APA: Aprenentatge Automàtic (TEMA 1)
Grau en Enginyeria Informàtica - UPC (2018/19)

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

Entrega: 2 Octubre 2018

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

Objectius:

1. Saber reconèixer i plantejar sistemes d'equacions resolubles per regressió lineal regularitzada;
2. Saber calcular estimadors de màxima versemblança per distribucions de probabilitat senzilles i conèixer les seves propietats fonamentals;
3. Saber relacionar les tècniques de màxima versemblança amb la regressió lineal regularitzada.

Problema 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:

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

on $f(x;\boldsymbol{c}) = \sum_{j=0}^{M} c_j x^j$. Recordeu que $\boldsymbol{c}^T\boldsymbol{c} = \|\boldsymbol{c}\|^2$. 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:

1. Expliqueu en què consisteixen els dos termes de l'error i el paper que juga el paràmetre λ.
2. Expliqueu perquè el model f és un model **lineal**.
3. Doneu les equacions del sistema que caldria resoldre per trobar el vector de paràmetres $\boldsymbol{c}$ (NO cal que el resolgueu):
   (a) Calculeu $\frac{\partial E_\lambda}{\partial \boldsymbol{c}}$. Com que tots els c_j juguen el mateix paper a f, podeu calcular $\frac{\partial E_\lambda}{\partial c_j}$ per un $j = 0, \ldots, M$ arbitrari.
   (b) Igualeu a 0 les derivades anteriors i obtindreu un sistema d'equacions on les c_j són les incògnites. Quantes equacions i incògnites tenim? Manipuleu el sistema de manera que resulti obvi que és un sistema *lineal* d'equacions. Suggeriment: expresseu el sistema com:

$$\sum_{i=0}^{M} A_{ij}c_i = B_j, j = 0, \ldots, M$$

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

. . . . . .

Problema 2 Ridge regression 2

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

$$E_\lambda(\boldsymbol{c}) = \frac{1}{2}\sum_{i=1}^{n}(y_i - f(\boldsymbol{x}_i;\boldsymbol{c}))^2 + \frac{\lambda}{2}\boldsymbol{c}^T\boldsymbol{c}, \qquad \boldsymbol{x}_i \in \mathbb{R}^d, y_i \in \mathbb{R}, \boldsymbol{c} \in \mathbb{R}^{d+1}$$

on $f(\boldsymbol{x};\boldsymbol{c}) = \sum_{j=1}^{d} c_j x_j + c_0$. Recordeu que $\boldsymbol{c}^T\boldsymbol{c} = \|\boldsymbol{c}\|^2$. 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:

1. Expliqueu en què consisteixen els dos termes de l'error i el paper que juga el paràmetre λ.
2. Expliqueu perquè el model f és un model lineal.
3. Doneu les equacions del sistema que caldria resoldre per trobar el vector de paràmetres $\boldsymbol{c}$ (NO cal que el resolgueu):
   (a) Calculeu $\frac{\partial E_\lambda}{\partial \boldsymbol{c}}$. Com que tots els c_j juguen el mateix paper a f, podeu calcular $\frac{\partial E_\lambda}{\partial c_j}$ per un $j = 0, \ldots, M$ arbitrari.
   (b) Igualeu a 0 les derivades anteriors i obtindreu un sistema d'equacions on les c_j són les incògnites. Quantes equacions i incògnites tenim? Manipuleu el sistema de manera que resulti obvi que és un sistema *lineal* d'equacions. Suggeriment: el mateix que pel Problema 1 (3b).

. . . . . .

Problema 3 Màxima versemblança 1 [G]

Considerem un experiment aleatori en què mesurem una determinada variable aleatòria X, que segueix una distribució Gaussiana univariada, cosa que escrivim $X \sim \mathcal{N}(\mu, \sigma^2)$. Prenem n mesures independents de X i obtenim una mostra aleatòria simple $\{x_1, \ldots, x_n\}$, on cada x_i és una realització de X, per $i = 1, \ldots, n$. Es demana:

1. Escriviu la funció de densitat de probabilitat per un x_i qualsevol i construïu la funció log-versemblança (negativa) de la mostra.
2. Trobeu els estimadors de màxima versemblança $\hat{\mu}$ i $\hat{\sigma}^2$ per μ i per σ^2, a partir de la mostra.
3. Demostreu que realment són mínims (i no extrems qualsevol).
4. Calculeu els biaixos dels dos estimadors. Determineu si l'estimador per μ és consistent.
5. Calculeu la variança de l'estimador per μ, de 2 maneres (que han de coincidir), a triar de les 3 següents. Pista: useu que si $X_i, X_j \sim \mathcal{N}(\mu, \sigma^2)$, llavors:

$$\mathbb{E}[X_i \cdot X_j] = \begin{cases} \mu^2 & \text{si } i \neq j; \\ \mu^2 + \sigma^2 & \text{si } i = j. \end{cases}$$

   (a) Insertant directament el valor de l'estimador i utilitzant les propietats de la variança;
   (b) Usant la coneguda fòrmula $Var[\hat{\theta}] = \mathbb{E}[\hat{\theta}^2] - (\mathbb{E}[\hat{\theta}])^2$;
   (c) Utilitzant la definició directa de la variança $Var[\hat{\theta}] = \mathbb{E}[(\mathbb{E}[\hat{\theta}] - \hat{\theta})^2]$.

6. Per determinar la velocitat màxima d'un prototip d'avió es van fer 15 proves independents (i caríssimes!) amb els resultats

   $$422.2, 418.7, 425.6, 420.3, 425.8, 423.1, 431.5, 428.2, 438.3, 434.0, 411.3, 417.2, 413.5, 441.3, 423.0$$

   on els valors venen expressats en m/s. Suposant que aquesta velocitat màxima és Gaussiana, quins valors estimaríeu per μ i per σ?

7. **[R]** Fixeu μ i σ^2 al vostre gust i genereu 1000 mostres i.i.d. de mida $n = 50$ (noteu que no cal emmagatzemar-les); calculeu els valors teòrics del biaix i la variança. Utilitzeu les mostres (que han de ser independents entre si) per *estimar* els biaixos i variances i compareu els dos resultats.

. . . . . .

Problema 4 Màxima versemblança 2 [G]

Una seqüència d'ADN està formada per mitjà de quatre bases: adenina (A), citosina (C), guanina (G) i timina (T). La seqüència d'aquestes quatre bases al llarg de la cadena codifica la informació. Estem interessats en l'estimació de la probabilitat de cada base que apareix en una seqüència d'ADN. Tractem cada lloc d'ADN com una variable aleatòria seguint una distribució categòrica de 4 valors; llavors les seqüències es formen per mostreig repetitiu (independent) d'aquesta distribució n cops. La distribució té k paràmetres, que denotem $p_1, \ldots, p_k$, tals que $p_i > 0$ i $\sum_{i=1}^{k} p_i = 1$ (en el cas que ens ocupa, $k = 4$; es redueix a la Bernoulli quan $k = 2$). Tenim una mostra $x_1, \ldots, x_n$, on $x_i \in \{A, C, G, T\}$.
Pista: consulteu el llibre d'en Bishop, pàgina 75.

1. Construïu la funció log-versemblança de la mostra.
2. Trobeu estimadors de màxima versemblança pels p_i, $i = 1, \ldots, 4$, a partir de la mostra i demostreu que realment són màxims (i no extrems qualsevols).
3. Calculeu els seus biaixos i variances. Concluïu si són biaixats (o no) i consistents (o no).
4. Observant un tros de seqüència d'ADN de mida 20, hem mesurat les bases GCGACGTAGTGTTGAGAACT. Quins valor estimaríeu pels p_i?

. . . . . .

Problema 5 Màxima versemblança 3 [G]

La distribució de Poisson és una distribució discreta sobre comptatges positius. Es pot aplicar a sistemes amb un gran nombre de possibles esdeveniments, cadascun dels quals és poc freqüent (per exemple, el nombre de cotxes que passen un peatge en una hora). Té com a funció de probabilitat:

$$p(x; \lambda) = e^{-\lambda} \frac{\lambda^x}{x!}, \qquad x = 0, 1, 2, \ldots$$

on x és el valor i $\lambda > 0$ el paràmetre de la Poisson. Considerem un experiment aleatori en què mesurem una determinada variable aleatòria X, que segueix una distribució de Poisson, cosa que escrivim $X \sim Pois(\lambda)$. Prenem n mesures independents de X i obtenim una mostra aleatòria simple $\{x_1, \ldots, x_n\}$, on cada x_i és una realització de X, per $i = 1, \ldots, n$. Es demana:

1. Construïu la funció log-versemblança (negativa) de la mostra.
2. Trobeu l'estimador de màxima versemblança per λ a partir de la mostra i demostreu que realment és un mínim (i no un extrem qualsevol).
3. Calculeu el seu biaix i la seva variança. Concluïu si aquest estimador és biaixat (o no) i consistent (o no). En cas que sigui biaixat, proposeu-ne un que corregeixi el biaix.

4. **[R]** Observant un mateix peatge d'autopista durant 24 dies consecutius a la mateixa hora hem realitzat els comptatges horaris

$$8, 17, 14, 21, 11, 14, 14, 11, 12, 11, 6, 10, 17, 12, 12, 22, 14, 10, 13, 11, 8, 10, 9, 13$$

Quin valor estimaríeu per λ? Com trobaríeu una estimació per la probabilitat que no passi cap cotxe en 1 hora?

5. L'experiment Irvine-Michigan-Brookhaven va permetre al 1989 la mesura del nombre de neutrins provinents de la supernova S1987, detectats en intervals de 10 segons:

nombre de neutrins	0	1	2	3	4	5	6	7	8
nombre d'intervals	1042	860	307	78	15	3	0	0	0

La hipòtesi més freqüent és suposar que el nombre de neutrins segueix una distribució de Poisson. Doneu una estimació del seu paràmetre λ.

. . . . . .

Problema 6 Màxima versemblança 4

La distribució de Bernoulli és una distribució binària sobre el fet que un esdeveniment tingui èxit (amb una certa probabilitat p) o fracassi (amb probabilitat $1-p$); la funció de distribució es:

$$P(X=x) = \begin{cases} p & \text{si } x=1; \\ 1-p & \text{si } x=0. \end{cases}, \ p \in (0,1)$$

on x és el valor i p el paràmetre de la Bernoulli. Considerem un experiment aleatori en què mesurem una determinada variable aleatòria X, que segueix una distribució de Bernoulli, cosa que escrivim $X \sim Ber(p)$. Prenem n mesures independents de X i obtenim una mostra aleatòria simple $\{x_1, \ldots, x_n\}$, on cada x_i és una realització de X, per $i = 1, \ldots, n$. Noteu que la variable aleatòria Y definida com "número d'èxits en la mostra" segueix la coneguda distribució binomial $Y \sim B(n,p)$. Es demana:

1. Construïu la funció log-versemblança de la mostra.
2. Trobeu l'estimador de màxima versemblança per p a partir de la mostra i demostreu que realment és un màxim (i no un extrem qualsevol).
3. Calculeu el seu biaix i la seva variança. Concluïu si aquest estimador és biaixat (o no) i consistent (o no). En cas que sigui biaixat, proposeu-ne un que corregeixi el biaix.
4. Un amic (de poc fiar) ens mostra una moneda (amb cara i creu) i la llença 100 cops davant nostre. Surten 73 cares i 27 creus. Quin valor estimaríeu per p, la probabilitat de cara?

. . . . . .

Problema 7 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 variances de la part estocàstica $\epsilon_1, \ldots, \epsilon_n$ són diferents (i independents entre si), és a dir, $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ (es diu *heterocedasticitat*).

1. Escriviu la nova funció log-versemblança (negativa) pel vector de paràmetres $\boldsymbol{c}$ de la regressió.

2. Mostreu que la minimització d'aquesta funció log-versemblança negativa és equivalent a la minimització d'un error quadràtic, només que ara és *ponderat*:

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

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

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

. . . . . .

Problema 8 Mínims quadrats

Tenim una seqüència de números reals $x_1, \ldots, x_n$. Definim la funció

$$f(x) = \frac{1}{n}\sum_{i=1}^{n}(x - x_i)^2$$

1. Demostreu que f té un únic mínim, doneu-lo i interpreteu el resultat.

2. Considerem ara la nova funció

$$f(x) = \sum_{i=1}^{n} p_i(x - x_i)^2$$

   on ponderem cada terme per un factor $p_i > 0$, de manera que $\sum_{n=1}^{N} p_n = 1$. Recalculeu la solució.

3. **[R]** Sigui $n = 100$. Apliqueu el resultat precedent a una seqüència de nombres triada per vosaltres. Useu una ponderació basada en nombres independents uniformes en $(0, 1)$. Dibuixeu la funció i el mínim.

. . . . . .

Problema 9 Error quadràtic mitjà

L'error quadràtic mitjà (MSE) d'un estimador $\hat{\theta}$ d'un paràmetre θ es defineix com:

$$\text{MSE}(\hat{\theta}) := \mathbb{E}[(\hat{\theta} - \theta)^2]$$

on se sobreentén que totes les esperances són sobre mostres de mida N. Es demana demostrar l'important resultat $\text{MSE}(\hat{\theta}) = bias(\hat{\theta})^2 + Var(\hat{\theta})$.

1. Introduïu el terme $-\mathbb{E}[\hat{\theta}]+\mathbb{E}[\hat{\theta}]$ i desplegueu el quadrat; surten tres esperances a les que anomenarem A, B i C, essent C el producte creuat.

2. Demostreu que A i B corresponen a $bias(\hat{\theta})^2$ i $Var(\hat{\theta})$.

3. Demostreu que C és zero.

. . . . . .

Problema 10 L'ajust de regressió per màxima versemblança

Considerem un experiment en què mesurem dues variables aleatòries X i T, totes dues contínues, on sabem que el valor de X determina (en part) el de T. Volem predir T a partir de X amb una certa classe de funcions que depenen d'un vector de coeficients $\boldsymbol{c}$. Assumim la relació $t = f(x; \boldsymbol{c}) + \epsilon$, on $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

Prenem n mesures independents de parelles (X, T) i obtenim una mostra aleatòria simple que escrivim $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, de manera que $P(y|x, \boldsymbol{c}, \sigma^2) = \mathcal{N}(y; f(x; \boldsymbol{c}), \sigma^2)$. Es demana:

1. Construïu la funció log-versemblança (negativa) de la mostra, prenent com a paràmetres $\boldsymbol{c}$ i σ^2.
2. Argumenteu quina és la funció d'error (que caldria minimitzar) per tal d'obtenir estimacions dels coeficients $\boldsymbol{c}$.
3. Obteniu una estimació del paràmetre σ^2 i doneu-ne una interpretació.
4. Considereu ara un nou punt x^* (que no és cap dels x_i). Expresseu la predicció y^* per aquest punt. Expliqueu com es podria fer per donar també un interval de confiança al voltant de la predicció y^*.

. . . . . .

Problema 11 Màxima versemblança 5 [G]

La distribució exponencial és una distribució continua definida sobre valors reals positius. Es una distribució versàtil on la densitat de probabilitat cau de manera monòtona. Té com a funció de probabilitat:

$$p(x; \lambda) = \lambda e^{-\lambda x}, \qquad x > 0$$

on x és el valor i $\lambda > 0$ el paràmetre. Considerem un experiment aleatori en què mesurem una determinada variable aleatòria X, que segueix una distribució exponencial, cosa que escrivim $X \sim \text{Exp}(\lambda)$. Prenem n mesures independents de X i obtenim una mostra aleatòria simple $\{x_1, \ldots, x_n\}$, on cada x_i és una realització de X, per $i = 1, \ldots, n$. Es demana:

1. Construïu la funció log-versemblança (negativa) de la mostra.
2. Trobeu l'estimador de màxima versemblança per λ a partir de la mostra i demostreu que realment és un mínim (i no un extrem qualsevol).
3. Calculeu el biaix de l'estimador. Concluïu si aquest estimador és biaixat. Pista: és útil saber que, si $X_1, \ldots, X_n$ es distribueixen $\text{Exp}(\lambda)$, llavors la seva suma es distribueix com $\text{Gamma}(n, \lambda^{-1})$.
4. Un amic es vanta de que usa una marca de pneumàtics que dura molt. Fins i tot ha près nota de la durada (en milers de km): $\{35.2, 41, 44.7, 38.6, 41.5\}$. Quin valor estimaríeu per λ?
5. Com trobaríeu una estimació per la probabilitat que els pneumàtics li durin més de 50.000 km?

. . . . . .

Problema 12 Qui té un amic ...

Tenim una mostra aleatòria simple de mida $N = 2$, formada per valors que provenen de dues variables aleatòries X_1, X_2, amb la mateixa distribució de probabilitat (desconeguda) i tals que $\mathbb{E}[X_1] = \mathbb{E}[X_2] = \mu$ i $Var(X_1) = Var(X_2) = \sigma^2$. Usant coneixements vistos a classe, sabem que l'EMV de μ és $(X_1 + X_2)/2$, que és no biaixat. Un amic ens suggereix que podríem trobar-ne un de millor, combinant els valors de les dues variables d'una manera més general. Ens proposa l'estimador $\alpha_1 X_1 + \alpha_2 X_2$, i ho deixa així ...

1. Quina condició cal imposar sobre les alfes per tal que aquest estimador sigui no biaixat?
2. Demostrar que aquest mètode no dóna un estimador millor que el que ja tenim.

. . . . . .