Universität Salzburg — Florian Graf

Machine Learning

Übungsblatt **6** — 25 Punkte

Aufgabe 1. *Ridge / Lasso Regression – Theorie I* 9 P.

Gegeben sei eine Stichprobe $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\} \subset \mathbb{R}^d$. Wir wollen ein lineares Regressionsmodell $\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{w}}$ lernen, wobei $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top \in \mathbb{R}^{n \times d}$ und $\mathbf{y} = (y_1, \ldots, y_n) \in \mathbb{R}^n$.

Im Falle von Least-Squares Regression wird $\hat{\mathbf{w}}$ durch das Optimierungsproblem $\hat{\mathbf{w}} = \operatorname{argmin}_{\mathbf{w}} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$ bestimmt. Wie in der Vorlesung gezeigt wurde, ist $\nabla_{\mathbf{w}} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 = 2(\mathbf{X}^\top\mathbf{X}\mathbf{w} - \mathbf{X}^\top\mathbf{y})$ und $\mathbf{H}(\mathbf{w}) = 2\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 = \mathbf{X}^\top\mathbf{X}$. Folglich ist, sofern $\mathbf{X}^\top\mathbf{X}$ invertierbar ist, $\hat{\mathbf{w}}_{\text{LS}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$.

(a) Ridge Regression entspricht dem Minimierungsproblem $\hat{\mathbf{w}}_{\text{Ridge}} = \operatorname{argmin}_{\mathbf{w}} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 + \lambda\|\mathbf{w}\|_2^2$, wobei $\lambda > 0$.

Zeigen Sie, dass die Funktion $\mathbf{w} \mapsto \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 + \lambda\|\mathbf{w}\|_2^2$ ein Minimum an der Stelle $\hat{\mathbf{w}}_{\text{Ridge}} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}$ hat.

Zeigen Sie außerdem, dass die Matrix $\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}$ invertierbar ist. Tipp: Jede positiv definite Matrix ist invertierbar.

(b) Lasso Regression entspricht dem Minimierungsproblem $\hat{\mathbf{w}}_{\text{Lasso}} = \operatorname{argmin}_{\mathbf{w}} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 + \lambda\|\mathbf{w}\|_1$, wobei $\lambda > 0$.

Berechnen Sie zunächst den Gradienten $\nabla f(\mathbf{w})$ für den Fall das alle Koordinaten $w_1, \ldots, w_D \neq 0$. Zeigen Sie, dass

$$\nabla f(\mathbf{w}) = 2\mathbf{X}^\top\mathbf{X}\,(\mathbf{w} - \hat{\mathbf{w}}_{\text{LS}}) + \lambda \operatorname{sign}\mathbf{w} \quad , \quad \text{wobei} \quad \operatorname{sign}\mathbf{w} = \begin{pmatrix} \operatorname{sign} w_1 \\ \vdots \\ \operatorname{sign} w_D \end{pmatrix} .$$

(c) Wir definieren

$$\hat{w}_i = \begin{cases} \mathbf{x}_{:,\mathbf{i}}{}^\top\mathbf{y} + \lambda/2 & \text{falls } \mathbf{x}_{:,\mathbf{i}}{}^\top\mathbf{y} < -\lambda/2 \\ 0 & \text{falls } \mathbf{x}_{:,\mathbf{i}}{}^\top\mathbf{y} \in [-\lambda/2, \lambda/2] \\ \mathbf{x}_{:,\mathbf{i}}{}^\top\mathbf{y} - \lambda/2 & \text{falls } \mathbf{x}_{:,\mathbf{i}}{}^\top\mathbf{y} > \lambda/2 \end{cases} ,$$

wobei $\mathbf{x}_{:,i}$ der i-te Spaltenvektor von $\mathbf{X}$ ist.

Zeigen Sie, dass falls $\mathbf{X}^\top\mathbf{X} = \mathbf{I}$, der Vektor $\hat{\mathbf{w}}_{\text{Lasso}} = (\hat{w}_1, \ldots \hat{w}_D)^\top$ das Minimum von f ist. Sie dürfen ohne Beweis verwenden, dass die Funktion f genau ein Minimum besitzt.

Aufgabe 2. *Ridge / Lasso Regression – Theorie II* 5 P.

Wir führen eine Lineare Regression durch, wobei die Spaltenvektoren $\mathbf{x}_{:,i}$ der Datenmatrix $\mathbf{X}$ orthonormal sind, d.h.

$$\mathbf{x}_{:,i}^\top\mathbf{x}_{:,j} = \begin{cases} 0 & \text{falls } i \neq j \\ 1 & \text{falls } i = j \end{cases} .$$

Es sei $\hat{w}_i$ die i-te Koordinate unseres Schätzers $\hat{\mathbf{w}}$ von $\mathbf{w}$ und es sei $c_i = \mathbf{x}_{:,i}^\top\mathbf{y}$. In Abb. 1 ist $\hat{w}_i$ gegen c_i dargestellt, wobei $\hat{w}_i$ durch 3 verschiedene Schätzmethoden bestimmt wurde: (a) Least-Square, (b) Ridge Regression mit Parameter λ_2 und (c) Lasso Regression mit Parameter λ_1.

(a) Ordnen Sie den Kurven in der Abbildung die Regressionsmodelle zu.

(b) Geben Sie die zugehörigen Werte von λ_1 und λ_2 an.

Begründen Sie Ihre Antworten. Tipp: Nutzen Sie die Zwischenergebnisse aus Aufgabe 1.

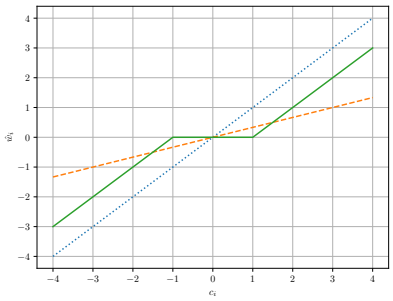

Abbildung 1:

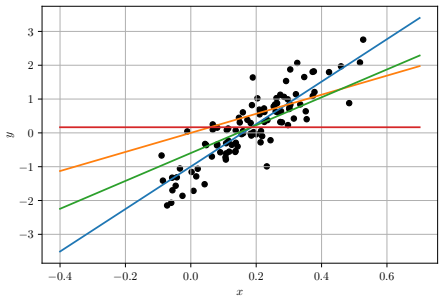

Abbildung 2:

Aufgabe 3. *Ridge / Lasso Regression – Regressionsgeraden* 4 P.

Gegeben seien skalare Daten $x_i \in \mathbb{R}$ und skalare Zielgrößen $y_i \in \mathbb{R}$. Wir bestimmen eine Regressionsgerade mittels (a) Least-Squares Regression, (b) Least-Squares Regression ohne konstanten Term (intercept), (c) Ridge Regression und (d) Lasso Regression. Die resultierenden Regressionsgeraden sind in Abb. 2 dargestellt. Ordnen Sie die Regressionsgeraden den Modellen zu. Begründen Sie Ihre Wahl.

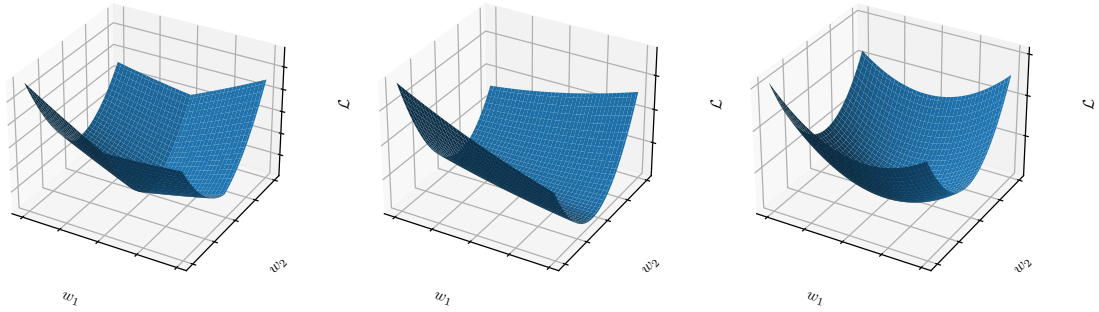

Abbildung 3:

Aufgabe 4. *Ridge / Lasso Regression – Lossgraphen* 3 P.

Gegeben seien zwei-dimensionale Daten $\mathbf{x}_i \in \mathbb{R}^2$ und skalare Zielgrößen $y_i \in \mathbb{R}$. Wir fitten ein lineares Regressionsmodell ohne konstanten Term mittels (a) Least-Squares Regression, (b) Ridge Regression und (c) Lasso Regression. Abb. 3 zeigt die Funktionsgraphen des Trainingslosses der drei Modelle, d.h. die RSS plus gegebenenfalls einen regularisierenden Strafterm. Ordnen Sie die Modelle den Funktionsgraphen zu. Begründen Sie Ihre Wahl.

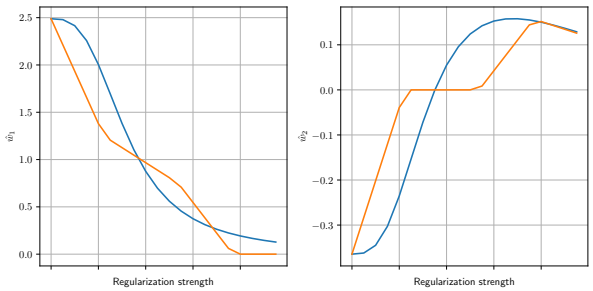

Abbildung 4:

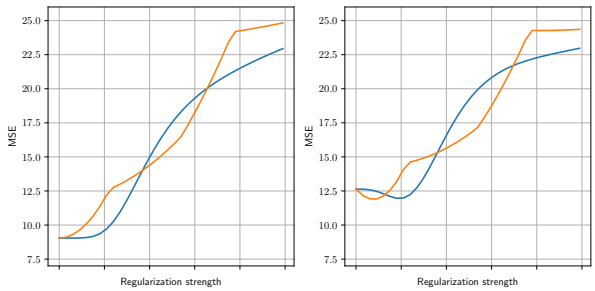

Abbildung 5:

Aufgabe 5. *Ridge / Lasso Regression – Regularisierungsstärke* 4 P.

Gegeben seien zwei-dimensionale Daten $\mathbf{x}_i \in \mathbb{R}^2$ und skalare Zielgrößen $y_i \in \mathbb{R}$. Wir fitten ein lineares Regressionsmodell $y = \mathbf{w}^\top \mathbf{x}$ ohne konstanten Term mittels (a) Ridge Regression und (b) Lasso Regression für verschiedene Regularisierungsstärken $\lambda > 0$.

(a) Abb. 4 zeigt die Koordinaten des geschätzen $\hat{\mathbf{w}}$. Ordnen Sie die Regressionsmodelle den Kurven zu.

(b) Abb. 3 zeigt den jeweiligen MSE auf (a) den Trainingsdaten und (b) Evaluierungsdaten. Ordnen Sie zu, welcher Plot zu den Trainings- bzw. Evaluierungsdaten gehört

Begründen Sie Ihre Wahl.