Problema 2

Domanda 1

Si dimostri che la stima della regressione *ridge*

$$\hat{\boldsymbol{\beta}}_\lambda^{\mathrm{R}} = (\mathbf{X}^\intercal \mathbf{X} + \lambda \mathbf{I})^{-1}\mathbf{X}^\intercal \mathbf{y}$$

basata sul vettore risposta (centrata) $\underset{n\times 1}{\mathbf{y}}$ e sulla matrice del disegno (centrata) $\underset{n\times p}{\mathbf{X}}$ e per un certo valore di $\lambda > 0$ fissato, si possono ottenere mediante la regressione OLS calcolata sui dati "aumentati"

$$\underset{(n+p)\times 1}{\tilde{\mathbf{y}}} = \begin{bmatrix} \underset{n\times 1}{\mathbf{y}} \\ \underset{p\times 1}{\mathbf{0}} \end{bmatrix}, \quad \underset{(n+p)\times p}{\tilde{\mathbf{X}}} = \begin{bmatrix} \underset{n\times p}{\mathbf{X}} \\ \sqrt{\lambda}\cdot \underset{p\times p}{\mathbf{I}} \end{bmatrix}.$$

Domanda 2

1. Si supponga che vogliamo utilizzare la regressione *ridge* sui dati $\underset{n\times 1}{\mathbf{y}}$ e $\underset{n\times p}{\mathbf{X}}$ con $n \gg p$. Una complicazione è che i dati sono stati suddivisi in m dataset di dimensione $n/m \in \mathbb{N}$,

$$\underset{n\times 1}{\mathbf{y}} = \begin{bmatrix} \underset{n/m\times 1}{\mathbf{y}}^{(1)} \\ \underset{n/m\times 1}{\mathbf{y}}^{(2)} \\ \vdots \\ \underset{n/m\times 1}{\mathbf{y}}^{(m)} \end{bmatrix}, \quad \underset{n\times p}{\mathbf{X}} = \begin{bmatrix} \underset{n/m\times p}{\mathbf{X}}^{(1)} \\ \underset{n/m\times p}{\mathbf{X}}^{(2)} \\ \vdots \\ \underset{n/m\times p}{\mathbf{X}}^{(m)} \end{bmatrix}.$$

   e ciascun dataset $\underset{n/m\times 1}{\mathbf{y}}^{(k)}$, $\underset{n/m\times p}{\mathbf{X}}^{(k)}$ è disponibile su server diverso. Siccome lo spostamento di grandi quantità di dati tra server è costoso, si spieghi come ottenere la stima ridge $\hat{\boldsymbol{\beta}}_\lambda^{\mathrm{R}}$ comunicando ad un server centrale al massimo $p(p+1)$ numeri in totale da ciascun server (ad esempio, se inviate da un server una matrice $p \times p$ ed un vettore $p \times 1$, questo corrisponde a $p^2 + p$ numeri). Si spieghi quindi come ottenere $\hat{\boldsymbol{\beta}}_\lambda^{\mathrm{R}}$ specificando:

   a) i calcoli da effettuare su ciascun server;
   b) i calcoli da effettuare sul server centrale.

2. Si supponga ora che $p \gg n$ e che le colonne di $\mathbf{X}$ siano state distribuite su m server diversi, quindi ciascun server contiene un numero $p/m \in \mathbb{N}$ di colonne di $\mathbf{X}$. Il server centrale invece contiene il vettore di risposta $\mathbf{y}$. Si spieghi come si può ottenere la stima $\mathbf{X}\hat{\boldsymbol{\beta}}_\lambda^{\mathrm{R}}$ comunicando al server centrale solo n^2 numeri in totale da ciascun server. Si spieghi quindi come ottenere $\mathbf{X}\hat{\boldsymbol{\beta}}_\lambda^{\mathrm{R}}$ specificando:

   a) i calcoli da effettuare su ciascun server;
   b) i calcoli da effettuare sul server centrale.

Suggerimento: Per 1., potrebbe tornare utile la lettura del paragrafo 2.2 del libro Azzalini e Scarpa (2004).

Domanda 3

Ipotizzando che il vero modello che genera i dati sia $\underset{n\times 1}{\mathbf{y}} = \underset{n\times p}{\mathbf{X}}\underset{p\times 1}{\boldsymbol{\beta}^0} + \underset{n\times 1}{\boldsymbol{\varepsilon}}$ dove $n > p$, $\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$, $\mathbb{V}\text{ar}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$ e la matrice $\mathbf{X}$ (fissata) ha rango pieno p.

Sia $\underset{n\times p}{\mathbf{X}} = \underset{n\times p}{\mathbf{U}}\underset{p\times p}{\mathbf{D}}\underset{p\times p}{\mathbf{V}}^\mathsf{T}$ la decomposizione a valori singolari (SVD in forma ridotta) di $\mathbf{X}$ dove $\mathbf{U}$ ha colonne ortonormali ($\mathbf{U}^\mathsf{T}\mathbf{U} = \mathbf{I}$), $\mathbf{D}$ è una matrice diagonale in cui gli elementi della diagonale sono i valori singolari ordinati in modo decrescente ($D_{11} \geq D_{22} \geq \ldots \geq D_{pp} \geq 0$), e $\mathbf{V}$ è una matrice ortogonale ($\mathbf{V}^\mathsf{T}\mathbf{V} = \mathbf{V}\mathbf{V}^\mathsf{T} = \mathbf{I}$).

Sia $\boldsymbol{\gamma} = \mathbf{U}^\mathsf{T}\mathbf{X}\boldsymbol{\beta}^0$, e fissato un certo valore di $\lambda > 0$, si può dimostrare che

$$\frac{1}{n}\mathbb{E}\|\mathbf{X}\boldsymbol{\beta}^0 - \mathbf{X}\hat{\boldsymbol{\beta}}^{\mathrm{R}}_\lambda\|_2^2 = \frac{1}{n}\sum_{j=1}^p \left(\frac{\lambda}{\lambda + D_{jj}^2}\right)^2 \gamma_j^2 + \frac{\sigma^2}{n}\sum_{j=1}^p \frac{D_{jj}^4}{(\lambda + D_{jj}^2)^2}.$$

Si supponga che $\|\mathbf{X}\boldsymbol{\beta}^0\|_2^2 = n$. Quale $\boldsymbol{\gamma}$ massimizza $\frac{1}{n}\mathbb{E}\|\mathbf{X}\boldsymbol{\beta}^0 - \mathbf{X}\hat{\boldsymbol{\beta}}^{\mathrm{R}}_\lambda\|_2^2$? Quale $\boldsymbol{\gamma}$ lo minimizza?

Regolamento

1. Bisogna consegnare entro la scadenza prevista (per i gruppi, quella indicata, per i lavori individuali, almeno una settimana prima della della data di esame) un **UNICO** file in formato **.PDF** (non sono ammessi altri tipi di file) contenente le risposte alle domande. Per le risposte, è ammissibile scannerizzare un testo scritto a mano oppure utilizzare un editor di testo (e.g. LaTex, Microsoft Word, etc.) e salvare il file in formato .PDF. Il file deve essere nominato nel seguente modo: [MATRICOLA]_HW2.pdf (e.g. 2575_HW2.pdf) per i lavori individuali e [NOME DEL GRUPPO]_HW2.pdf per i lavori di gruppo. Il file dovrà essere caricato sulla pagina MOODLE in corrispondenza all'HOMEWORK2. Per i lavori di gruppo, **tutti** i componenti del gruppo devono caricare lo stesso file. Sarà possibile effettuare **una sola** sottomissione finale (vi verrà chiesta conferma, non sono ammesse consegne via e-mail). Per tutti gli studenti, il mancato rispetto della scadenza prevista corrisponde ad un punteggio di 0.

2. La valutazione si baserà sulla correttezza, chiarezza e precisione delle risposte fornite. Non è previsto un limite di pagine per il file da consegnare, ma verrà premiata la capacità di sintesi, ovvero una struttura argomentativa ben articolata, soluzioni eleganti e leggibili, con le conclusioni che rispondono in modo specifico e puntuale alla domanda iniziale. E' possibile utilizzare fonti (libri, Internet, persone e così via) ma è richiesto di citarle nel testo. L'uso di fonti senza citarle si traduce in un voto nullo.

3. Il docente si riserva la possibilità di chiedere a qualunque studente di spiegare le risposte fornite (ed eventualmente del codice utilizzato). Per i lavori individuali, questa spiegazione (se richiesta) avverrà il giorno della prova scritta. Il punteggio ottenuto scade alla fine dell'Anno Accademico 2020/21. Tutti gli studenti sono tenuti ad aderire ad un codice di condotta, che vieta il plagio, la falsificazione, l'assistenza non autorizzata, imbrogli e altri atti gravi di disonestà accademica. Comportamenti non corretti possono essere soggetti a provvedimenti disciplinari come da Art. 35 e 36 del Regolamento Didattico di Ateneo.