i)}$

3.3. Para η grande (es decir, η cercano a 1) y $0.5 < \lambda\eta < 1$. ¿cuál de las siguientes figuras corresponde a una única actualización de SGD realizada en respuesta al punto marcado '+' arriba?

Como se calsifica correctamente y fuera del margen $y^{(i)}(\theta \cdot x^{(i)}) > 1 \Rightarrow \nabla_\theta [\mathbf{Loss}_h(y^{(i)}\theta \cdot x^{(i)})] = 0$

$$\Rightarrow \theta \to \theta - \eta \nabla_\theta [\mathbf{Loss}_h(y^{(i)}\theta \cdot x^{(i)})] - \eta\lambda\theta$$

$$\Rightarrow \theta \to \theta - \eta * 0 - \eta\lambda\theta$$

$$\Rightarrow \theta \to \theta(1 - \eta\lambda)$$

Y como $0.5 < \lambda\eta < 1 \Rightarrow 0 < (1 - \eta\lambda) < \frac{1}{2}$

Entonces la actualización en θ sera lo que ya tenia, multiplicado por algo menor a $\frac{1}{2}$ es decir, se hara mas pequeña. Por lo tanto el margen se hara mas grande y no rotara.

La respuesta que considero correcta es e)

3.4. ¿cuál de las siguientes figuras corresponde a una única actualización de SGD realizada en respuesta al punto marcado '+' arriba?

En esta ocasión, el punto esta dentro del margen. Por lo cual

$$y^{(i)}(\theta \cdot x^{(i)}) < 1$$

$$\Rightarrow \nabla_\theta[\mathbf{Loss}_h(y^{(i)}\theta \cdot x^{(i)})] = -y^{(i)} \cdot x^{(i)}$$

Esto implica que habra un cambio en θ y por lo tanto una rotación en el vector ortogonal a θ, como solo hay una respuesta presentando una rotación.

La respuesta que considero correcta es a)

4. ¿Cuál de las siguientes afirmaciones es correcta?

4.1. El gradiente calculado en el algoritmo de retropropagación consiste en las derivadas parciales de la función de pérdida con respecto a cada peso de red. Verdadero Falso Justifique.

Verdadero

El gradiente calculado en el algoritmo de retropropagación consiste en derivadas parciales de la función de pérdida respecto a los pesos de la red.

4.2. El descenso de gradiente estocástico difiere del (verdadero) descenso de gradiente en actualizar solo un peso de la red durante cada paso de descenso de gradiente. Verdadero Falso Justifique.

Falso

El descenso de gradiente estocástico unicamente hace actualizaciónes en una entrada del gradiente por iteración, mientras que el descenso por gradiente actualiza todo el gradiente de la función objetivo a minimizar.

4.3. Lo más importante en aprendizaje de máquina es clasificar de manera correcta el conjunto de datos de entrenamiento, Verdadero Falso Justifique.

Falso

Lo más importante en aprendizaje de máquina es la precisión de las predicciones. Si lo que más importara fuera el conjunto de datos de entrenamiento, se sobre entrenaría para los datos de entrenamiento y al hacer las predicciones estas presentarían muchas inconsistencias no deseadas.

5. Desarrolle

5.1. Explique en que consiste la maldición de la dimensionalidad

La maldición de la dimensionalidad ocurre cuando incorporamos variables/características a un modelo de clasificación. Lo lógico es pensar que añadir más características hará un modelo más preciso, lo cual podría ser cierto. Pero si añadimos más características, aumentamos la dimensión del modelo. Habrá más posibles combinaciones entre las características. Esto hace que aumente el número de datos con los que tenemos que entrenar el modelo. Este crecimiento es exponencial, por lo que añadir una sola característica nos hará necesitar muchos datos mas con los que posiblemente no contemos.

5.2. ¿Cuál es la función de regresión y que minimiza?

La función de regresión es $\hat{Y(x)} = E[Y|X = x]$ que por lo general es $\hat{Y} = x\theta + \theta_0$. Minimiza una función de error, por lo general el error cuadrático medio (ECM) $E[(Y - \hat{Y(x)})^2|X = x]$ sobre todas las posibles funciones $\hat{Y(x)}$ en las variables observadas $X = x$.

5.3. Explica en qué consiste el trade-off entre sesgo y varianza

El error de una aproximación, en particular el ECM se puede ver de la siguiente forma.

$$E[(Y - \hat{Y})^2] = E[(\hat{Y} - E[\hat{Y}])^2] + (E[\hat{Y}] - Y)^2$$

$$ECM[\hat{Y}] = \mathbf{VAR}(\hat{Y}) + \mathbf{Sesgo}(\hat{Y})^2$$

Es decir podemos ver el error como la suma de dos componentes. Donde el Sesgo es una medida de precisión, y la Varianza es una medida de dispersión. A la hora de modelar es casuístico, puede que haya modelos más complejos que parecería que podrían mostrar mejores resultados, es decir errores más pequeños. Pero a la hora de ejecutarlos, modelos simples suelen tener menor varianza, con mayor sesgo. Y por lo mismo un menor error.

6. Codificación en Python

Vease el archivo magic_loop_gridsearch.ipynb