Conceptos básicos

Lección 02

Dr. Pablo Alvarado Moya

CE5506 Introducción al reconocimiento de patrones
Área de Ingeniería en Computadores
Tecnológico de Costa Rica

II Semestre, 2019

Contenido

Aprendizaje supervisado

- Aprendizaje **supervisado**: métodos entrenados con
  - conjunto de entrenamiento: pares ordenados $(\underline{\mathbf{x}}^{(i)}, y^{(i)})$,
  - $\underline{\mathbf{x}}^{(i)}$ es el i-ésimo vector de entrada y
  - $y^{(i)}$ es la correspondiente etiqueta (*label*) "correcta" que se desea predecir posteriormente.
- Es el tipo de aprendizaje automático más común
- Surgen empresas en venta de datos para entrenamiento
- Ejemplo de hace $\approx$ 30 años (Carnegie Mellon):
  - Carnegie Mellon 80s: ALVINN

Regresión

- ALVINN muestra problema de **regresión**

Regresión

- ALVINN muestra problema de **regresión**
- Regresión: produce valores continuos

Regresión

- ALVINN muestra problema de **regresión**
- Regresión: produce valores continuos
- Por ejemplo: ALVINN aprende valores para ajustar dirección

Regresión
Precios de casa en Escazú

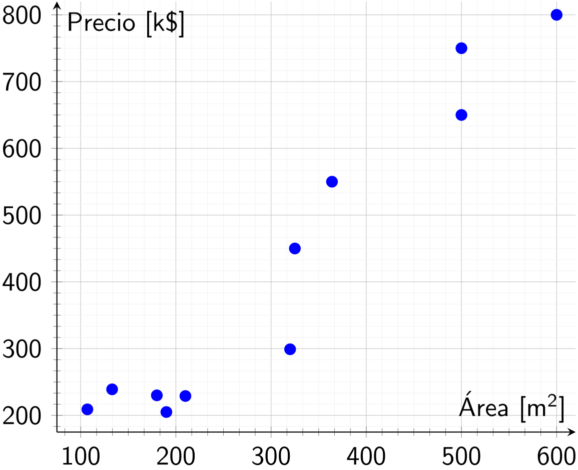
Área [m^2]	Plantas	Hab.	Precio [k$]
600	3	5	800
190	2	2	205
210	2	2	229
364	2	2	550
325	2	4	450
180	2	2	230
133	2	2	239
500	2	3	650
107	1	2	209
320	2	3	299
500	2	4	750

Regresión

Precios de casa en Escazú

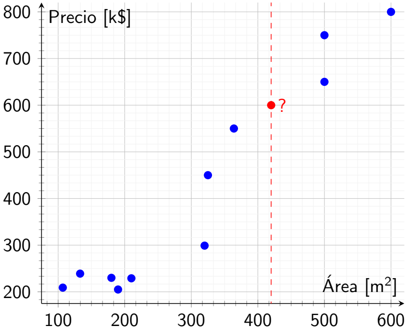
Área [m^2]	Plantas	Hab.	Precio [k$]
600	3	5	800
190	2	2	205
210	2	2	229
364	2	2	550
325	2	4	450
180	2	2	230
133	2	2	239
500	2	3	650
107	1	2	209
320	2	3	299
500	2	4	750

Problema en una dimensión

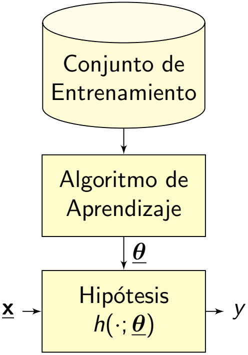


Problema

Dado un conjunto de entrenamiento como el anterior, ¿cómo se puede encontrar relación de salida (*precio*) en términos de la entrada (*área*)



Notación y modelo supervisado



- m: número de ejemplos de entrenamiento
- $\underline{\mathbf{x}}$: variables de entrada (x si es un escalar)
- n: dimensión de la entrada $\underline{\mathbf{x}}$ (número de características)
- y: variable de salida u objetivo (*target*)
  - **Clasificación**: $y \in \{C_1, \ldots C_k\}, k \in \mathbb{N}$
  - **Regresión**: $y \in \mathbb{R}$
- $(\underline{\mathbf{x}}^{(i)}, y^{(i)})$: i-ésimo ejemplo de entrenamiento
- $\underline{\boldsymbol{\theta}}$: parámetros

Hipótesis para regresión lineal

- Ejemplo de hipótesis: regresión lineal

$$y = h(\underline{\mathbf{x}}; \underline{\boldsymbol{\theta}}) = h_{\underline{\boldsymbol{\theta}}}(\underline{\mathbf{x}}) = \theta_0 + \theta_1 x_1 + \ldots + \theta_n x_n$$

- En ejemplo $n = 3$ con
  - x_1: área de casa
  - x_2: # de habitaciones
  - x_3: # de pisos
- Convención para simplificar notación: $x_0 = 1$

$$y = h(\underline{\mathbf{x}}; \underline{\boldsymbol{\theta}}) = h_{\underline{\boldsymbol{\theta}}}(\underline{\mathbf{x}}) = \theta_0 x_0 + \theta_1 x_1 + \ldots + \theta_n x_n = \sum_{i=0}^{n} \theta_i x_i$$

$$= \underline{\boldsymbol{\theta}}^T \underline{\mathbf{x}} = \langle \underline{\boldsymbol{\theta}}, \underline{\mathbf{x}} \rangle = \underline{\boldsymbol{\theta}} \cdot \underline{\mathbf{x}}$$

$$\underline{\boldsymbol{\theta}} = [\theta_0, \theta_1, \ldots, \theta_n]^T$$

$$\underline{\mathbf{x}} = [x_0, x_1, \ldots, x_n]^T$$

Función objetivo y minimización de cuadrados

- Para encontrar $\underline{\boldsymbol{\theta}}$ minimizamos función de error $J(\underline{\boldsymbol{\theta}})$ con

$$J(\underline{\boldsymbol{\theta}}) = \frac{1}{2}\sum_{i=1}^{m}\left(h_\theta(\underline{\mathbf{x}}^{(i)}) - y^{(i)}\right)^2$$

- El factor $1/2$ se coloca por conveniencia
- Planteamos problema de optimización de mínimos cuadrados ordinarios (*OLS, ordinary least squares*):

$$\underline{\boldsymbol{\theta}}^* = \arg\min_{\underline{\boldsymbol{\theta}}} J(\underline{\boldsymbol{\theta}})$$

Se buscan parámetros $\underline{\boldsymbol{\theta}}$ que producen el menor valor de $J(\underline{\boldsymbol{\theta}})$

Ejemplo de regresión de precios de casas

El caso general de regresión **lineal** minimiza entonces a

$$\begin{aligned} J(\underline{\boldsymbol{\theta}}) &= \frac{1}{2}\sum_{i=1}^{m}\left(h_\theta(\underline{\mathbf{x}}^{(i)}) - y^{(i)}\right)^2 \\ &= \frac{1}{2}\sum_{i=1}^{m}\left(\underline{\boldsymbol{\theta}}^T\underline{\mathbf{x}}^{(i)} - y^{(i)}\right)^2 \end{aligned}$$

y para el caso de precio$=f$(área)

$$J(\theta_0,\theta_1) = \frac{1}{2}\sum_{i=1}^{m}\left(\theta_0 + \theta_1 x_1^{(i)} - y^{(i)}\right)^2$$

Ver ejemplo `fobj.m`

Minimización de la función objetivo

- Hay varias posibilidades para minimizar $J(\underline{\boldsymbol{\theta}})$
- En general, las técnicas de aprendizaje
  - Toman un valor inicial de $\underline{\boldsymbol{\theta}}$ (p. ej. $\underline{\mathbf{0}}$)
  - Modifican iterativamente $\underline{\boldsymbol{\theta}}$ para reducir $J(\underline{\boldsymbol{\theta}})$

Descenso por gradiente
Gradient descent

- Caso particular **descenso por gradiente**:
  1. Tome un valor $\underline{\theta}^{(0)}$ inicial, con $t = 0$

Descenso por gradiente
Gradient descent

- Caso particular **descenso por gradiente**:
  1. Tome un valor $\underline{\boldsymbol{\theta}}^{(0)}$ inicial, con $t = 0$
  2. Calcule en $\underline{\boldsymbol{\theta}}^{(t)}$ el gradiente (máxima dirección de cambio)
  $\nabla_{\underline{\boldsymbol{\theta}}} J(\underline{\boldsymbol{\theta}}^{(t)}) = \begin{bmatrix} \frac{\partial J}{\partial \theta_0} & \frac{\partial J}{\partial \theta_1} & \cdots & \frac{\partial J}{\partial \theta_n} \end{bmatrix}^T$

Descenso por gradiente
Gradient descent

- Caso particular **descenso por gradiente**:
  1. Tome un valor $\underline{\boldsymbol{\theta}}^{(0)}$ inicial, con $t = 0$
  2. Calcule en $\underline{\boldsymbol{\theta}}^{(t)}$ el gradiente (máxima dirección de cambio)
     $\nabla_{\underline{\boldsymbol{\theta}}} J(\underline{\boldsymbol{\theta}}^{(t)}) = \left[\frac{\partial J}{\partial \theta_0} \quad \frac{\partial J}{\partial \theta_1} \quad \cdots \quad \frac{\partial J}{\partial \theta_n}\right]^T$
  3. Calcule la nueva posición

$$\underline{\boldsymbol{\theta}}^{(t+1)} := \underline{\boldsymbol{\theta}}^{(t)} - \alpha \nabla J(\underline{\boldsymbol{\theta}}^{(t)})$$

o de forma equivalente para cada θ_j, $j \in 1 \ldots n$

$$\theta_j^{(t+1)} := \theta_j^{(t)} - \alpha \frac{\partial J(\underline{\boldsymbol{\theta}}^{(t)})}{\partial \theta_j}$$

Descenso por gradiente
Gradient descent

- Caso particular **descenso por gradiente**:
  1. Tome un valor $\underline{\boldsymbol{\theta}}^{(0)}$ inicial, con $t = 0$
  2. Calcule en $\underline{\boldsymbol{\theta}}^{(t)}$ el gradiente (máxima dirección de cambio)
     $\nabla_{\underline{\boldsymbol{\theta}}} J(\underline{\boldsymbol{\theta}}^{(t)}) = \left[\frac{\partial J}{\partial \theta_0} \quad \frac{\partial J}{\partial \theta_1} \quad \cdots \quad \frac{\partial J}{\partial \theta_n}\right]^T$
  3. Calcule la nueva posición

$$\underline{\boldsymbol{\theta}}^{(t+1)} := \underline{\boldsymbol{\theta}}^{(t)} - \alpha \nabla J(\underline{\boldsymbol{\theta}}^{(t)})$$

o de forma equivalente para cada θ_j, $j \in 1 \ldots n$

$$\theta_j^{(t+1)} := \theta_j^{(t)} - \alpha \frac{\partial J(\underline{\boldsymbol{\theta}}^{(t)})}{\partial \theta_j}$$

- Ejemplos: `peaksDescent.m`, `step_normalized.m`

Estrategias de parada

Para detener búsqueda de mínimo:

- Usualmente se utilizan tasas de cambio

Estrategias de parada

Para detener búsqueda de mínimo:

- Usualmente se utilizan tasas de cambio
- Primera opción: $J(\underline{\boldsymbol{\theta}}^{(t)}) - J(\underline{\boldsymbol{\theta}}^{(t+1)}) < \epsilon$

Estrategias de parada

Para detener búsqueda de mínimo:

- Usualmente se utilizan tasas de cambio
- Primera opción: $J(\underline{\boldsymbol{\theta}}^{(t)}) - J(\underline{\boldsymbol{\theta}}^{(t+1)}) < \epsilon$
    - hasta que el cambio del error sea menor que umbral ϵ

Estrategias de parada

Para detener búsqueda de mínimo:

- Usualmente se utilizan tasas de cambio
- Primera opción: $J(\underline{\boldsymbol{\theta}}^{(t)}) - J(\underline{\boldsymbol{\theta}}^{(t+1)}) < \epsilon$
  - hasta que el cambio del error sea menor que umbral ϵ
  - nótese que si error crece, también se detiene

Estrategias de parada

Para detener búsqueda de mínimo:

- Usualmente se utilizan tasas de cambio
- Primera opción: $J(\underline{\boldsymbol{\theta}}^{(t)}) - J(\underline{\boldsymbol{\theta}}^{(t+1)}) < \epsilon$
  - hasta que el cambio del error sea menor que umbral ϵ
  - nótese que si error crece, también se detiene
- Segunda opción: $\|\underline{\boldsymbol{\theta}}^{(t)} - \underline{\boldsymbol{\theta}}^{(t+1)}\| < \epsilon$

Estrategias de parada

Para detener búsqueda de mínimo:

- Usualmente se utilizan tasas de cambio
- Primera opción: $J(\underline{\boldsymbol{\theta}}^{(t)}) - J(\underline{\boldsymbol{\theta}}^{(t+1)}) < \epsilon$
  - hasta que el cambio del error sea menor que umbral ϵ
  - nótese que si error crece, también se detiene
- Segunda opción: $\|\underline{\boldsymbol{\theta}}^{(t)} - \underline{\boldsymbol{\theta}}^{(t+1)}\| < \epsilon$
  - hasta que cambio de posición sea inferior a umbral ϵ

Estrategias de parada

Para detener búsqueda de mínimo:

- Usualmente se utilizan tasas de cambio
- Primera opción: $J(\underline{\boldsymbol{\theta}}^{(t)}) - J(\underline{\boldsymbol{\theta}}^{(t+1)}) < \epsilon$
  - hasta que el cambio del error sea menor que umbral ϵ
  - nótese que si error crece, también se detiene
- Segunda opción: $\|\underline{\boldsymbol{\theta}}^{(t)} - \underline{\boldsymbol{\theta}}^{(t+1)}\| < \epsilon$
  - hasta que cambio de posición sea inferior a umbral ϵ
- Tercera opción: Número máximo de iteraciones

Estrategias de parada

Para detener búsqueda de mínimo:

- Usualmente se utilizan tasas de cambio
- Primera opción: $J(\underline{\boldsymbol{\theta}}^{(t)}) - J(\underline{\boldsymbol{\theta}}^{(t+1)}) < \epsilon$
    - hasta que el cambio del error sea menor que umbral ϵ
    - nótese que si error crece, también se detiene
- Segunda opción: $\|\underline{\boldsymbol{\theta}}^{(t)} - \underline{\boldsymbol{\theta}}^{(t+1)}\| < \epsilon$
    - hasta que cambio de posición sea inferior a umbral ϵ
- Tercera opción: Número máximo de iteraciones
- Opción usual: combinación de anteriores

Normalización de datos

- Si el gradiente es fuertemente asimétrico (como en el caso actual), la tasa de aprendizaje α debe elegirse muy pequeña y proceso necesitará demasiadas iteraciones para converger
- ¡Datos deben normalizarse (preprocesamiento) para evitar estos problemas!

Descenso de gradiente para regresión lineal (1)

Cálculo del gradiente

Partiendo del caso concreto:

$$J(\theta_0,\theta_1) = \frac{1}{2}\sum_{i=1}^{m}\left(\theta_0 + \theta_1 x_1^{(i)} - y^{(i)}\right)^2$$

podemos calcular el gradiente fácilmente

$$\begin{aligned}\nabla_{\underline{\boldsymbol{\theta}}} J(\theta_0,\theta_1) &= \begin{bmatrix} \frac{\partial J(\theta_0,\theta_1)}{\partial \theta_0} \\ \frac{\partial J(\theta_0,\theta_1)}{\partial \theta_1} \end{bmatrix} \\ &= \begin{bmatrix} \sum_{i=1}^{m}\left(\theta_0 + \theta_1 x_1^{(i)} - y^{(i)}\right)\cdot 1 \\ \sum_{i=1}^{m}\left(\theta_0 + \theta_1 x_1^{(i)} - y^{(i)}\right)\cdot x_1^{(i)} \end{bmatrix}\end{aligned}$$

Descenso de gradiente para regresión lineal (2)

Cálculo del gradiente

Observe que para el caso general de regresión lineal se tiene

$$J(\underline{\boldsymbol{\theta}}) = \frac{1}{2}\sum_{i=1}^{m}\left(\underline{\boldsymbol{\theta}}^T\underline{\mathbf{x}}^{(i)} - y^{(i)}\right)^2$$
$$= \frac{1}{2}\sum_{i=1}^{m}\left((\theta_0 x_0 + \theta_1 x_1 + \ldots + \theta_n x_n) - y^{(i)}\right)^2$$

La j-ésima componente del gradiente $\nabla_{\underline{\boldsymbol{\theta}}} J(\underline{\boldsymbol{\theta}})$ es

$$\frac{\partial J(\underline{\boldsymbol{\theta}})}{\partial \theta_j} = \sum_{i=1}^{m}\left(\underline{\boldsymbol{\theta}}^T\underline{\mathbf{x}}^{(i)} - y^{(i)}\right) \cdot x_j^{(i)}$$

Descenso de gradiente para regresión lineal (3)

Cálculo del gradiente

lo que finalmente implica que el gradiente es

$$\nabla J(\underline{\boldsymbol{\theta}}) = \sum_{i=1}^{m} \left(\underline{\boldsymbol{\theta}}^T \underline{\mathbf{x}}^{(i)} - y^{(i)} \right) \underline{\mathbf{x}}^{(i)}$$

Descenso de gradiente por lotes

- Combinando todos los resultados anteriores tenemos el algoritmo de **descenso de gradiente por lotes** (*batch gradient descent*):

$$\theta_j^{(t+1)} := \theta_j^{(t)} - \alpha \sum_{i=1}^{m} \left(\underline{\boldsymbol{\theta}}^T \underline{\mathbf{x}}^{(i)} - y^{(i)} \right) \cdot x_j^{(i)}$$

- **Lotes**: cada paso usa **todo** el conjunto de entrenamiento
- α recibe el nombre de **tasa de aprendizaje** (*learning rate*)
- Su ajuste es "delicado":
  - Si α es muy grande, oscila alrededor de mínimo
  - Si α es muy pequeño, necesita muchos pasos para converger
- El descenso de gradiente converge a extremos locales, que dependen del punto inicial

Descenso por gradiente estocástico

Stochastic Gradient Descent

- Descenso por gradiente **estocástico** o **incremental** usa un ejemplo del conjunto de entrenamiento a la vez:

  1: **repeat**
  2:     **for** each $(\underline{\mathbf{x}}^{(i)}, y^{(i)})$ in training set **do**
  3:         $\underline{\boldsymbol{\theta}}^{(t+1)} := \underline{\boldsymbol{\theta}}^{(t)} - \alpha \left({\underline{\boldsymbol{\theta}}^{(t)}}^T \underline{\mathbf{x}}^{(i)} - y^{(i)} \right) \underline{\mathbf{x}}^{(i)}$
  4:         $t := t + 1$
  5:     **end for**
  6: **until** (convergence)

- No asegura convergencia, pero "se mueve" inmediatamente
- Trajectoria hacia el mínimo "divaga" pero en general se acerca al mínimo
- Útil para conjuntos de entrenamiento gigantescos
- Ejemplo: `stoch_all_steps_normalized.m`

Método por lotes contra estocástico

- Método estocástico produce soluciones acertadas más pronto
- Método por lotes es trivialmente paralelizable

Ecuaciones normales

- Caso particular de regresión lineal se resuelve con **ecuaciones normales**
- En curso de Análisis Numérico se derivaron algunas de ellas
- Extenderemos notación matemática de derivación de matrices, para facilitar cálculos

Derivadas con matrices y la traza

- Sean $\mathbf{A}$ una matriz de tamaño $m \times n$ y $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ una función que mapea matrices como $\mathbf{A}$ a valores reales. El gradiente de f es la matriz

$$\nabla_{\mathbf{A}} f(\mathbf{A}) = \begin{bmatrix} \frac{\partial f}{\partial a_{11}} & \cdots & \frac{\partial f}{\partial a_{1n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial a_{m1}} & \cdots & \frac{\partial f}{\partial a_{mn}} \end{bmatrix} \qquad \mathbf{A} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}$$

- La traza de la matriz cuadrada $\mathbf{A}$ de tamaño $n \times n$ es

$$\operatorname{tr} \mathbf{A} = \sum_{i=1}^{n} a_{ii}$$

- **Un escalar s (matriz 1×1) tiene $\operatorname{tr} s = s$**

Propiedades de la traza

- $\operatorname{tr}(\mathbf{A}+\mathbf{B}) = \operatorname{tr}\mathbf{A} + \operatorname{tr}\mathbf{B}$
- $\operatorname{tr}(c\mathbf{A}) = c\operatorname{tr}\mathbf{A}$
- $\operatorname{tr}\mathbf{A} = \operatorname{tr}\mathbf{A}^T$
- Si $\mathbf{AB}$ es cuadrada entonces $\operatorname{tr}\mathbf{AB} = \operatorname{tr}\mathbf{BA}$

$$\operatorname{tr}\mathbf{AB} = \sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}b_{ji} = \sum_{j=1}^{m}\sum_{i=1}^{n} b_{ji}a_{ij} = \operatorname{tr}\mathbf{BA}$$

- Corolarios:
    - $\operatorname{tr}\mathbf{ABC} = \operatorname{tr}\mathbf{CAB} = \operatorname{tr}\mathbf{BCA}$
    - $\operatorname{tr}\mathbf{ABCD} = \operatorname{tr}\mathbf{DABC} = \operatorname{tr}\mathbf{CDAB} = \operatorname{tr}\mathbf{BCDA}$

Combinando trazas con derivadas

1. $\nabla_{\mathbf{A}} \operatorname{tr} \mathbf{AB} = \mathbf{B}^T$
2. $\nabla_{\mathbf{A}^T} f(\mathbf{A}) = (\nabla_{\mathbf{A}} f(\mathbf{A}))^T$
3. $\nabla_{\mathbf{A}} \operatorname{tr} \mathbf{ABA}^T\mathbf{C} = \mathbf{CAB} + \mathbf{C}^T\mathbf{AB}^T$
4. $\nabla_{\mathbf{A}} |\mathbf{A}| = |\mathbf{A}|(\mathbf{A}^{-1})^T$

Combinando (2) y (3)

- $\nabla_{\mathbf{A}^T} \operatorname{tr} \mathbf{ABA}^T\mathbf{C} = \mathbf{B}^T\mathbf{A}^T\mathbf{C}^T + \mathbf{BA}^T\mathbf{C}$

Replanteando los mínimos cuadrados (1)

- Buscamos $\underline{\boldsymbol{\theta}}$ que minimiza $J(\underline{\boldsymbol{\theta}})$
- Reescribamos $J(\underline{\boldsymbol{\theta}})$ de forma matricial:
  - Sea $\mathbf{X}$ la **matriz de diseño** de tamaño $m \times n+1$

$$\mathbf{X} = \begin{bmatrix} \underline{\mathbf{x}}^{(1)\,T} \\ \underline{\mathbf{x}}^{(2)\,T} \\ \vdots \\ \underline{\mathbf{x}}^{(m)\,T} \end{bmatrix}$$

  - Sea $\underline{\mathbf{y}}$ el vector de valores objetivo:

$$\underline{\mathbf{y}} = \begin{bmatrix} y^{(1)} & y^{(2)} & \cdots & y^{(m)} \end{bmatrix}^T$$

Replanteando los mínimos cuadrados (2)

- Puesto que

$$\mathbf{X}\underline{\boldsymbol{\theta}} - \underline{\mathbf{y}} = \begin{bmatrix} {\underline{\mathbf{x}}^{(1)}}^T \underline{\boldsymbol{\theta}} \\ {\underline{\mathbf{x}}^{(2)}}^T \underline{\boldsymbol{\theta}} \\ \vdots \\ {\underline{\mathbf{x}}^{(m)}}^T \underline{\boldsymbol{\theta}} \end{bmatrix} - \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{bmatrix} = \begin{bmatrix} {\underline{\mathbf{x}}^{(1)}}^T \underline{\boldsymbol{\theta}} - y^{(1)} \\ {\underline{\mathbf{x}}^{(2)}}^T \underline{\boldsymbol{\theta}} - y^{(2)} \\ \vdots \\ {\underline{\mathbf{x}}^{(m)}}^T \underline{\boldsymbol{\theta}} - y^{(m)} \end{bmatrix}$$

Entonces usando $\underline{\mathbf{v}}^T \underline{\mathbf{v}} = \|\underline{\mathbf{v}}\|^2 = \sum_i v_i^2$

$$J(\underline{\boldsymbol{\theta}}) = \frac{1}{2} \sum_{i=1}^{m} \left({\underline{\mathbf{x}}^{(i)}}^T \underline{\boldsymbol{\theta}} - y^{(i)} \right)^2 = \frac{1}{2} (\mathbf{X}\underline{\boldsymbol{\theta}} - \underline{\mathbf{y}})^T (\mathbf{X}\underline{\boldsymbol{\theta}} - \underline{\mathbf{y}})$$

Replanteando los mínimos cuadrados (3)

- El mínimo se encuentra buscando $\nabla_{\underline{\boldsymbol{\theta}}} J(\underline{\boldsymbol{\theta}}) = 0$

$$\begin{aligned}\nabla_{\underline{\boldsymbol{\theta}}} J(\underline{\boldsymbol{\theta}}) &= \nabla_{\underline{\boldsymbol{\theta}}} \frac{1}{2}(\mathbf{X}\underline{\boldsymbol{\theta}} - \underline{\mathbf{y}})^T(\mathbf{X}\underline{\boldsymbol{\theta}} - \underline{\mathbf{y}}) \overset{!}{=} \underline{\mathbf{0}} \\ &= \frac{1}{2}\nabla_{\underline{\boldsymbol{\theta}}}\left(\underline{\boldsymbol{\theta}}^T\mathbf{X}^T\mathbf{X}\underline{\boldsymbol{\theta}} - \underline{\boldsymbol{\theta}}^T\mathbf{X}^T\underline{\mathbf{y}} - \underline{\mathbf{y}}^T\mathbf{X}\underline{\boldsymbol{\theta}} + \underline{\mathbf{y}}^T\underline{\mathbf{y}}\right)\end{aligned}$$

y puesto que término entre paréntesis es un escalar real

$$\begin{aligned}&= \frac{1}{2}\nabla_{\underline{\boldsymbol{\theta}}} \operatorname{tr}\left(\underline{\boldsymbol{\theta}}^T\mathbf{X}^T\mathbf{X}\underline{\boldsymbol{\theta}} - \underline{\boldsymbol{\theta}}^T\mathbf{X}^T\underline{\mathbf{y}} - \underline{\mathbf{y}}^T\mathbf{X}\underline{\boldsymbol{\theta}} + \underline{\mathbf{y}}^T\underline{\mathbf{y}}\right) \\ &= \frac{1}{2}\nabla_{\underline{\boldsymbol{\theta}}}\left(\operatorname{tr}\underline{\boldsymbol{\theta}}^T\mathbf{X}^T\mathbf{X}\underline{\boldsymbol{\theta}} - 2\operatorname{tr}\underline{\mathbf{y}}^T\mathbf{X}\underline{\boldsymbol{\theta}}\right)\end{aligned}$$

Replanteando los mínimos cuadrados (4)

y usando $\nabla_{\mathbf{A}^T} \operatorname{tr} \mathbf{A}\mathbf{B}\mathbf{A}^T\mathbf{C} = \mathbf{B}^T\mathbf{A}^T\mathbf{C}^T + \mathbf{B}\mathbf{A}^T\mathbf{C}$, con $\mathbf{A} = \underline{\boldsymbol{\theta}}^T$, $\mathbf{B} = \mathbf{X}\mathbf{X}^T$ y $\mathbf{C} = \mathbf{I}$

$$\begin{aligned}
\nabla_{\underline{\boldsymbol{\theta}}} J(\underline{\boldsymbol{\theta}}) &= \frac{1}{2}\nabla_{\underline{\boldsymbol{\theta}}}\left(\operatorname{tr} \underline{\boldsymbol{\theta}}^T\mathbf{X}^T\mathbf{X}\underline{\boldsymbol{\theta}} - 2\operatorname{tr} \underline{\mathbf{y}}^T\mathbf{X}\underline{\boldsymbol{\theta}}\right) \\
&= \frac{1}{2}\left(\mathbf{X}^T\mathbf{X}\underline{\boldsymbol{\theta}} + \mathbf{X}^T\mathbf{X}\underline{\boldsymbol{\theta}} - 2\mathbf{X}^T\underline{\mathbf{y}}\right) \\
&= \mathbf{X}^T\mathbf{X}\underline{\boldsymbol{\theta}} - \mathbf{X}^T\underline{\mathbf{y}} \\
&\stackrel{!}{=} \underline{\mathbf{0}}
\end{aligned}$$

Replanteando los mínimos cuadrados (5)

con lo que se obtienen las **ecuaciones normales**

$$\begin{aligned} \mathbf{X}^T\mathbf{X}\underline{\boldsymbol{\theta}} &= \mathbf{X}^T\underline{\mathbf{y}} \\ \underline{\boldsymbol{\theta}} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\underline{\mathbf{y}} \end{aligned}$$

- $\underline{\boldsymbol{\theta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\underline{\mathbf{y}}$ es la solución cerrada
- $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ se conoce como la matriz seudoinversa de $\mathbf{X}$ o seudoinversa de Moore-Penrose, denotada con $\mathbf{X}^\dagger$:

$$\underline{\boldsymbol{\theta}} = \mathbf{X}^\dagger\underline{\mathbf{y}}$$

Resumen

1. Aprendizaje Supervisado
2. Regresión Lineal
3. Descenso por gradiente
4. Ecuaciones normales

Este documento ha sido elaborado con software libre incluyendo LATEX, Beamer, GNUPlot, GNU/Octave, XFig, Inkscape, GNU-Make y Subversion en GNU/Linux