La Régression Linéaire Pénalisée (en Grande Dimension)

Cours#2 - M2 MIAGE Unité "Données Massives"

Clément Lejeune

Institut de Recherche en Informatique de Toulouse,
Université Toulouse III Paul Sabatier

8 Décembre 2021

Rappels

Avantages de la LR

- La prédiction du modèle est facile à expliquer (linéaire),

Rappels

Avantages de la LR

- La prédiction du modèle est facile à expliquer (linéaire),
- en basse dimension, l'apprentissage des paramètres est rapide.

Rappels

Avantages de la LR

- La prédiction du modèle est facile à expliquer (linéaire),
- en basse dimension, l'apprentissage des paramètres est rapide.

Inconvénients

- En grande dimension, la prédiction devient incertaine,

Rappels

Avantages de la LR

- La prédiction du modèle est facile à expliquer (linéaire),
- en basse dimension, l'apprentissage des paramètres est rapide.

Inconvénients

- En grande dimension, la prédiction devient incertaine,
- les variables non-explicatives sont considérées au même titre que les "bonnes" variables explicatives,

Rappels

Avantages de la LR

- La prédiction du modèle est facile à expliquer (linéaire),
- en basse dimension, l'apprentissage des paramètres est rapide.

Inconvénients

- En grande dimension, la prédiction devient incertaine,
- les variables non-explicatives sont considérées au même titre que les "bonnes" variables explicatives,
- la présence de corrélation entre certaines variables explicatives diminuent le pouvoir prédictif du modèle (*c-à-d* l'erreur de généralisation).

Sommaire

La malédiction / le fléau de la dimension

On parle de malédiction de la dimension quand le nombre de variables explicatives d devient très "grand" devant le nombre d'observations n *c-à-d* $d >> n$. Pourquoi ?

- Plus il y a de paramètres à apprendre, plus les $\hat{y}_i$ deviennent incertaines (grande variance), hors-scope,
- Plus il y a de paramètres à apprendre, plus l'apprentissage peut être long,
- Le minimiseur $\hat{\mathbf{w}}$:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{1}{2n} \sum_{i=1}^{n} \underbrace{(y_i - \langle \mathbf{w}, x_i \rangle)^2}_{\ell(y_i, h(x_i))} \quad (1)$$

  n'est pas unique.

Rappel et conséquences des hypothèses la LR classique

Pourquoi $\hat{\mathbf{w}}$ n'est pas unique quand $d > n$? Au cours précédent, on a appris $\widehat{\mathbf{w}}$ en résolvant (selon $\mathbf{x}$) sous l'hypothèse $n > d$:
$\nabla g = \left(\frac{\partial g}{\partial w_1}, \ldots, \frac{\partial g}{\partial w_d}\right)^\top = \frac{1}{n} X^\top (X\mathbf{w} - \mathbf{y}) = \mathbf{0}$ que l'on réécrit:

$$\underbrace{X^\top X}_{G} \mathbf{w} = X^\top \mathbf{y}$$

Le $\mathbf{w}$ vérifiant cette équation est:

$$\widehat{\mathbf{w}} = G^{-1} X^\top \mathbf{y}$$

G^{-1} existe si G est inversible (*c-à-d* G^{-1} est l'unique matrice telle que $G^{-1}G = GG^{-1} = I$).

La malédiction / le fléau de la dimension

Inversibilité d'une matrice

Les conditions suivantes sont équivalentes sur une matrice carrée $A \in \mathbb{R}^{k \times k}$:

- A est inversible si ses k colonnes sont linéairement indépendantes,

La malédiction / le fléau de la dimension

Inversibilité d'une matrice

Les conditions suivantes sont équivalentes sur une matrice carrée $A \in \mathbb{R}^{k \times k}$:

- A est inversible si ses k colonnes sont linéairement indépendantes,
- Le rang de A (nombre de colonnes et lignes linéairement indépendantes) vaut k,

La malédiction / le fléau de la dimension

Inversibilité d'une matrice

Les conditions suivantes sont équivalentes sur une matrice carrée $A \in \mathbb{R}^{k \times k}$:

- A est inversible si ses k colonnes sont linéairement indépendantes,
- Le rang de A (nombre de colonnes et lignes linéairement indépendantes) vaut k,
- Le déterminant de A est non-nul.

On a les propriétés suivantes sur le rang d'une matrice arbitraire $B \in \mathbb{R}^{k \times q}$ (non carrée):

- $rang(B) \leqslant \min(k, q)$
- $rang(B) = rang(B^\top)$
- $rang(AB) \leqslant \min(rang(A), rang(B))$

La malédiction / le fléau de la dimension

Donc dans notre cas, puisque $n < d$, on a:

$$rang(X^\top X) = rang(G) \leqslant \min((d,n),(n,d)) = n < d$$

Conclusion: $G \in \mathbb{R}^{d \times d}$ est au plus de rang n, donc pas de rang d, et ainsi **n'est pas inversible**: dans ce cas G^{-1} n'existe pas !

Conséquences du fléau de la dimension pour une RL

En grande dimension, $n < d$ (plus de variables que d'observations):

- L'apprentissage des paramètres $\mathbf{w} = (w_1, \dots, w_d)$ d'un modèle linéaire ne se résout pas analytiquement avec la formule $\widehat{\mathbf{w}} = (X^\top X)^{-1} X^\top \mathbf{y}$,

La malédiction / le fléau de la dimension

Donc dans notre cas, puisque $n < d$, on a:

$$rang(X^\top X) = rang(G) \leqslant \min((d,n),(n,d)) = n < d$$

Conclusion: $G \in \mathbb{R}^{d\times d}$ est au plus de rang n, donc pas de rang d, et ainsi **n'est pas inversible**: dans ce cas G^{-1} n'existe pas !

Conséquences du fléau de la dimension pour une RL

En grande dimension, $n < d$ (plus de variables que d'observations):

- L'apprentissage des paramètres $\mathbf{w} = (w_1, \dots, w_d)$ d'un modèle linéaire ne se résout pas analytiquement avec la formule $\widehat{\mathbf{w}} = (X^\top X)^{-1} X^\top \mathbf{y}$,
- le problème de minimisation défini au cours#1: $g(\mathbf{w}) = \frac{1}{2n}\sum_{i=1}^{n} \underbrace{(y_i - \langle \mathbf{w}, x_i\rangle)^2}_{\ell(y_i, h(x_i))}$ n'admet pas de solution unique g n'est plus *strictement convexe*.

La malédiction / le fléau de la dimension

"faible" inversibilité en basse dimension

La non-inversibilité de $G = X^\top X$ peut aussi avoir lieu en basse dimension, $n > d$. En effet, dans le cas où les colonnes de X (variables prédictives) sont linéairement corrélées, G devient "faiblement inversible" car son déterminant tend vers 0.

Comment remédier au fléau de la dimension ?

On peut envisager plusieurs stratégies en grande dimension, $n < d$:

- Une première est de diminuer le nombre de variables prédictives "à la main": regarder si certaine(s) variable(s) ont une influence négligeable (voire nulle) sur la prédiction puis les enlever des données $=>$ long, incertain et fastidieux,

Comment remédier au fléau de la dimension ?

On peut envisager plusieurs stratégies en grande dimension, $n < d$:

- Une première est de diminuer le nombre de variables prédictives "à la main": regarder si certaine(s) variable(s) ont une influence négligeable (voire nulle) sur la prédiction puis les enlever des données $=>$ long, incertain et fastidieux,
- une seconde: appliquer une procédure itérative (descente de gradient) à la fonction objective $g(\mathbf{w}) =>$ long, ne résout que le problème de non-inversibilité, (le minimiseur trouvé n'est pas unique),

Comment remédier au fléau de la dimension ?

On peut envisager plusieurs stratégies en grande dimension, $n < d$:

- Une première est de diminuer le nombre de variables prédictives "à la main": regarder si certaine(s) variable(s) ont une influence négligeable (voire nulle) sur la prédiction puis les enlever des données $=>$ long, incertain et fastidieux,
- une seconde: appliquer une procédure itérative (descente de gradient) à la fonction objective $g(\mathbf{w}) =>$ long, ne résout que le problème de non-inversibilité, (le minimiseur trouvé n'est pas unique),
- une troisième: modifier le risque empirique $\frac{1}{n}\sum_{i=1}^{n} \ell\big(y_i, h(x_i)\big)$ de manière à diminuer l'effet de l'*ensemble* variables sur la prédiction.

On va s'intéresser à la troisième stratégie et voir qu'elle peut adresser toutes les limitations de l'apprentissage "classique" de la RL (appelée régression linéaire par *moindres carrés ordinaires*.

La régularisation *ridge* du risque empirique

Tout le problème de la grande dimension réside dans la non-inversibilité de $G = X^\top X$ (et encore pire si les prédicteurs sont corrélés). Idéalement, ce qu'on aimerait, c'est:

- apprendre les paramètres de façon analytique (comme avec la RL ordinaire)

La régularisation *ridge* du risque empirique

Tout le problème de la grande dimension réside dans la non-inversibilité de $G = X^\top X$ (et encore pire si les prédicteurs sont corrélés). Idéalement, ce qu'on aimerait, c'est:

- apprendre les paramètres de façon analytique (comme avec la RL ordinaire)
- atténuer la corrélation des prédicteurs (qui peut aussi avoir lieu en basse dimension, cf. TP),

La régularisation *ridge* du risque empirique

Tout le problème de la grande dimension réside dans la non-inversibilité de $G = X^\top X$ (et encore pire si les prédicteurs sont corrélés). Idéalement, ce qu'on aimerait, c'est:

- apprendre les paramètres de façon analytique (comme avec la RL ordinaire)
- atténuer la corrélation des prédicteurs (qui peut aussi avoir lieu en basse dimension, cf. TP),
- intégrer une contrainte de "sélection de variables" de façon à choisir les variables explicatives directement pendant l'apprentissage.

Sélection de variables ? 🤔

Ne pas sélectionner la variable X_j revient à avoir le poids appris $\hat{w}_j = 0$:

$$\hat{y}_i = \hat{w}_1 x_{i1} + \cdots + \underbrace{0}_{=\hat{w}_j} \times x_{ij} + \cdots + \hat{w}_d x_{id}$$

Introduction d'une contrainte dans l'apprentissage

On va voir que la troisième stratégie résout tous nos problèmes (quasi). Cette (famille de) stratégie s'appelle la *régularisation*, qui eu lieu de minimiser le risque empirique classique:

$$g(\mathbf{w} = (w_1, \dots, w_d)) = \frac{1}{2n} \sum_{i}^{n} (y_i - \sum_{j=1}^{d} w_j x_{ij})^2$$

On va voir que la troisième stratégie résout tous nos problèmes (quasi). Cette (famille de) stratégie s'appelle la *régularisation*, qui eu lieu de minimiser le risque empirique classique:

$$g(\mathbf{w} = (w_1, \dots, w_d)) = \frac{1}{2n} \sum_{i}^{n} (y_i - \sum_{j=1}^{d} w_j x_{ij})^2$$

vise à minimiser le risque empirique régularisé:

$$g_\lambda(\mathbf{w} = (w_1, \dots, w_d)) = \underbrace{\frac{1}{2n} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{d} w_j x_{ij})^2}_{\text{risque empirique}} + \underbrace{\frac{\lambda}{2} \sum_{j=1}^{d} w_j^2}_{\text{régularisation } \textit{ridge}}$$

où $\lambda > 0$. Ce problème est un type de régression pénalisée et la régularisation est appelée *ridge* Hastie, Tibshirani, and Friedman 2009; James et al. 2013 ou bien de Tikhonov Boyd and Vandenberghe 2004; Hoerl and Kennard 1970.

Introduction d'une contrainte dans l'apprentissageEffet de la pénalité/régularisation ridge

Introduction d'une contrainte dans l'apprentissage

Pourquoi ce terme supplémentaire est une bonne nouvelle ? 🤔
Réécrivons le problème sous forme vectorielle 🤓:

$$\text{MINIMIZE } g_\lambda(\mathbf{w}) = \frac{1}{2n}\|\mathbf{y} - X\mathbf{w}\|_2^2 + \frac{\lambda}{2}\|\mathbf{w}\|_2^2$$

où $\|\mathbf{u}\|_2 = \sqrt{\sum_{j=1}^d u_j^2}$ dénote la norme Euclidienne. Puis en concaténant $\mathbf{y}$ avec $\mathbf{0}$, et X avec $\sqrt{\lambda}I$, g_λ devient:

$$g_\lambda(\mathbf{w}) = \frac{1}{2n}\left\|\begin{pmatrix}\mathbf{y}\\ \mathbf{0}\end{pmatrix} - \begin{pmatrix}X\\ \sqrt{\lambda}I\end{pmatrix}\mathbf{w}\right\|_2^2$$

où $I =$ est la matrice identité de $\mathbb{R}^{d\times d}$.

Introduction d'une contrainte dans l'apprentissage

Par cette réécriture, g_λ est exactement de même forme qu'avec les moindres carrés ordinaires. Il suffit de poser:

$$\tilde{\mathbf{y}} = \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix} ; \tilde{X}_\lambda = \begin{pmatrix} X \\ \sqrt{\lambda} I \end{pmatrix}$$

et donc:

$$g_\lambda(\mathbf{w}) = \frac{1}{2n} \| \underbrace{\tilde{\mathbf{y}}}_{cible} - \underbrace{\tilde{X}_\lambda}_{prdicteurs} \mathbf{w} \|_2^2$$

Introduction d'une contrainte dans l'apprentissage

Par cette réécriture, g_λ est exactement de même forme qu'avec les moindres carrés ordinaires. Il suffit de poser:

$$\tilde{\mathbf{y}} = \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix} ; \tilde{X}_\lambda = \begin{pmatrix} X \\ \sqrt{\lambda} I \end{pmatrix}$$

et donc:

$$g_\lambda(\mathbf{w}) = \frac{1}{2n} \| \underbrace{\tilde{\mathbf{y}}}_{cible} - \underbrace{\tilde{X}_\lambda}_{prdicteurs} \mathbf{w} \|_2^2$$

Qu'est-ce qui a changé entre $g(\mathbf{w})$ et $g_\lambda(\mathbf{w})$? Le rang de $\tilde{X}_\lambda$ n'est plus le même que celui de X: on a ajouté des lignes linéairement indépendantes à X pour obtenir $\tilde{X}_\lambda$! On a donc:

$$rang(\tilde{X}_\lambda^\top \tilde{X}_\lambda) = rang(\tilde{G}_\lambda) \leqslant \min((d, n+d), (n+d, d)) = d$$

Apprentissage des paramètres régularisés

$\tilde{G}_\lambda \in \mathbb{R}^{d \times d}$ est au plus de rang $d > n$. Or $\tilde{G}_\lambda = X^\top X + \lambda I$ est inversible (matrice semi-définie-positive + matrice définie-positive). Conclusion: $\tilde{G}_\lambda$ est inversible donc de rang d.

Apprentissage des paramètres régularisés

$\tilde{G}_\lambda \in \mathbb{R}^{d\times d}$ est au plus de rang $d > n$. Or $\tilde{G}_\lambda = X^\top X + \lambda I$ est inversible (matrice semi-définie-positive + matrice définie-positive). Conclusion: $\tilde{G}_\lambda$ est inversible donc de rang d.
Cette propriété nous permet d'avoir les mêmes conditions qu'avec la RL ordinaire (stricte convexité, quadratique, différentiable). Pour calculer le minimiseur de g_λ, on procède de donc la même façon:

- calcul du gradient: $\nabla g_\lambda = \frac{1}{n}\tilde{X}_\lambda^\top(\tilde{X}_\lambda \mathbf{w} - \tilde{\mathbf{y}})$,

Apprentissage des paramètres régularisés

$\tilde{G}_\lambda \in \mathbb{R}^{d \times d}$ est au plus de rang $d > n$. Or $\tilde{G}_\lambda = X^\top X + \lambda I$ est inversible (matrice semi-définie-positive + matrice définie-positive). Conclusion: $\tilde{G}_\lambda$ est inversible donc de rang d.
Cette propriété nous permet d'avoir les mêmes conditions qu'avec la RL ordinaire (stricte convexité, quadratique, différentiable). Pour calculer le minimiseur de g_λ, on procède de donc la même façon:

- calcul du gradient: $\nabla g_\lambda = \frac{1}{n} \tilde{X}_\lambda^\top (\tilde{X}_\lambda \mathbf{w} - \tilde{\mathbf{y}})$,
- trouver le $\hat{\mathbf{w}}_{ridge}$ tel que $\nabla g_\lambda = \mathbf{0}$.

Apprentissage des paramètres régularisés

$\tilde{G}_\lambda \in \mathbb{R}^{d\times d}$ est au plus de rang $d > n$. Or $\tilde{G}_\lambda = X^\top X + \lambda I$ est inversible (matrice semi-définie-positive + matrice définie-positive). Conclusion: $\tilde{G}_\lambda$ est inversible donc de rang d.

Cette propriété nous permet d'avoir les mêmes conditions qu'avec la RL ordinaire (stricte convexité, quadratique, différentiable). Pour calculer le minimiseur de g_λ, on procède de donc la même façon:

- calcul du gradient: $\nabla g_\lambda = \frac{1}{n}\tilde{X}_\lambda^\top(\tilde{X}_\lambda \mathbf{w} - \tilde{\mathbf{y}})$,
- trouver le $\hat{\mathbf{w}}_{ridge}$ tel que $\nabla g_\lambda = \mathbf{0}$.

$$\nabla g_\lambda = \mathbf{0} \iff \frac{1}{n}\tilde{X}_\lambda^\top(\tilde{X}_\lambda \mathbf{w} - \tilde{\mathbf{y}}) = \mathbf{0} \iff \tilde{X}_\lambda^\top \tilde{X}_\lambda \mathbf{w} - \tilde{X}_\lambda^\top \tilde{\mathbf{y}} = \mathbf{0}$$

Or $\tilde{X}_\lambda^\top \tilde{X}_\lambda = X^\top X + \lambda I$ (inversible) et $\tilde{X}_\lambda^\top \tilde{\mathbf{y}} = X^\top \mathbf{y}$

Comment remédier au fléau de la dimension ?

Solution ridge

Pour $\lambda > 0$, $X \in \mathbb{R}^{n \times d}$ telle que $rang(X^\top X) \leqslant n < d$, les paramètres de la régression linéaire ridge se calculent selon:

$$\hat{\mathbf{w}}_{ridge} = (X^\top X + \lambda I)^{-1} X^\top \mathbf{y}.$$

Comment remédier au fléau de la dimension ?

Solution ridge

Pour $\lambda > 0$, $X \in \mathbb{R}^{n \times d}$ telle que $rang(X^\top X) \leqslant n < d$, les paramètres de la régression linéaire ridge se calculent selon:

$$\hat{\mathbf{w}}_{ridge} = (X^\top X + \lambda I)^{-1} X^\top \mathbf{y}.$$

=> Application en TP !

[1] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[2] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. 2009. DOI: 10.1007/978-0-387-84858-7. URL: https://link.springer.com/content/pdf/10.1007%2F978-0-387-84858-7.pdf%0Ahttp://link.springer.com/10.1007/978-0-387-84858-7.

[3] Arthur E Hoerl and Robert W Kennard. "Ridge regression: Biased estimation for nonorthogonal problems". In: *Technometrics* 12.1 (1970), pp. 55–67.

[4] Gareth James et al. *An Introduction to Statistical Leaning with Applications in R*. 2013, p. 441. ISBN: 9781461471370. URL: http://www-bcf.usc.edu/$\sim$gareth/ISL/ISLRFirstPrinting.pdf.