ML 14 Géométrie de la régression ridge et SVD

1 Coefficients de la régression ridge en fonction du SVD

Exprimons le calcul des coefficients d'une régression ridge en fonction de la décomposition en valeurs singulières de la matrice des données observées $\mathbf{X} \in \mathbb{R}^{M \times N}$.

$$
\begin{aligned}
& \hat{\beta}_\lambda \\
= & \{\text{Voir la dérivation de la régression ridge dans un précédent module.}\} \\
& (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} \\
= & \{\text{SVD de } \mathbf{X} : \mathbf{UDV}^T \quad \mathbf{U} \in \mathbb{R}^{M \times M} \ , \ \mathbf{D} \in \mathbb{R}^{M \times N} \ , \ \mathbf{V} \in \mathbb{R}^{N \times N}\} \\
& (\mathbf{VD}^T\mathbf{U}^T\mathbf{UDV}^T + \lambda\mathbf{I})^{-1}\mathbf{VD}^T\mathbf{U}^T\mathbf{y} \\
= & \{\mathbf{U} \text{ et } \mathbf{V} \text{ sont orthogonales : } \mathbf{I} = \mathbf{VV}^T\} \\
& (\mathbf{VD}^T\mathbf{DV}^T + \lambda\mathbf{VV}^T)^{-1}\mathbf{VD}^T\mathbf{U}^T\mathbf{y} \\
= & \\
& (\mathbf{V}(\mathbf{D}^T\mathbf{D} + \lambda\mathbf{I})\mathbf{V}^T)^{-1}\mathbf{VD}^T\mathbf{U}^T\mathbf{y} \\
= & \{(\mathbf{XY})^{-1} = \mathbf{Y}^{-1}\mathbf{X}^{-1} \quad \mathbf{V} \text{ est orthogonale : } \mathbf{V}^{-1} = \mathbf{V}^T\} \\
& \mathbf{V}(\mathbf{D}^T\mathbf{D} + \lambda\mathbf{I})^{-1}\mathbf{V}^T\mathbf{VD}^T\mathbf{U}^T\mathbf{y} \\
= & \\
& \mathbf{V}(\mathbf{D}^T\mathbf{D} + \lambda\mathbf{I})^{-1}\mathbf{D}^T\mathbf{U}^T\mathbf{y} \\
= & \{\text{Soit } d_j \text{ le jème élément sur la diagonale de } \mathbf{D}, \mathbf{u_j} \text{ et } \mathbf{v_j} \text{ les jèmes colonnes de respectivement } \mathbf{U} \text{ et } \mathbf{V}\} \\
& \sum_{d_j > 0} \mathbf{v_j} \frac{d_j}{d_j^2 + \lambda} \mathbf{u_j}^T \mathbf{y}
\end{aligned}
$$

Remarquons ainsi que la décomposition en valeurs singulières de $\mathbf{X}$ donne les coefficients de la régression ridge pour toutes les valeurs possibles du coefficient de régularisation λ.

2 Régression ridge et géométrie

Observons la relation entre les étiquettes prédites $\hat{\mathbf{y}}_\lambda$ et les étiquettes observées $\mathbf{y}$.

$$
\begin{aligned}
\hat{\mathbf{y}}_\lambda &= \mathbf{X}\hat{\beta}_\lambda \\
&= \mathbf{UDV}^T\mathbf{V}(\mathbf{D}^T\mathbf{D} + \lambda\mathbf{I})^{-1}\mathbf{D}^T\mathbf{U}^T\mathbf{y} \\
&= \sum_{d_j > 0} \mathbf{u_j} \frac{d_j^2}{d_j^2 + \lambda} \mathbf{u_j}^T \mathbf{y}
\end{aligned}
$$

Nous remarquons qu'en labsence de régularisation, $\lambda = 0$, les valeurs estimées $\hat{\mathbf{y}}$ sont les projections sur les axes principaux $\mathbf{u_j}$ – qui couvrent l'espace des colonnes de $\mathbf{X}$, i.e. $Im(\mathbf{X})$ – des valeurs observées $\mathbf{y}$.

En présence de régularisation, $\lambda > 0$, les coordonnées, sur les axes principaux, de l'estimation $\hat{\mathbf{y}}_\lambda$ sont de plus en plus contractées lorsqu'on progresse vers les axes qui expliquent de moins en moins la variabilités des données.