TD 7 : RÉGULARISATION

COURS D'APPRENTISSAGE, ECOLE NORMALE SUPÉRIEURE, PRINTEMPS 2013

Remi Lajugie
remi.lajugie@ens.fr

RÉSUMÉ. Dans ce TP/TD, on va commencer par chercher à illustrer le paradoxe de James-Stein. Ensuite on s'intéressera à la régression ridge, déjà vue lors du TP4 du point de vue de l'optimisation. On essaiera ici de s'intéresser aux aspects plus statistiques de cette régularisation.

1. Exercice : Estimateur de James-Stein

Dans cet exercice, on cherche a illustrer le paradoxe de James-Stein evoque en cours. Pour cela on considere un n echantillon gaussien de dimension p, de variance unite et de moyenne $\theta \in \mathbb{R}$. Pour les simulations numeriques on pourra prendre n = 500 et p = 40

1) Quel est l'estimateur du maximum de vraisemblance de θ, $\hat{\theta}_{MV}$? Quel est le risque quadratique moyen qui lui est associe ?

2) Rappelez la de nition de l'estimateur de James-Stein $\hat{\theta_{JS}}$ vu en cours ainsi que le paradoxe qui lui est associe.

3) Generez un vecteur de moyenne θ et simulez des donnees suivant une loi normale de moyenne θ et de variance unite. Tracez la courbe donnant le risque quadratique evalue empiriquement sur les donnees simulees en fonction de la norme de θ.

4) Expliquez le comportement de l'estimateur quand $\|\theta\|$ tend vers 0.

5) BONUS : (a faire a la n) On rappelle le resultat suivant, vu en cours concernant l'estimateur de James-Stein :

$$\mathbb{E}[(\theta_{JS} - \theta)^2] \leq 2\sigma^2 + \frac{\sigma^2(p-2)\|\theta\|^2}{(p-2)\sigma^2 + \|\theta\|^2}$$

. A l'aide de simulations illustrez cette borne.

2. Exercice : Sélection de modèle avec le C_p de Mallows

On considere le cas de la regression lineaire multivariee de $\mathbb{R}^p$ dans $\mathbb{R}$. On appelle X la matrice de design des donnees. On va chercher a utiliser les idees de la selection de modele telles que vous les avez vues en cours. On se place dans le même cadre que le cours, a savoir que l'on considere la classe des ETL (Estimateurs par transformation lineaire)

Tout d'abord, quelques rappels sur cette statistique. On appelle dans la suite σ^2 la variance globale associee au predicteur de la regression lineaire complete (avec tous les regresseurs), on appelle ensuite, pour un modele impliquant seulement p variables explicatives, SSE(M_p), SSE, la somme des erreurs quadratiques entre les observations y_i et les predictions faites par le modele $\hat{f}$. On de nit alors la quantite :

$$C_p = SSE/n + 2 * \sigma^2 p/n.$$

6) Sur des donnees de dimension p = 8, implementez une methode de selection des variables pour les predicteurs de la regression lineaire de modele en se fondant sur C_p. Pour chaque $i \leq p$ vous chercherez le meilleur modele impliquant i variable et le C_{i^*} associe. Vous tracerez l'evolution de ces C_{i^*} en fonction du nombre de regresseurs.

3. Exercice : Régression ridge

Dans cet exercice on considere la regression lineaire des moindres carres penalisee de $\mathbb{R}^p$ dans $\mathbb{R}$, connue aussi sous le nom de regression ridge. X designe la matrice de design des donnees, qui comporte n lignes et p colonnes.

$$\min_{w \in \mathbb{R}^p} \frac{1}{2n} \|y - X^T w\|_2^2 + \lambda \|w\|_2^2$$

7) Rappelez l'expression analytique de l'estimateur de la regression ridge ci-dessus. (Savez-vous le demontrer ?)

8) On suppose que les donnees y_i sont generees selon $y_i = x_i^T \beta + \epsilon_i$ ou ϵ_i sont des bruits i.i.d de variance σ^2. On se place dans le cadre du design fixe.

a) Montrez que l'estimateur de la regression ridge, sous ces hypotheses est un ETL. Explicitez la matrice associee que l'on notera A.

b) En utilisant la Decomposition en Valeurs Singulieres (SVD en Anglais et `svd` en MATLAB/Octave) de la matrice de design (on ecrira $X = UDV^T$ avec D la matrice des valeurs singulieres, U et V respectivement unitaires sur $\mathbb{R}^n$ et $\mathbb{R}^p$), Exprimez l'exces de risque quadratique du predicteur avec seulement σ^2, le parametre de regularisation λ et les valeurs singulieres d_k.

9) En notant $\tilde{\beta}$ la representation de β dans une base bien choisie, le biais de la regression ridge s'exprime en fonction de $\tilde{\beta}$, σ^2, λ et les coefficients de la svd de A.

10) On considere maintenant p = 8. Generez 100 points dans $\mathbb{R}^p$ suivant une loi uniforme. Generez un vecteur y de reponses dans $\mathbb{R}$ en les generant suivant une fonction deterministe des donnees d'entree que vous augmenterez d'un bruit.

11) En faisant varier la valeur du parametre de regularisation (sur une echelle logarithmique par exemple), representez en fonction de lambda, la norme du vecteur donne par la regression ridge (on dit aussi que l'on represente l'evolution de la norme sur le \chemin de regularisation").

12) Representez l'evolution des coefficients de la regression le long du chemin de regularisation (utilisez l'expression analytique des questions precedentes), que constatez vous ?