Aprendizaje estadístico

Formulario · Primavera 2021

Carlos Lezama

EST · 25134
ITAM

Introducción

¿Cuándo necesitamos aprendizaje de máquina?

Dos aspectos de un problema dado pueden requerir el uso de programas que aprendan y mejoren sobre la base de su "experiencia": la **complejidad** del problema y la necesidad de **adaptabilidad**.

Complejidad

- Tareas realizadas por animales o humanos.
- Tareas más allá de las capacidades humanas.

Adaptabilidad

Una característica limitante de las herramientas programadas es su rigidez: una vez que el programa se ha escrito e instalado, permanece sin cambios. Sin embargo, muchas tareas cambian con el tiempo, o de un usuario a otro. Las herramientas de aprendizaje de máquina (programas cuyo comportamiento se adapta a sus datos de entrada) ofrecen una solución a estos problemas; estos son, por naturaleza, adaptables a los cambios en el entorno con el que interactúan.

Tipos de aprendizaje

- Supervisado o no supervisado.
- Por refuerzo.
- Agentes pasivos o activos.
- Maestro.
- Protocolo por bloques o continuo.

1. Marco formal de aprendizaje

Conjunto de dominio ($\mathcal{X}$)

$$\mathcal{X} \subseteq \mathbb{R}^d \quad \text{tal que} \quad d < \infty.$$

Conjunto de etiquetas ($\mathcal{Y}$)

En el caso de etiquetado binario,

$$\mathcal{Y} = \{0, 1\} \quad \text{ó} \quad \mathcal{Y} = \{-1, +1\}.$$

Conjunto de entrenamiento (S)

Una sucesión $S = \{(x_i, y_i)\}_{i=1}^m$ tal que $m < \infty$ y $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$.

Reglas de predicción (h)

$$h : \mathcal{X} \to \mathcal{Y},$$

también llamado *predictor*, *hipótesis* o *clasificador*.

Algoritmo de aprendizaje (A)

Denotamos $A(S)$ a la hipótesis que el algoritmo de aprendizaje A genera al observar el conjunto de entrenamiento S. Asimismo, asumimos que $\mathcal{X}$ tiene una medida de probabilidad desconocida $\mathcal{D}$ y que existe una función desconocida f que etiqueta los datos de manera correcta, es decir:

$$\exists f : \mathcal{X} \to \mathcal{Y} \quad \text{tal que} \quad f(x_i) = y_i, \ \forall i.$$

Métricas de éxito

Definición 1.1 (Error de un clasificador). *Dado un subconjunto de dominio $A \subseteq \mathcal{X}$ y su probabilidad de observarlo $\mathcal{D}(A)$. En muchos casos, nos referimos a A como un evento y lo expresamos usando una función $\pi : \mathcal{X} \to \{0, 1\}$ tal que $A = \{x \in \mathcal{X} : \pi(x) = 1\}$. Así pues, definimos el **error del clasificador** $h : \mathcal{X} \to \mathcal{Y}$ como sigue:*

$$L_{\mathcal{D},f}(h) \overset{\text{def}}{=} \mathop{\mathbb{P}}_{x \sim \mathcal{D}} [h(x) \neq f(x)] \overset{\text{def}}{=} \mathcal{D}(\{x : h(x) \neq f(x)\}).$$

*También se le conoce como **error de generalización** o **riesgo** de h.*

1.1. Minimización de riesgo empírico

El objetivo de nuestro algoritmo es encontrar $h_S : \mathcal{X} \to \mathcal{Y}$ que minimice el error con respecto a las desconocidas $\mathcal{D}$ y f.

Definición 1.2 (Error de entrenamiento). *El error en el que incurre el clasificador sobre el conjunto de entrenamiento está dado por:*

$$L_S(h) \overset{\text{def}}{=} \frac{|i \in [m] : h(x_i) \neq y_i|}{m},$$

donde

$$[m] = \{1, \ldots, m\}.$$

*A este **error de entrenamiento** también se le conoce como **error empírico** o **riesgo empírico**.*

Definición 1.3 (Minimización de riesgo empírico). *Al predictor h que minimiza $L_S(h)$ se le conoce como **minimizador de riesgo empírico** o **ERM**, por sus siglas en inglés.*

Definición 1.4 (Sobreajuste). *Cuando un sistema se sobreentrena, o se entrena con datos extraños, el algoritmo de aprendizaje puede quedar ajustado a unas características muy específicas de los datos de entrenamiento que no tienen relación causal con la función objetivo. A este fenómeno le decimos **sobreajuste**.*

1.2. Minimización de riesgo empírico con sesgo inductivo

Definición 1.5 (Familia de hipótesis). *Sea Ω el espacio de todos los clasificadores, definimos una **familia** (o clase) **de hipótesis** como el conjunto $\mathcal{H} \subset \Omega$ que el agente elige de antemano como un espacio de búsqueda restringido.*

Dada una clase $\mathcal{H}$ y un conjunto de entrenamiento S, el agente $\text{ERM}_{\mathcal{H}}$ usa la regla de *minimización de riesgo empírico* para escoger un predictor $h \in \mathcal{H}$ con el menor error posible sobre S.

$$\text{i.e.} \quad \text{ERM}_{\mathcal{H}}(S) \in \mathop{\arg\min}_{h \in \mathcal{H}} L_S(h).$$

Así pues, a tales restricciones se les conoce como **sesgo inductivo**.

1.3. Familia de hipótesis finita

Observación 1.1

Denotamos por h_S al resultado generado de aplicar $\text{ERM}_{\mathcal{H}}$ al conjunto de entrenamiento S.

$$\text{i.e.} \quad h_S \in \mathop{\arg\min}_{h \in \mathcal{H}} L_S(h).$$

Definición 1.6 (Hipótesis de realizabilidad). *Existe $h^* \in \mathcal{H}$ tal que $L_{\mathcal{D},f}(h^*) = 0$. Asimismo, podemos decir que $L_S(h^*) = 0$ con probabilidad 1 sobre S cuando las muestras de S se distribuyen $\mathcal{D}$ y se etiquetan f.*

Definición 1.7 (Hipótesis de independencia y distribución idéntica). *Asumimos que cada $x_i \in S$ se distribuye $\mathcal{D}$, es decir:*

$$S \sim \mathcal{D}^m,$$

donde $m = |S|$.

Observación 1.2: Parámetro de confianza

Denotamos por δ la probabilidad de obtener una muestra poco representativa de $\mathcal{D}$. Así pues, podemos ver a $(1 - \delta)$ como nuestro **parámetro de confianza**.

Observación 1.3: Parámetro de precisión

Denotamos por ε a la probabilidad de encontrar errores de etiquetado. Es decir, interpretamos $L_{\mathcal{D},f}(h_S) > \varepsilon$ como un fracaso para el agente y $L_{\mathcal{D},f}(h_S) \leq \varepsilon$ como un clasificador aproximadamente correcto.

Lema 1.1

Para dos conjuntos cualesquiera A y B, y una distribución $\mathcal{D}$, tenemos:

$$\mathcal{D}(A \cup B) \leq \mathcal{D}(A) + \mathcal{D}(B).$$

Corolario 1.1

Sean $\mathcal{H}$ una familia de hipótesis finita, $\delta \in (0, 1)$, $\varepsilon > 0$ y m un entero, se cumple:

$$m \geq \frac{\log(|\mathcal{H}|/\delta)}{\varepsilon}.$$

Entonces, para toda función de etiquetado f y toda distribución $\mathcal{D}$, para las cuales se cumple la *hipótesis de realizabilidad*, tenemos — con probabilidad de al menos $(1-\delta)$ sobre nuestra muestra independiente e idénticamente distribuida S de tamaño m — que toda solución del minimizador de riesgo empírico h_S satisface:

$$L_{\mathcal{D},f}(h_S) \leq \varepsilon.$$

2. Un modelo formal de aprendizaje

2.1. Aprendizaje PAC

Definición 2.1 (Aprendizaje PAC). *Una familia de hipótesis $\mathcal{H}$ es **PAC**[†] **aprendible** si existe una función $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ y un algoritmo de aprendizaje con la siguiente propiedad: para toda $\varepsilon, \delta \in (0,1)$, toda distribución $\mathcal{D}$ sobre $\mathcal{X}$ y toda función de etiquetado $f : \mathcal{X} \to \{0,1\}$, si se satisface la hipótesis de realizabilidad con respecto a $\mathcal{H}$, $\mathcal{D}$ y f, al ejecutar el algoritmo de aprendizaje con $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ elementos independientes e idénticamente distribuidos por $\mathcal{D}$ y etiquetados por f, el algoritmo genera una hipótesis h tal que, con probabilidad de al menos $(1-\delta)$ sobre la elección de elementos, $L_{\mathcal{D},f}(h) \leq \varepsilon$.*

[†] Probablemente aproximadamente correcto, del inglés: *Probably Approximately Correct*.

Definición 2.2 (Complejidad muestral). *La **complejidad muestral** de un algoritmo de aprendizaje representa la cantidad de muestras de entrenamiento que necesita para aprender con* éxito *una función objetivo.*

Corolario 2.1

Toda familia finita de hipótesis es PAC aprendible con complejidad muestral:

$$m_{\mathcal{H}}(\varepsilon, \delta) \leq \left\lceil \frac{\log(|\mathcal{H}|/\delta)}{\varepsilon} \right\rceil .$$

2.2. Aprendizaje PAC agnóstico

Definición 2.3 (Error verdadero de un clasificador). *Para alguna distribución de probabilidad $\mathcal{D}$ sobre $\mathcal{X} \times \mathcal{Y}$ se puede medir la probabilidad de que h cometa un error cuando los puntos etiquetados se extraen aleatoriamente con respecto a $\mathcal{D}$. Así pues, definimos el **error** (o riesgo) **verdadero del clasificador** h, como sigue:*

$$L_{\mathcal{D}}(h) \stackrel{\text{def}}{=} \mathop{\mathbb{P}}_{(x,y)\sim\mathcal{D}}[h(x) \neq y] \stackrel{\text{def}}{=} \mathcal{D}(\{(x,y) : h(x) \neq y\}) .$$

Nota: la definición de **error de entrenamiento** no se modifica.

Definición 2.4 (Clasificador bayesiano óptimo). *Dada una distribución de probabilidad $\mathcal{D}$ sobre $\mathcal{X} \times \{0,1\}$, el mejor clasificador es:*

$$f_{\mathcal{D}}(x) = \begin{cases} 1, & \text{si } \mathbb{P}[y = 1 \mid x] \geq 1/2 \\ 0, & \text{en otros casos} \end{cases} .$$

Observación 2.1

El *clasificador bayesiano* es óptimo porque cualquier otro clasificador $g : \mathcal{X} \to \{0,1\}$ tiene un error mayor.

$$\text{i.e.} \quad L_{\mathcal{D}}(f_{\mathcal{D}}) \leq L_{\mathcal{D}}(g), \ \forall g.$$

Definición 2.5 (Aprendizaje PAC agnóstico). *Una familia de hipótesis $\mathcal{H}$ es **PAC aprendible** en un sentido **agnóstico** si existe una función $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ y un algoritmo de aprendizaje con la siguiente propiedad: para toda $\varepsilon, \delta \in (0,1)$ y toda distribución $\mathcal{D}$ sobre $\mathcal{X} \times \mathcal{Y}$, al ejecutar el algoritmo de aprendizaje con $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ elementos independientes e idénticamente distribuidos por $\mathcal{D}$, el algoritmo genera una hipótesis h tal que, con probabilidad de al menos $(1-\delta)$ sobre la elección de m elementos de entrenamiento, $L_{\mathcal{D}}(h) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \varepsilon$.*

2.3. El alcance de los problemas de aprendizaje modelados

Observación 2.2: Error cuadrático medio

Podemos evaluar la calidad de un clasificador $h : \mathcal{X} \to \mathcal{Y}$ por el **error cuadrático medio** entre el etiquetado correcto y los valores predichos.

$$\text{i.e.} \quad L_{\mathcal{D}}(h) \stackrel{\text{def}}{=} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} (h(x) - y)^2 .$$

Definición 2.6 (Función de pérdida generalizada). *Dado cualquier conjunto $\mathcal{H}$ y algún dominio Z. ℓ es cualquier función de $\mathcal{H} \times Z$ a los reales no negativos;* i.e. *$\ell : \mathcal{H} \times Z \to \mathbb{R}_+$. A dichas funciones las llamamos **funciones de pérdida.***
Nótese que, para los problemas de predicción, $\mathcal{H}$ es nuestra familia de hipótesis y $Z = \mathcal{X} \times \mathcal{Y}$.

Definición 2.7 (Función de riesgo). *Definimos la **función de riesgo** como la pérdida esperada de un clasificador $h \in \mathcal{H}$ con respecto a una distribución de probabilidad $\mathcal{D}$ sobre Z. Es decir:*

$$L_{\mathcal{D}}(h) \stackrel{\text{def}}{=} \mathop{\mathbb{E}}_{z\sim\mathcal{D}} [\ell(h,z)] .$$

Definición 2.8 (Riesgo empírico). *Definimos el **riesgo empírico** como la pérdida esperada sobre una muestra $S = (z_1, \ldots, z_m) \in Z^m$. Es decir:*

$$L_S(h) \stackrel{\text{def}}{=} \frac{1}{m} \sum_{i=1}^{m} \ell(h, z_i).$$

Proposición 2.1

El *riesgo empírico* $L_S(h)$ es un estimador insesgado.

Definición 2.9 (Pérdida 0-1).

$$\ell_{0-1}(h,(x,y)) \stackrel{\text{def}}{=} \begin{cases} 0, & \text{si } h(x) = y \\ 1, & \text{si } h(x) \neq y \end{cases} .$$

Definición 2.10 (Pérdida cuadrática).

$$\ell_{\text{sq}}(h,(x,y)) \stackrel{\text{def}}{=} (h(x) - y)^2 .$$

Definición 2.11 (Aprendizaje PAC agnóstico para funciones de pérdida generalizada). *Una familia de hipótesis $\mathcal{H}$ es **PAC aprendible** en un sentido **agnóstico** con respecto a Z y $\ell : \mathcal{H} \times Z \to \mathbb{R}_+$ (una función de pérdida) si existe una función $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ y un algoritmo de aprendizaje con la siguiente propiedad: para toda $\varepsilon, \delta \in (0,1)$ y toda distribución $\mathcal{D}$ sobre Z, al ejecutar el algoritmo de aprendizaje con $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ elementos independientes e idénticamente distribuidos por $\mathcal{D}$, el algoritmo genera una hipótesis $h \in \mathcal{H}$ tal que, con probabilidad de al menos $(1-\delta)$ sobre la elección de m elementos de entrenamiento, $L_{\mathcal{D}}(h) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \varepsilon$, donde $L_{\mathcal{D}}(h) = \mathop{\mathbb{E}}_{z\sim\mathcal{D}}[\ell(h,z)]$*

Observación 2.3

En aprendizaje PAC agnóstico para funciones de pérdida generalizada, necesitamos que la función $\ell(h,\cdot)$ sea *medible*. Es decir, al asumir un σ-álgebra de subconjuntos de Z sobre el cuál la probabilidad $\mathcal{D}$ está definida, y que la preimagen de cada segmento inicial en $\mathbb{R}_+$ está en este σ-álgebra. En el caso específico de clasificación binaria con pérdida 0-1, la σ-álgebra está sobre $\mathcal{X} \times \{0,1\}$ y nuestra suposición en ℓ es equivalente a asumir que sobre toda h, el conjunto $\{(x, h(x)) : x \in \mathcal{X}\}$ está en el σ-álgebra.

Observación 2.4: Aprendizaje adecuado o independiente de la representación

En algunos casos, $\mathcal{H}$ es un subconjunto de algún conjunto $\mathcal{H}'$ y la función de pérdida puede extenderse a ser una función de $\mathcal{H}' \times Z$ a los reales. En este caso, podemos permitir que el algoritmo genere una hipótesis $h' \in \mathcal{H}'$ siempre y cuando cumpla con $L_{\mathcal{D}}(h') \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \varepsilon$. Al permitir que el algoritmo genere una hipótesis de $\mathcal{H}'$, se dice que dicho aprendizaje es de *representación independiente* o *inadecuado*; mientras que, en aprendizaje *adecuado*, el algoritmo debe producir una hipótesis de $\mathcal{H}$.

3. Convergencia uniforme

Definición 3.1 (Muestra ε-representativa). *Un conjunto de entrenamiento S es **ε-representativo** respecto al dominio Z, la familia de hipótesis $\mathcal{H}$, la función de pérdida ℓ y la distribución $\mathcal{D}$ si:*

$$\forall h \in \mathcal{H}, \quad |L_S(h) - L_{\mathcal{D}}(h)| \leq \varepsilon.$$

Lema 3.1

Al asumir S un conjunto de entrenamiento $\varepsilon/2$-representativo, cualquier $h_S \in \arg\min_{h \in \mathcal{H}} L_S(h)$ satisface:

$$L_{\mathcal{D}}(h_S) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \varepsilon.$$

Definición 3.2 (Convergencia uniforme). *Decimos que una familia de hipótesis $\mathcal{H}$ tiene la propiedad de **convergencia uniforme** con respecto al dominio Z y la función de pérdida ℓ si existe una función $m_{\mathcal{H}}^{\text{UC}} : (0,1)^2 \to \mathbb{N}$ tal que, para toda $\varepsilon, \delta \in (0,1)$ y toda distribución de probabilidad $\mathcal{D}$ sobre Z, si S es una muestra de $m \geq m_{\mathcal{H}}^{\text{UC}}$ elementos independientes e idénticamente distribuidos por $\mathcal{D}$, entonces, con probabilidad de al menos $(1-\delta)$, S es ε-representativa.*

Corolario 3.1

Si una familia de hipótesis $\mathcal{H}$ tiene convergencia uniforme con una función $m_{\mathcal{H}}^{\text{UC}}$, entonces la familia es PAC aprendible en sentido agnóstico con una complejidad muestral $m_{\mathcal{H}}(\varepsilon, \delta) \leq m_{\mathcal{H}}^{\text{UC}}(\varepsilon/2, \delta)$.

Lema 3.2: Desigualdad de Hoeffding

Sea $\theta_1, \ldots, \theta_m$ una sucesión de variables aleatorias independientes e idénticamente distribuidas. Al asumir que, para toda i, $\mathbb{E}[\theta_i] = \mu$ y $\mathbb{P}[a \leq \theta_i \leq b] = 1$; para toda $\varepsilon > 0$:

$$\mathbb{P}\left[\left|\frac{1}{m}\sum_{i=1}^{m}\theta_i - \mu\right| > \varepsilon\right] \leq 2\exp\left(\frac{-2m\varepsilon^2}{(b-a)^2}\right).$$

Corolario 3.2

Sean $\mathcal{H}$ un clase de hipótesis finita, Z el dominio y $\ell : \mathcal{H} \times Z \to [0,1]$ una función de pérdida. Entonces, $\mathcal{H}$ posee la propiedad de convergencia uniforme con complejidad muestral:

$$m_{\mathcal{H}}^{\text{UC}}(\varepsilon, \delta) \leq \left\lceil \frac{\log(2|\mathcal{H}|/\delta)}{2\varepsilon^2} \right\rceil.$$

Asimismo, la clase es PAC aprendible en un sentido agnóstico al utilizar el algoritmo ERM con complejidad muestral:

$$m_{\mathcal{H}}(\varepsilon, \delta) \leq m_{\mathcal{H}}^{\text{UC}}(\varepsilon/2, \delta) \leq \left\lceil \frac{2\log(2|\mathcal{H}|/\delta)}{\varepsilon^2} \right\rceil.$$

4. Predictores lineales

Definición 4.1 (Clase de transformaciones afines). *Definimos la* ***clase de transformaciones afines*** *como:*

$$L_d = \{h_{\mathrm{w},b} : \mathrm{w} \in \mathbb{R}^d, b \in \mathbb{R}\},$$

donde

$$h_{\mathrm{w},b}(\mathrm{x}) = \langle \mathrm{w}, \mathrm{x} \rangle + b = \left(\sum_{i=1}^{d} w_i x_i \right) + b.$$

Asimismo, es conveniente usar la siguiente notación:

$$L_d = \{\mathrm{x} \mapsto \langle \mathrm{w}, \mathrm{x} \rangle + b : \mathrm{w} \in \mathbb{R}^d, b \in \mathbb{R}\}.$$

Observación 4.1

En algunos casos, resulta conveniente incorporar el *sesgo* b en w como una coordenada y agregar una coordenada extra con valor de 1 a todas las $\mathrm{x} \in \mathcal{X}$. Es decir, sean $\mathrm{w}' = (b, w_1, w_2, \ldots, w_d) \in \mathbb{R}^{d+1}$ y $\mathrm{x}' = (1, x_1, x_2, \ldots, x_d) \in \mathbb{R}^{d+1}$. Por lo tanto:

$$h_{\mathrm{w},b}(\mathrm{x}) = \langle \mathrm{w}, \mathrm{x} \rangle + b = \langle \mathrm{w}', \mathrm{x}' \rangle .$$

4.1. Semiespacios

Definición 4.2 (Clase de semiespacios). *La* ***clase de semiespacios****, diseñada para problemas de clasificación binaria con* $\mathcal{X} = \mathbb{R}^d$ *y* $\mathcal{Y} = \{-1, +1\}$*, la definimos como:*

$$HS_d = \mathrm{sgn} \circ L_d = \{\mathrm{x} \mapsto \mathrm{sgn}(h_{\mathrm{w},b}(\mathrm{x})) : h_{\mathrm{w},b} \in L_d\}.$$

Podemos considerar las siguientes tres situaciones:

1. Casos separables, al asumir que se cumple la *hipótesis de realizabilidad*.
2. Casos no separables en un sentido agnóstico.
3. Cuando una función lineal no es suficiente y necesitamos una transformación no lineal.

4.1.1. Programación lineal

Definición 4.3 (Programas lineales). *Los* ***programas lineales*** *son problemas que pueden expresarse como la maximización de una función lineal sujeta a restricciones lineales. Es decir:*

$$\begin{aligned} \max_{\mathrm{w} \in \mathbb{R}^d} \quad & \langle \mathrm{u}, \mathrm{w} \rangle \\ \text{sujeto a} \quad & A\mathrm{w} \geq \mathrm{v}, \end{aligned}$$

donde w *es el vector de variables que deseamos determinar,* A *es una matriz* $m \times d$ *y* $\mathrm{v} \in \mathbb{R}^m$, $\mathrm{u} \in \mathbb{R}^d$ *son vectores.*

Proposición 4.1

Sea $S = \{(\mathrm{x}_i, y_i)\}_{i=1}^m$ un conjunto de entrenamiento de tamaño m. Al asumir casos separables, existe $\mathrm{w} \in \mathbb{R}^d$ tal que:

$$y_i \langle \mathrm{w}, \mathrm{x}_i \rangle \geq 1, \quad \forall i = 1, \ldots, m,$$

o bien,

$$A\mathrm{w} \geq \mathrm{v},$$

donde $A \in \mathbb{R}^{m \times d}$, $A_{i,j} = y_i \, x_{i,j}$ y $\mathrm{v} = (1, \ldots, 1) \in \mathbb{R}^m$; y w es un predictor ERM.

4.1.2. Algoritmo de perceptrones

Input: Un conjunto de entrenamiento $S = \{(\mathrm{x}_i, y_i)\}_{i=1}^m$

```
w^(1) = (0, ..., 0)
for t = 1, 2, ... do
    if ∃i tal que y_i ⟨w^(t), x_i⟩ ≤ 0 then
        w^(t+1) = w^(t) + y_i x_i
    else
        Output: w^(t)
    end
end
```

Algoritmo 1: Perceptrón por bloques

Teorema 4.1

Al asumir $\{(\mathrm{x}_i, y_i)\}_{i=1}^m$ separables, y sean:

$$B = \min \{\|\mathrm{w}\| : y_i \langle \mathrm{w}, \mathrm{x}_i \rangle \geq 1\},$$

y

$$R = \max_i \{\|\mathrm{x}_i\|\}.$$

Entonces, el algortimo de perceptrones se detiene a lo más $(RB)^2$ iteraciones después y devuelve $\mathrm{w}^{(t)}$ tal que $y_i \langle \mathrm{w}^{(t)}, \mathrm{x}_i \rangle > 0$.