APA: Aprenentatge Automàtic (TEMA 4)
Grau en Enginyeria Informàtica - UPC (2017/18)

Lluís A. Belanche, `belanche@cs.upc.edu`

Entrega: 13 Novembre 2017

Els problemes marcats **[G]** són de grup; els problemes/apartats marcats **[R]** són per fer-se en `R`

Objectius:

1. Conèixer el compromís biaix/variança i la descomposició del risc total en el cas de la regressió
2. Saber plantejar problemes de mínims quadrats senzills i resoldre'ls per diferents métodes

Problema 1 Descomposició del risc en regressió [G]

Per tasques de regressió plantejades de la forma usual $t = f(\boldsymbol{x}) + \epsilon$, on $\mathbb{E}[\epsilon] = 0$ i $Var[\epsilon] = \sigma^2 < \infty$; pel cas on $\epsilon \sim N(0, \sigma^2)$, que porta a l'error quadràtic, el funcional de **risc total** d'un model y és:

$$R(y) = \int_{\mathbb{R}} \int_{\mathbb{R}^d} (t - y(\boldsymbol{x}))^2 p(t, \boldsymbol{x})\, d\boldsymbol{x}\, dt,$$

que cal minimitzar respecte y. Sumant i restant $f(\boldsymbol{x})$ dins del quadrat, $R(y)$ es pot descomposar en la suma de tres termes, que hem anomenat A, B i C, on:

- $A = \int_{\mathbb{R}} \int_{\mathbb{R}^d} (t - f(\boldsymbol{x}))^2 p(t, \boldsymbol{x})\, d\boldsymbol{x}\, dt$
- $B = \int_{\mathbb{R}^d} (f(\boldsymbol{x}) - y(\boldsymbol{x}))^2 p(\boldsymbol{x})\, d\boldsymbol{x}$
- $C = 2 \int_{\mathbb{R}} \int_{\mathbb{R}^d} (t - f(\boldsymbol{x}))(f(\boldsymbol{x}) - y(\boldsymbol{x})) p(t, \boldsymbol{x})\, d\boldsymbol{x}\, dt$

1. Demostreu que $A = \sigma^2$. Pista: cal un canvi de variable per introduir ϵ; després, si us cal, podeu usar que per $a > 0$, $\int_0^\infty u^2 e^{-a^2 u^2}\, du = \frac{1}{4}\sqrt{\pi} a^{-3}$.
2. Demostreu que $C = 0$.
3. Interpreteu el resultat final en termes de A, B. Useu els conceptes de biaix i variança.

. . . . . .

Problema 2 Regressió lineal ponderada

En regressió lineal, sota la suposició que el soroll gaussià és homocedàstic, la maximització de la funció log-versemblança és equivalent a la minimització de l'error quadràtic. En el cas d'heterocedasticitat, fem $t_n = f(\boldsymbol{x}_n) + \epsilon_n$, on $\epsilon_n \sim N(0, \sigma_n^2)$. Llavors la maximització de la funció log-versemblança és equivalent a la minimització de l'error quadràtic *ponderat* (vegeu el problema 7 del TEMA 1):

$$E(\boldsymbol{w}) = \frac{1}{2} \sum_{n=1}^{N} a_n (t_n - y(\boldsymbol{x}_n; \boldsymbol{w}))^2$$

on $a_n = \sigma_n^{-2}$, de manera que l'optimització dona menys importància a aquells exemples amb més variabilitat condicional (i, per tant, menys representatius de la mitjana). Volem ara resoldre aquest tipus de problemes d'una forma més general, de la següent manera. Suposem que partim del problema de mínims quadrats:

$$\min_{\boldsymbol{w}} E(\boldsymbol{w}) = (\boldsymbol{t} - \Phi\boldsymbol{w})^{\mathrm{T}} A (\boldsymbol{t} - \Phi\boldsymbol{w})$$

on Φ és la matriu de disseny i A és una matriu simètrica i semi-definida positiva (PSD).

1. Determineu una matriu A tal que s'obtingui el resultat del problema 10 del TEMA 1.
2. Resoleu (en $\boldsymbol{w}$) el problema general de mínims quadrats per matrius A PSD arbitràries.

. . . . . .

Problema 3 Interacció entre partícules [R,G]

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 estàndar (`sd`) de la secció transversal (`st`) mesurada, com indica la Taula 1.

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

Taula 1: Interacció entre partícules

Plantegeu el problema de predir la secció transversal amb la inversa de l'energia com una regressió lineal ponderada (vegeu el problema 7 del TEMA 1). Resoleu-lo numèricament usant la rutina `lm()` (ó `glm()` si especifiqueu `family=gaussian`). Feu un gràfic del resultat amb i sense la ponderació; compareu els resultat i expliqueu la raó de les diferències.

. . . . . .

Problema 4 Propietats elàstiques d'una molla [R]

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

1. Plantegeu el problema com un problema de mínims quadrats
2. Resoleu-lo amb el métode de la matriu pseudo-inversa
3. Resoleu-lo amb el métode basat en la SVD

. . . . . .

Problema 5 Ajustant un petit polinomi [R]

En un problema de regressió univariant es tenen les parelles d'exemples $\{(-1,2),(1,1),(2,1),(3,0),(5,3)\}$. Es vol ajustar un polinomi de grau dos de la forma $y(x) = c_0 + c_1 x + c_2 x^2$.

1. Plantegeu el problema com un problema de mínims quadrats
2. Resoleu-lo amb el métode de la matriu pseudo-inversa
3. Feu un gràfic amb les dades i la solució obtinguda

. . . . . .

Problema 6 Càlcul d'òrbites [R,G]

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 les parelles (r,ϕ) indiquen les diferents posicions observades (en coordenades polars amb centre en el Sol).

Els astrònoms han reunit un conjunt de coordenades:

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

1. Escriviu el problema com un sistema lineal
2. Trobeu les dues constants p, e per mínims quadrats

. . . . . .

Problema 7 Equivalència de solucions

La solució del problema $\min_{\boldsymbol{x}} ||\boldsymbol{y} - A\boldsymbol{x}||^2 + \lambda \boldsymbol{x}^{\mathrm{T}} C \boldsymbol{x}$ és $\boldsymbol{x}^* = \left(A^{\mathrm{T}}A + \lambda C\right)^{-1} A^{\mathrm{T}}\boldsymbol{y}$, on C és una matriu semi-definida positiva. Utilitzeu aquest resultat per demostrar formalment que la solució del problema $\min_{\boldsymbol{x}} ||\boldsymbol{y} - A\boldsymbol{x}||^2 + ||B\boldsymbol{x}||^2$ és $\boldsymbol{x}^* = \left(A^{\mathrm{T}}A + B^{\mathrm{T}}B\right)^{-1} A^{\mathrm{T}}\boldsymbol{y}$.

. . . . . .

Problema 8 Producció anual de minerals [R]

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 Taula 2.

1. Plantegeu i resoleu numèricament el problema de predir la producció en funció de l'any usant la rutina `lm()` (ó `glm()` si especifiqueu `family=gaussian`). Feu un gràfic amb les dades i la solució obtinguda
2. Si no hi ha hagúes cap influència externa forta que provoqués variacions substancials en la producció, quina seria la previssió de producció per 1984? Doneu un interval de confiança al 95%.

. . . . . .

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

Taula 2: Producció de minerals

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

Taula 3: Viatjant pels EEUU

Problema 9 Viatjant pels EEUU [R]

La Taula 3 mostra les distàncies (en milles) entre Baltimore i altres 12 ciutats dels EEUU, juntament amb el preu del bitllet d'avió (en dòlars) entre elles.

1. Plantegeu i resoleu numèricament el problema de predir la `Tarifa` amb la `Distància` usant la rutina `lm()` (ó `glm()` si especifiqueu `family=gaussian`). Feu un gràfic amb les dades i la solució obtinguda
2. Observareu que algunes ciutats tenen tarifes anormalment baixes per la distància a la que es troben. Dissenyeu una manera de reduir la influència d'aquests casos i recalculeu la solució.

. . . . . .

Problema 10 Interpolació polinòmica

El problema d'*interpolar* un conjunt de punts $\{(\boldsymbol{x}_1, t_1), \ldots, (\boldsymbol{x}_N, t_N)\}$ amb una funció f consisteix a forçar que $f(\boldsymbol{x}_n) = t_n$, per tot n. Assumint que tots els $\boldsymbol{x}_n$ són diferents entre sí, aquesta tasca és resoluble amb un polinomi de grau $N-1$. Es demana:

1. Definiu el vector $\boldsymbol{t} = (t_1, \ldots, t_n)^{\mathrm{T}}$ 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*.
2. Apliqueu el resultat a dades de la vostra elecció amb $N = 10$ i comproveu la qualitat de la solució.

. . . . . .