APRENENTATGE AUTOMÀTIC (APA)

Llista de Problemes 2

Lluís A. Belanche
Javier Béjar

Curso 2020-2021

Grau en Enginyeria Informàtica - UPC

Instruccions:

Per l'entrega de grup heu de triar un problema del capítol de problemes de grup.

Per l'entrega individual heu de triar un problema del capítol de problemes individuals. **Cada membre del grup ha de triar un problema diferent**.

Els problemes/apartats marcats amb **[PROG]** requereixen fer servir python per a ser resolts.

Heu de fer l'entrega pujant la solució en un arxiu **PDF**. Si escanegeu una solució feta a mà, comproveu que **sigui llegible**.

Objectius:

1. Saber plantejar problemes de mínims quadrats senzills i resoldre'ls per diferents mètodes
2. Saber reconèixer i plantejar sistemes d'equacions resolubles per regressió lineal regularitzada

Avaluació:

La nota d'aquesta entrega es calcularà com 1/3 de la nota del problema de grup mes 2/3 de la nota del problema individual.

CAPÍTOL 1

Problemes de Grup

1. **[PROG] Interacció entre partícules**

   S'ha dissenyat un experiment per provar una teoria sobre la naturalesa de la interacció entre certs tipus de partícules elementals en col·lisió amb protons. Es creu que la secció transversal està linealment relacionada amb la inversa de l'energia. A tal efecte, s'han determinat submostres per diferents nivells de la inèrcia de la partícula. En cada submostra es van prendre un gran nombre d'observacions i això ha permès estimar la desviació estàndard (sd) de la secció transversal (st) mesurada, com indica la següent taula:

energia	st	sd
2.899	367	17
3.484	311	9
3.984	295	9
4.444	268	7
4.831	253	7
5.376	239	6
6.211	220	6
7.576	213	6
11.905	193	5
16.667	192	5

   Plantegeu el problema de predir la secció transversal amb la inversa de l'energia com una regressió lineal ponderada. Resoleu-lo numèricament usant la rutina `LinearRegression()`. Feu un gràfic del resultat amb la ponderació i sense; compareu els resultats i expliqueu la raó de les diferències.

2. **[PROG] Càlcul d'òrbites**

   El cometa Tentax es va descobrir al 1968 i té una òrbita quadràtica (el·líptica, parabòlica o hiperbòlica) d'acord a les lleis de Kepler. L'òrbita té l'equació:

$$R = \frac{p}{1 - e\cos\phi}$$

on p és un coeficient específic per aquest cometa, e és l'excentricitat (totes dues desconegudes) i parelles (r, ϕ) indiquen les diferents posicions observades (en coordenades polars amb centre en el Sol).

Els astrònoms han reunit un conjunt de coordenades:

$$\{(2.70, 48°), (2.00, 67°), (1.61, 83°), (1.20, 108°), (1.02, 126°)\}$$

(a) Escriviu el problema com un sistema lineal

(b) Trobeu les dues constants p y e per mínims quadrats i feu un gràfic amb les dades i la solució obtinguda.

3. **[PROG] Robust Regression**

One problem with using the squared error loss for linear regression is that outliers will have a large effect on the result, and depending on their number, it can result in a very bad model. To reduce this problem loss functions less sensitive to outliers can be used, this is known as robust regression. That is the purpose, for instance, of the *Huber Loss*, that minimizes the squared error when the error is smaller than an ϵ and the absolute error otherwise. This method is implemented in scikit-learn by the function `HuberRegressor`. The tasks of this exercise are:

(a) Create a random one variable regression using the function `make_regression`, around 250 samples will be enough.

(b) Split the data in train and test sets (80%/20%) and modify some of the examples of the training set so they get far from the rest. For instance you can add a constant to the independent and dependent variables and some gaussian noise. Check that the examples are far enough from the regression.

(c) Compute the Hubber regression with different values for ϵ. Compute the linear regression and the Ridge regression (use `RidgeCV` to obtain the best regularization parameter)

(d) Plot the different regressions and compute their mean squared error using the train and test set and comment the results.

CAPÍTOL 2

Problemes Individuals

1. **Ridge Regression 1**

   En la regressió ridge amb funcions polinòmiques de grau M en una variable, l'error empíric regularitzat a minimitzar se sol expressar:

   $$E_\lambda(c) = \frac{1}{2}\sum_{i=1}^{n}(y_i - f(x_i;c))^2 + \frac{\lambda}{2}c^T c, \quad x_i, y_i \in \mathbb{R}, c \in \mathbb{R}^{M+1}$$

   on $f(x;c) = \sum_{j=0}^{M} c_j x^j$. Recordeu que $c^T c = ||c||^2$. Usualment el terme c_0 no es regularitza, ja que correspon a l'estimació de la mitjana de les y_i ; per simplicitat, ignorem aquest aspecte en el problema. Suposem que hem fixat $\lambda > 0$. Es demana:

   (a) Expliqueu en què consisteixen els dos termes de l'error i el paper que juga el paràmetre λ.

   (b) Expliqueu perquè el model f és un model lineal

   (c) Doneu les equacions del sistema que caldria resoldre per trobar el vector de paràmetres c (NO cal que el resolgueu):

      i. Calculeu $\frac{\partial E_\lambda}{\partial C}$. Com que tots els c_k juguen el mateix paper a f, podeu calcular per un $\frac{\partial E_\lambda}{\partial c_k}$ per un $k = 0, \ldots, M$ arbitrari.

      ii. Igualeu a 0 les derivades anteriors i obtindreu un sistema d'equacions on les c_k són les incògnites. Quantes equacions i incògnites tenim? Manipuleu el sistema de manera que resulti obvi que és un sistema *lineal* d'equacions. Acabeu expressant el sistema com:

      $$\sum_{i=0}^{M} A_{ij} C_i = B_i, j = 0, \ldots, M$$

      Cal que doneu les expressions per A_{ij} i B_j.

2. **Ridge Regression 2**

   En la regressió ridge amb funcions lineals en d variables, l'error empíric regularitzat a minimitzar se sol expressar:

$$E_\lambda(c) = \frac{1}{2}\sum_{i=1}^{n}(y_i - f(x_i;c))^2 + \frac{\lambda}{2}c^T c, \quad x_i, y_i \in \mathbb{R}, c \in \mathbb{R}^{M+1}$$

on $f(x;c) = \sum_{j=0}^{M} c_j x_j + c_0$. Recordeu que $c^T c = ||c||^2$. Usualment el terme c_0 no es regularitza, doncs correspon a l'estimació de la mitjana de les y_i; per simplicitat, ignorem aquest aspecte en el problema. Suposem que hem fixat $\lambda > 0$. Es demana:

(a) Expliqueu en què consisteixen els dos termes de l'error i el paper que juga el paràmetre λ.

(b) Expliqueu perquè el model f és un model lineal

(c) Doneu les equacions del sistema que caldria resoldre per trobar el vector de paràmetres c (NO cal que el resolgueu):

i. Calculeu $\frac{\partial E_\lambda}{\partial C}$. Com que tots els c_k juguen el mateix paper a f, podeu calcular per un $\frac{\partial E_\lambda}{\partial c_k}$ per un $k = 0, \dots, M$ arbitrari.

ii. Igualeu a 0 les derivades anteriors i obtindreu un sistema d'equacions on les c_k són les incògnites. Quantes equacions i incògnites tenim? Manipuleu el sistema de manera que resulti obvi que és un sistema lineal d'equacions. Acabeu expressant el sistema com:

$$\sum_{i=0}^{M} A_{ij} C_i = B_i, j = 0, \dots, M$$

Cal que doneu les expressions per A_{ij} i B_j.

3. **Regressió lineal ponderada**

Quan hem parlat de regressió lineal, normalment hem suposat que el soroll Gaussià és homocedàstic. Això ens ha portat a obtenir que la maximització de la funció log-versemblança és equivalent a la minimització de l'error quadràtic. Ara suposem que les respectives variàncies de la part estocàstica $\epsilon_1, \dots, \epsilon_n$ són diferents (i independents entre si), és a dir, $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ (es diu heterocedasticitat).

(a) Escriviu la nova funció log-versemblança (negativa) pel vector de paràmetres c de la regressió.

(b) Mostreu que la minimització d'aquesta funció log-versemblança negativa és equivalent a la minimització d'un error quadràtic, només que ara és ponderat:

$$E(c) = \frac{1}{2}\sum_{i=1}^{n} a_i (y_i - f(x_i;c))^2$$

Expresseu a_i en funció de σ_i^2 i interpreteu el resultat.

(c) Deriveu una expressió per l'estimador de σ_i^2 i interpreteu-la.

4. **[PROG] Propietats elàstiques d'una molla**

Volem determinar les propietats elàstiques d'una molla usant diferents pesos i mesurant la deformació que es produeix. La llei de Hooke relaciona la longitud l i la força F que exerceix el pes com:

$$e + kF = l$$

on e, k són constants de la llei, que es volen determinar. S'ha realitzat un experiment i obtingut les dades:

F	1	2	3	4	5
l	7.97	10.2	14.2	16.0	21.2

(a) Plantegeu el problema com un problema de mínims quadrats

(b) Resoleu-lo amb el mètode de la matriu pseudo-inversa

(c) Resoleu-lo amb el mètode basat en la SVD

5. **[PROG] Ajustant polinomis**

En un problema de regressió univariant es tenen les parelles d'exemples:

$$\{(-1,2),(1,1),(2,1),(3,0),(5,3),(12,6),(15,10),(21,0)\}$$

Es vol ajustar un polinomi de la forma $f(x) = c_0 + \sum_{j=1}^{M} c_j x^j$, per $M = 2, 3, 4$.

(a) Plantegeu el problema com un problema de mínims quadrats

(b) Resoleu-lo amb el mètode de la matriu pseudo-inversa

(c) Feu un gràfic amb les dades i les solucions obtingudes. Expliqueu com es podria fer per triar la millor solució.

6. **[PROG] Producció anual de minerals**

Una empresa extreia un mineral preciós i portava un registre anual de la massa extreta (en tones mètriques, equivalents a 1.000 quilos). Per la dècada dels 70 es va obtenir la producció de la següent taula:

Any	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979
Tones	9	9	10	10	8	6	6	4	6	4

(a) Plantegeu i resoleu numèricament el problema de predir la producció en funció de l'any usant la rutina `LinearRegression()`. Feu un gràfic amb les dades i la solució obtinguda.

(b) Si no hagués cap influència externa que provoqués variacions substancials en la producció, quina seria la previsió de producció per 1984? Doneu-ne un interval de confiança al 95%.

(c) Critiqueu el model i proposeu-ne un de millor (en el sentit de més realista).

7. **[PROG] Viatjant pels EE.UU.**

La següent taula mostra les distàncies (en milles) entre Baltimore i altres 12 ciutats dels EE.UU., juntament amb el preu del bitllet d'avió (en dòlars) entre elles.

Destí	Distància	Tarifa
Atlanta	576	178
Boston	370	138
Chicago	612	94
Dallas	1216	278
Detroit	409	158
Denver	1502	258
Miami	946	198
New Orleans	998	188
New York	189	98
Orlando	787	179
Pittsburgh	210	138
St. Louis	737	98

(a) Plantegeu i resoleu numèricament el problema de predir la Tarifa amb la Distància usant la rutina `LinearRegression()`. Feu un gràfic amb les dades i la solució obtinguda

(b) Observareu que algunes ciutats tenen tarifes anormalment baixes per la distància a la qual es troben. Dissenyeu una manera de reduir la influència d'aquests casos i recalculeu la solució.

8. **Interpolació polinòmica**

El problema d'interpolar un conjunt de punts $\{(x_1, y_1), \dots, (x_n, y_n)\}$ amb una funció f consisteix a forçar que $f(x_i) = y_i$, per tot i. Assumint que tots els xi són diferents entre si, aquesta tasca és resoluble amb un polinomi de grau $n - 1$. Es demana:

(a) Definiu el vector $y = (y_1, \cdot, y_n)^\top$ i la matriu de disseny Φ convenientment, expresseu el problema en format matricial i ressoleu-lo. Pista: la matriu Φ que en resulta es coneix com matriu de *Vandermonde*.

(b) Apliqueu el resultat a dades de la vostra elecció amb $n = 10$ i comproveu la qualitat de la solució.

9. **[PROG] Ridge Regression and Singular Value Decomposition**

We have seen that Linear Regression can be computed using singular value decomposition. The weights of Ridge Regression can be computed as:

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

(a) Show that using singular value decomposition the weights can also be computed as:

$$w = V(\Lambda^2 + \lambda I)^{-1} \Lambda U^\top y$$

Remenber that the U and V matrices from the SVD are orthonormal, so $A^\top A = AA^\top = A^{-1}A = AA^{-1} = I$ and the Λ matrix is diagonal.

(b) Generate a random regression problem using the function `make_regression` and solve it using the two expressions.