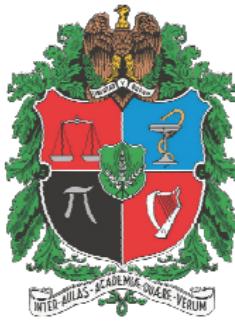


Modelo lineal y mínimos cuadrados

A. M. Alvarez-Meza, Ph.D.
amalvarezme@unal.edu.co

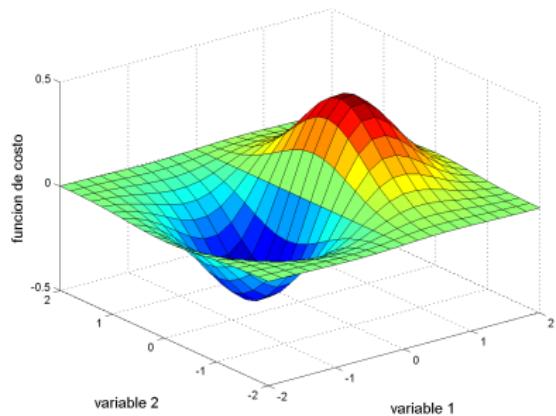
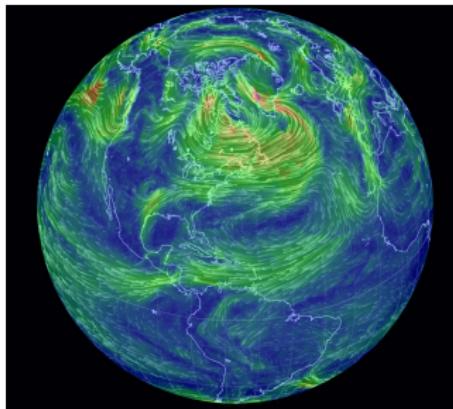
Departamento de ingeniería eléctrica, electrónica y computación
Universidad Nacional de Colombia-sede Manizales



Contenido

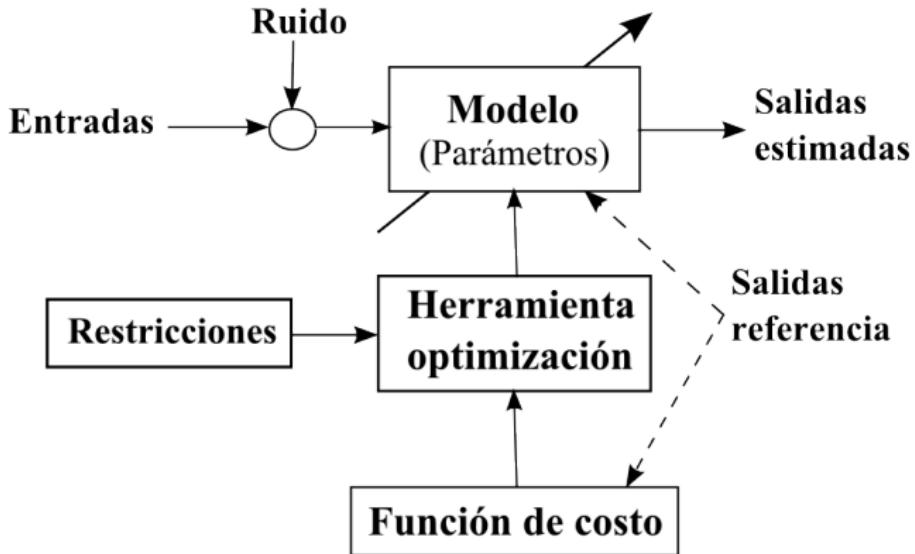
- 1 Nociones básicas de optimización
- 2 Modelo lineal y mínimos cuadrados
- 3 Modelo lineal regularizado
- 4 Extensión a representaciones no lineales

Por qué debemos resolver problemas de optimización?



Ciencia de datos, aprendizaje de máquina, minería de datos...
todos tienen que optimizar algo!

Aprendizaje de máquina y optimización



Buscamos información relevante en nuestros datos
a partir de la optimización de un modelo matemático

Optimización con y sin restricciones

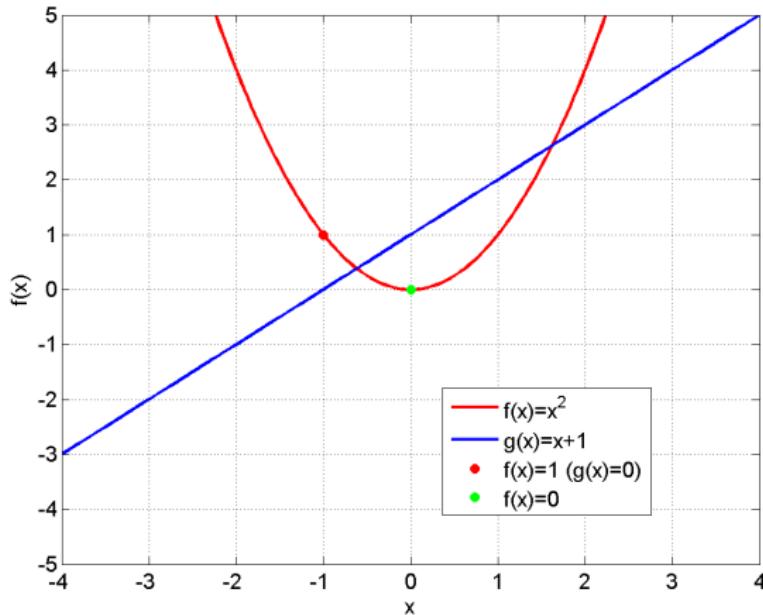
$$x^* = \arg \min_x f(x)$$

$$\text{s.t. } g_i(x) = c_i$$

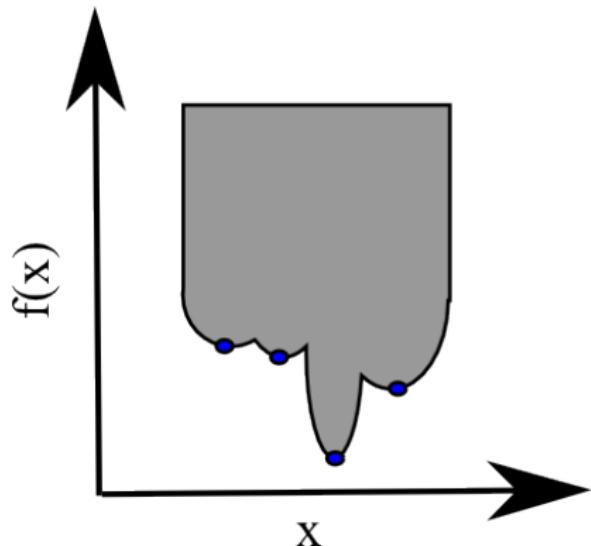
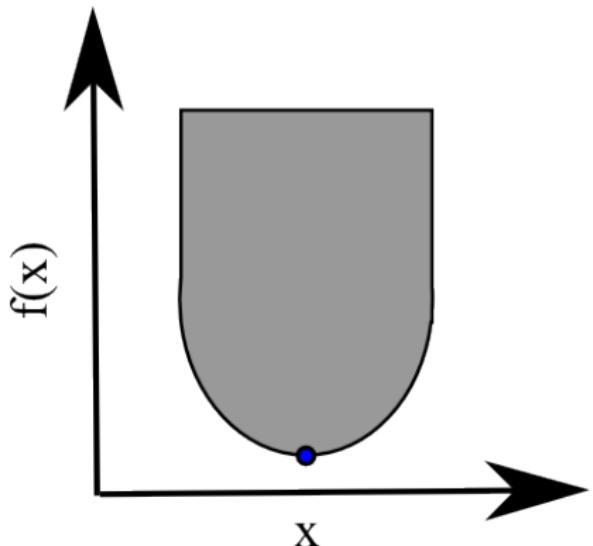
$$h_j(x) > d_j$$

- x^* : solución "óptima".
- $f(x)$: función de costo.
- $g_i(x)$: restricciones de igualdad ($i \in [1, D]$).
- $h_j(x)$: restricciones de desigualdad ($j \in [1, Q]$).
- x puede ser un escalar, un vector o una matriz.

Optimización con y sin restricciones



Convexa o no convexa?



- Función convexa: único mínimo.
- Función no convexa: mínimos locales.
- La convexidad del problema de optimización depende de la función de costo y de las restricciones.

La función Lagrangiano

$$L(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) - \sum_{i=1}^D \lambda_i g_i(\mathbf{x})$$

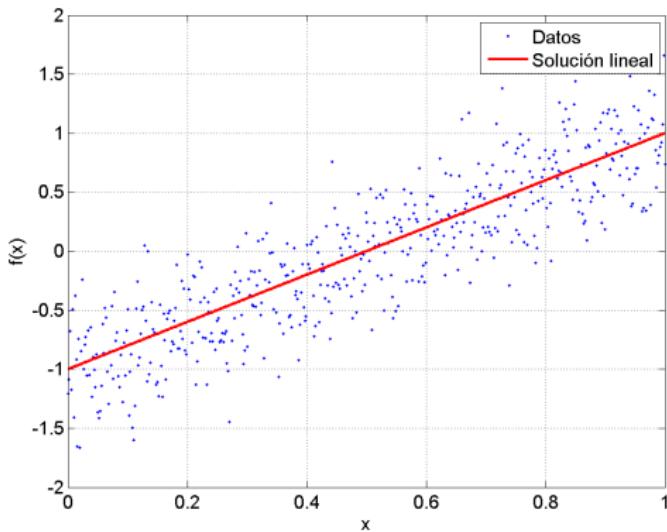
- Relaciona el gradiente de la función de costo y el gradiente de las restricciones.
- Solución de interés: minimizar la función de costo/maximizar las restricciones.

$$\frac{dL(\mathbf{x}, \boldsymbol{\lambda})}{d\mathbf{x}} = 0$$
$$\frac{dL(\mathbf{x}, \boldsymbol{\lambda})}{d\boldsymbol{\lambda}} = 0$$

Aprendizaje de máquina y optimización: resumiendo

- Ciencia de los datos: extraer información relevante desde datos (ejemplo: ajuste de datos).
- Optimización: herramienta matemática para ajustar algún modelo de "extracción de información" desde datos.
- Hipótesis \Rightarrow función de costo, restricciones.
- Función de costo: calidad de la solución.
- Restricciones: viabilidad de la solución.

Modelo lineal



- $y = f(x) = mx + b \Rightarrow$ noción lineal desde álgebra básica.
- $\mathbf{y} = f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + \mathbf{b} \Rightarrow$ extensión álgebra vectorial.
- Los datos no siempre llegan limpios y no siempre comparten relaciones lineales!

Modelo lineal: extensión matricial

$$\hat{\mathbf{y}} = f(\mathbf{X}) = \mathbf{X}\mathbf{w} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1P} \\ x_{21} & x_{22} & \dots & x_{2P} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \dots & x_{NP} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_P \end{bmatrix}$$

- $\hat{\mathbf{y}} \in \mathbb{R}^N$, $\mathbf{X} \in \mathbb{R}^{N \times P}$, $\mathbf{w} \in \mathbb{R}^P$
- N : # muestras.
- P : # características.

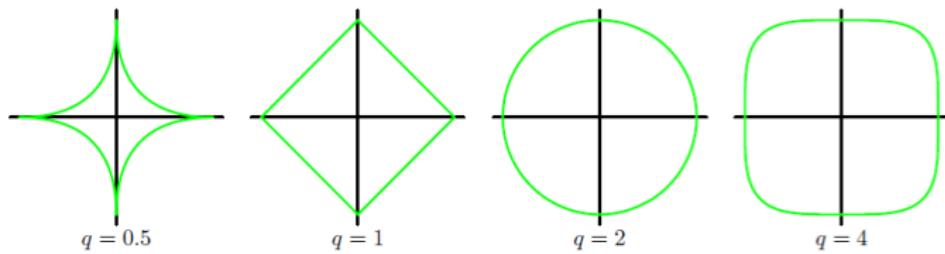
Cómo encontrar los parámetros del modelo lineal (\mathbf{w})?

Mínimos cuadrados como función de costo

$$\epsilon(\mathbf{w}) = \sum_{i=1}^N (y_i - \mathbf{x}_i \mathbf{w})^2$$

- La función de costo de mínimos cuadrados se entiende como el cálculo de la norma-2 del error de aproximación.
- $\mathbf{e} \in \mathbb{R}^N$: vector de error, donde $e_i = y_i - f(\mathbf{x}_i) = y_i - \mathbf{x}_i \mathbf{w}$.
- $\epsilon(\mathbf{w}) = \|\mathbf{e}\|_q = \left(\sum_{i=1}^N |e_i|^q \right)^{\frac{1}{q}}$.

Mínimos cuadrados como función de costo



- Mínimos cuadrados: $\|\mathbf{e}\|_q = \left(\sum_{i=1}^N |e_i|^q \right)^{\frac{1}{q}}$; para $q = 2$.
- Por facilidad matemática se trabaja con $\epsilon(\mathbf{w}) = \frac{1}{2} \|\mathbf{e}\|_2^2$.

Solución modelo lineal y mínimos cuadrados

$$\mathbf{w}^* = \arg \min_{\mathbf{w}} \epsilon(\mathbf{w})$$

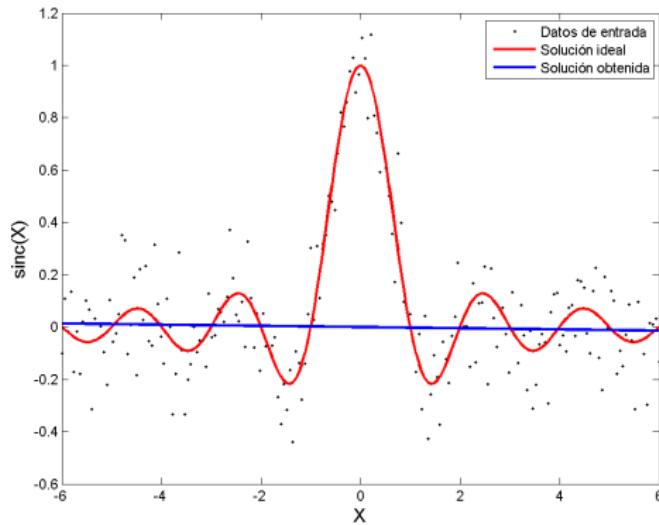
$$\mathbf{w}^* = \arg \min_{\mathbf{w}} \frac{1}{2} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2$$

$$\frac{d\epsilon(\mathbf{w})}{d\mathbf{w}} = \mathbf{y}^\top \mathbf{y} - 2\mathbf{y}^\top \mathbf{X}\mathbf{w} + \mathbf{w}^\top \mathbf{X}^\top \mathbf{X}\mathbf{w} = 0$$

$$\mathbf{w}^* = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

- $(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$: pseudoinversa de \mathbf{X} .
- $\mathbf{X}^\top \mathbf{X} \in \mathbb{R}^{P \times P}$: matriz que codifica las relaciones lineales entre las P características de \mathbf{X} .

Ejemplo modelo lineal



- Cómo tratamos con el ruido y relaciones no lineales?

La regularización del modelo lineal

- Se busca tratar con el mal condicionamiento de la matriz a invertir $(\mathbf{X}^\top \mathbf{X})^{-1}$.
- El mal condicionamiento se relaciona con el rango de la matriz - valores propios.
- Nociones básicas de factorización matricial - SVD y EIG.
- Se busca la solución más simple que se ajuste mejor a los datos.

La regularización y la solución de un sistema de ecuaciones

- $\mathbf{A}\mathbf{w} = \mathbf{b} \Rightarrow \mathbf{w} = \mathbf{A}^{-1}\mathbf{b}$: problema lineal clásico.
- En nuestro caso: $\mathbf{X}\mathbf{w} = \mathbf{y} \Rightarrow \mathbf{w} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$.
- $(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$ es la pseudoinversa de \mathbf{X} .
- Problema: la pseudoinversa no siempre es calculable/ no siempre es la mejor solución.
- Nuestros datos contienen ruido, falta información (rango de la matriz).
- Muchas variables - pocas ecuaciones linealmente independientes!

Repaso factorización en bases ortonormales

- La svd(\mathbf{X}) :

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^\top,$$

$\mathbf{U} \in \mathbb{R}^{N \times N}$, $\mathbf{S} \in \mathbb{R}^{N \times P}$, $\mathbf{V} \in \mathbb{R}^{P \times P}$.

- La eig(\mathbf{C}) :

$$\mathbf{C} = \mathbf{X}^\top \mathbf{X} = \mathbf{V}\Delta\mathbf{V}^\top,$$

$\mathbf{C} \in \mathbb{R}^{P \times P}$, $\Delta \in \mathbb{R}^{P \times P}$.

- La eig(\mathbf{K}) :

$$\mathbf{K} = \mathbf{X}\mathbf{X}^\top = \mathbf{U}\hat{\Delta}\mathbf{U}^\top,$$

$\mathbf{K} \in \mathbb{R}^{N \times N}$, $\hat{\Delta} \in \mathbb{R}^{N \times N}$.

- \mathbf{C} codifica relaciones entre características, \mathbf{K} codifica las relaciones entre las muestras.
- $\Delta = \hat{\Delta} = \mathbf{S}^2$, teniendo en cuenta el rango de la matriz.

El rango deficiente y el mal condicionamiento

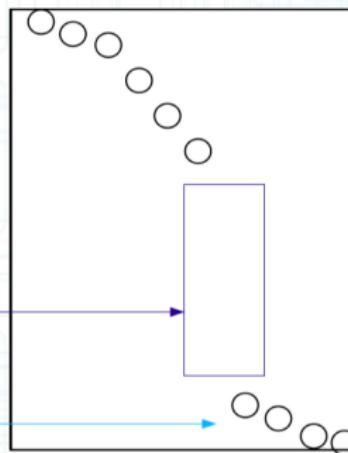
Rank Deficient

- Gap in the Singular Values



- Approx. Zero

- Noise!



Alistair Boyle, Feb 2009, SYS5906: Directed Studies -- Inverse Problems

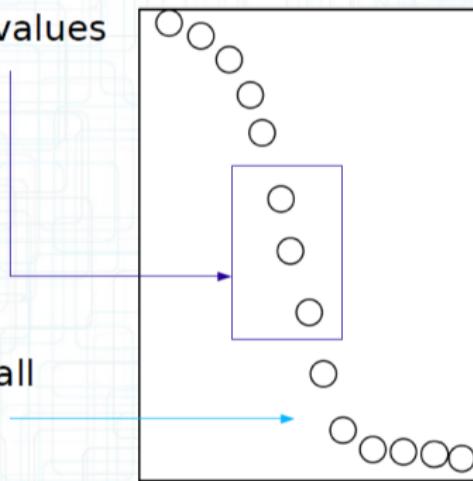
6

Los valores singulares (valores propios) me dan idea de que tanto ruido o que tanta información tengo en mis datos

El rango deficiente y el mal condicionamiento

Ill-Posed

- No gap in values



- Lots of small values

Alistair Boyle, Feb 2009, SYS5906: Directed Studies -- Inverse Problems

7

Los valores singulares (valores propios) me dan idea de que tanto ruido o que tanta información tengo en mis datos

Modelo lineal regularizado

$$\epsilon(\mathbf{w}, \lambda) = \frac{1}{2} \|\mathbf{e}\|_2^2 + \frac{1}{2} \lambda \|\mathbf{w}\|_2^2$$

- $\|\mathbf{e}\|_2^2$: cuantifica el desajuste entre los datos y el modelo lineal de aproximación.
- $\|\mathbf{w}\|_2^2$: cuantifica el sobreajuste (complejidad) de la solución.
- $\lambda \in \mathbb{R}^+$: parámetro de balance ("trade-off").
- Se necesita un λ que garantice una solución simple pero "exacta".

Solución del modelo lineal regularizado

$$\mathbf{w}^* = \arg \min_{\mathbf{w}} \frac{1}{2} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \frac{1}{2} \lambda \|\mathbf{w}\|_2^2$$

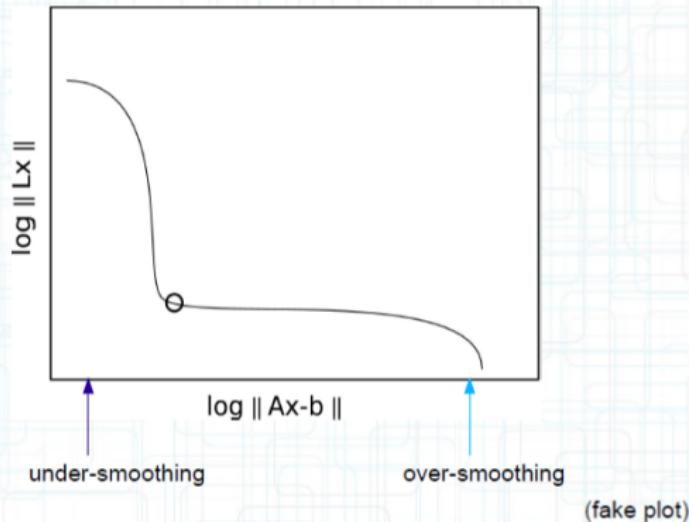
$$\frac{d\epsilon(\mathbf{w})}{d\mathbf{w}} = -\mathbf{X}^\top \mathbf{y} + \mathbf{X}^\top \mathbf{X}\mathbf{w} + \lambda \mathbf{w} = 0$$

$$\mathbf{w}^* = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}$$

- El término $\lambda \mathbf{I}$ ataca el mal condicionamiento de la matriz a invertir.
- Una solución alternativa es truncar la pseudoinversa (consultar que es TSVD).
- Sumarle un valor por la identica a la matriz "sube" los valores propios que están cercanos a cero.
- La "complejidad" de la solución se puede cuantificar con otras normas:
 - $\|\mathbf{w}\|_2, q = 2$: regularizador cuadrático.
 - $\|\mathbf{w}\|_1, q = 1$: lasso.

Selección de λ - L-curva

L-curve



Alistair Boyle, Feb 2009, SYS5906: Directed Studies -- Inverse Problems

Se busca el λ correspondiente al codo de la curva! ($L = \lambda I$)

Extensión a representaciones no lineales

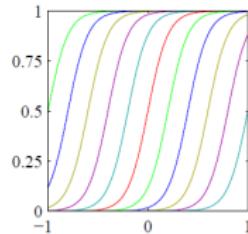
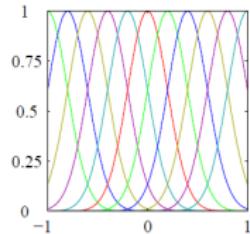
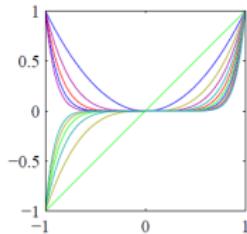
- La regularización nos ayuda a tratar con el ruido pero nuestro modelo sigue siendo lineal ("rectas").
- Una alternativa es extender las combinaciones lineales sobre funciones no lineales fijas de los datos de entrada:

$$\mathbf{y} = \Phi \mathbf{w},$$

donde $\Phi = [\phi^\top(\mathbf{x}_1), \phi^\top(\mathbf{x}_2), \dots, \phi^\top(\mathbf{x}_N)]^\top$, $\phi : \mathbb{R}^P \rightarrow \mathbb{R}^M$, $\mathbf{w} \in \mathbb{R}^Q$.

- El vector \mathbf{w} se estima sobre \mathbb{R}^Q utilizando mínimos cuadrados (regularizado) como función de costo.
- El modelo sigue siendo lineal pero aplicado sobre un nuevo espacio de representación Φ que puede codificar relaciones no lineales.

Ejemplos funciones base



- Polinomial: $\phi(\mathbf{x}) = [\mathbf{x}^j]_{j=1}^D$, D : grado del polinomio.
- Exponencial: $\phi\left(\mathbf{x} | \{\boldsymbol{\mu}_j\}_{j=1}^Q, \sigma\right) = \left[\exp\left(\frac{\|\mathbf{x} - \boldsymbol{\mu}_j\|_2^2}{2\sigma^2}\right)\right]_{j=1}^Q$.
- Sigmoidal:
$$\phi\left(\mathbf{x} | \{\boldsymbol{\mu}_j\}_{j=1}^Q, \sigma\right) = \left[1 / \left(1 + \exp\left(\|\mathbf{x} - \boldsymbol{\mu}_j\|_2^2 / (2\sigma^2)\right)\right)\right]_{j=1}^Q.$$

Solución del modelo lineal sobre representaciones no lineales

- Solución sobre \mathbf{X} :

$$\mathbf{w}^* = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}$$

$$y_i = \mathbf{x}_i \mathbf{w}$$

- Solución sobre Φ :

$$\mathbf{w}^* = (\Phi^\top \Phi + \lambda \mathbf{I})^{-1} \Phi^\top \mathbf{y}$$

$$\mathbf{X} \in \mathbb{R}^{N \times P}, \mathbf{w} \in \mathbb{R}^P, \mathbf{X}^\top \mathbf{X} \in \mathbb{R}^{P \times P}.$$

$$y_i = \phi(\mathbf{x}_i) \mathbf{w}$$

$$\Phi \in \mathbb{R}^{N \times Q}, \mathbf{w} \in \mathbb{R}^Q, \Phi^\top \Phi \in \mathbb{R}^{Q \times Q}.$$

Ejercicios Laboratorio

En un cuaderno (notebook) de jupyter responda a las siguientes preguntas con ejemplos concretos de implementación sobre Python 3. Envie su notebook al correo amalvarezme@unal.edu.co.

- Consultar el funcionamiento (modelo matemático, función de costo y optimización) de los siguientes algoritmos según su implementación en el paquete Scikit-Learn de Python:
 - `sklearn.linear_model.LinearRegression`
 - `sklearn.linear_model.Ridge`
 - `sklearn.linear_model.Lasso`
 - `sklearn.linear_model.ElasticNet`
 - `sklearn.kernel_ridge.KernelRidge` (describa los modelos no lineales disponibles).
- Utilizando las bases de datos estudiadas en la sesión 3, realice un análisis comparitivo de los métodos de regression mediante validación cruzada. Recuerde sintonizar los parámetros libres de cada algoritmo utilizando búsqueda por grilla y búsqueda aleatoria.

Referencias I



Hansen, P. C. (1998).

Rank-deficient and discrete ill-posed problems: numerical aspects of linear inversion.
(Vol. 4). Siam.



Bishop, C. (2006).

Pattern recognition.
Machine Learning, 128.