Annexe 1

Sofia HARROUCH

30 mars 2018

Les méthode de régularisation (Ridge, lasso et elastic net)

Dans le domaine des mathématiques et de la statistique, et plus particulièrement dans le domaine de l'apprentissage automatique, la régularisation fait référence à un processus consistant à ajouter de l'information à un problème pour éviter **le surapprentissage**. Cette information prend généralement la forme d'une pénalité.

Une méthode généralement utilisée est de pénaliser les valeurs extrêmes des paramètres, qui correspondent souvent à un surapprentissage. Pour cela, on va utiliser une norme sur ces paramètres, que l'on va ajouter à la fonction qu'on cherche à minimiser. Les normes les plus couramment employées pour cela sont L_1??? etL_2. L_1 offre l'avantage de faire une sélection de paramètres, mais elle n'est pas différentiable, ce qui peut être un inconvénient pour les algorithmes utilisant un calcul de gradient pour l'optimisation. l_2 ????? Cette régularisation ou rétrécissement permet alors d'avoir des coefficients qui peuvent ˆetre estimé exactement par zéro. Par conséquent, ces méthodes peuvent effectuer une sélection des variables importantes pour la variable réponse.

Dans le cadre d'une régression multiple $Y = X\beta + \epsilon$ tel que $\beta = (\beta_0, \beta_1, \beta_2, ..., \beta_p)^T$ et $X = (1, X_1, ..., X_p)$, on se trouve des fois dans le cas où le nombre de paramètre p est supérieur au nombre de données n. En présence de Ce nombre élevés de prédicteurs, On aura besoin de ces méthodes pour réduire le nombre de paramètres car sinon la solution donnée par la méthode de moindres carré ordinaire n'est pas unique et la variance a tendance d'être grande et le biais est petit. Alors que les méthodes de régularisation offrent une réduction de variance et une petite augmentation de biais.

La méthode de Ridge

La méthode de Ridge est une technique de régularisation qui se base sur la norme L_2. cette méthode a comme pénalité: $p(\beta) = ||\beta||_2^2$ avec $||\beta||_2^2 = \sum_{i=1}^{p} \beta_i^2$. L'estimateur de Ridge de β est défini par: $\hat{\beta}^{Ridge} = argmin_\beta\{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2\}$ sous la contrainte de $\sum_{j=1}^{p} \beta_j^2 \leq t$. L'estimateur de Ridge est donc: $\hat{\beta}^{Ridge} = (X^TX + \lambda I)^{-1}X^Ty$ Cette méthode estutilisée plus dans le cas où on a une corrélation entre les varibles exlpicatives; là où X^TX a des valeurs proche de zéro et la MMCO n'est pas satisfaisante. Cette méthode permet de rajouter un terme λ pour augmenter la valeur de X^TX pour les rendre stable. Et par là elle contrôle la variance des estimateurs en pénalisant les grandes valeurs de $\hat{\beta}$ ce qui a comme avantage l'obtention d'un erreur de prédiction moins faible. Par contre, elle ne permet pas d'avoir un modèle parcimonieux car cette méthode ne pénalise pas les variables nuisibles par des coefficients nuls, par la suite elle introduit tout les variables explicatives dans le modèle ce qui complique l'interprétation du modèle.

La méthode du LASSO

C'est une technique basée sur la norme L_1. La pénalité de cette méthode est donnée par :$p(\beta) = ||\beta||_1$ avec $||\beta||_1 = \sum_{i=1}^{p} |\beta_i|$. L'estimateur de β par la méthode Lasso est défini par:

$$\hat{\beta}^{Lasso} = argmin_\beta\{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2\}$$

tel que: $\sum_{j=1}^{p} |\beta|_j \leq t$ Par un calcul analytique, la soluton de ce problème est:

$$\hat{\beta}_j^{lasso} = signe\{\hat{\beta}_j^{MCO}\}(|\hat{\beta}_j^{MCO}| - \lambda)_+$$

Cette méthode permet de créér une parcimonie, car elle élimine les variables nuisibles dans le modèle en estimant leur coefficients par des zéros, donc elle rétrecit les coefficients de β à zéro et permet de choisir les varables qui contribuent le plus dans le modèle. Par contre, elle est innapropriée dans le cas où un ensemble de prédicteurs sont fortement corrélés car elle choisit un parmis eux et annules les autres. Sans oublié le fait qu'elle choist un maximum de n variables dans le cas où p > n. Ces derniers problèmes sont traités par la méthode qui suit.

La méthode d Elastique Net(EN)

Cette méthode de régularisation combine les deux norme (L_1 et L_2). Elle est une compromis entre la méthode du Lasso et la méthode du Ridge. Sa pénalité est donnée par:$p(\beta) = \lambda_1||\beta||_1 + \lambda_2||\beta||_2^2$. L'estimateur $\hat{\beta}^{EN} = argmin_\beta\{||y - X\beta||^2 \ + \lambda_1||\beta||_1 + \lambda_2||\beta||_2^2\}$. En introduisons la pénalité l_1, on s'assure de la génération d'un modèle parcimonieux. Alors que l'ajout de la pénalité quadratique l_2 assume la suppression de la limitation du nombre des variables selectionnées, enourage l'effet du groupe (groupe effect), car elle permet de selectionner un nombre de paramètre p supérieur à n. Sans oublié aussi le fait que cette méthode permet de tenir en compte la corrélation entre les prédicteurs.