Universität Salzburg — Florian Graf

Machine Learning

Übungsblatt **8** — 28 Punkte

Aufgabe 1. *Ridge- und Lasso Regression* 16 P.

Gegeben sei eine Stichprobe $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\} \subset \mathbb{R}^{d\times 1}$. 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. In dieser Aufgabe betrachten wir die regularisierten Modelle Ridge- und Lasso-Regression.

(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 $\lambda > 0$. Bestimmen Sie $\hat{\mathbf{w}}_{\text{Ridge}}$.

Hinweis: Sie haben bereits auf Blatt 3 gezeigt, dass jede positiv definite Matrix invertierbar ist.

(b) Lasso Regression entspricht dem Minimierungsproblem $\hat{\mathbf{w}}_{\text{Lasso}} = \operatorname{argmin}_{\mathbf{w}} f(\mathbf{w})$, wobei $f(\mathbf{w}) = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 + \lambda\|\mathbf{w}\|_1$ und $\lambda > 0$. Der Einfachheit halber fixieren wir den konstanten Term $w_0 = 0$. Bestimmen Sie zunächst $\hat{\mathbf{w}}_{\text{Lasso}}$ in dem Fall von nur einem normierten Merkmal, d.h. $x_i \in \mathbb{R}$ und $\frac{1}{n}\sum_{i=1}^n x_i^2 = 1$. Gehen Sie dabei folgendermaßen vor.

(i) Bestimmen Sie die Ableitung an den Stellen wo f differenzierbar ist.

(ii) Skizzieren Sie die Ableitung $f'(w)$ handschriftlich in einem Koordinatensystem.

(iii) Schließen Sie auf die Form von f. Beschreiben Sie die Form im Allgemeinen und skizzieren Sie f in den folgenden Fällen:

(A) $\hat{w}_{\text{OLS}} > 0$ und $\frac{\lambda}{2} < \hat{w}_{\text{OLS}}$

(B) $\hat{w}_{\text{OLS}} > 0$ und $\frac{\lambda}{2} = \hat{w}_{\text{OLS}}$

(C) $\hat{w}_{\text{OLS}} > 0$ und $\frac{\lambda}{2} > \hat{w}_{\text{OLS}}$

Hierbei ist $\hat{w}_{\text{OLS}}$ der entsprechende ordinary least squares Schätzer auf den Daten $(x_1, y_1), \ldots, (x_n, y_n)$.

(iv) Bestimmen Sie $\hat{\mathbf{w}}_{\text{Lasso}} = \operatorname{argmin}_{\mathbf{w}} f(\mathbf{w})$.

(c) Bestimmen Sie $\hat{\mathbf{w}}_{\text{Lasso}}$ im Fall von d Merkmalen ($\mathbf{x}_i \in \mathbb{R}^d$) unter der Annahme, das die Merkmale $\mathbf{x}_{:,j} \in \mathbb{R}^n$ orthonormal sind.

(d) Wir betrachten erneut den Fall, dass die Merkmale $\mathbf{x}_{:,j} \in \mathbb{R}^n$ orthonormal sind. Wir schätzen $\mathbf{w}$ mittels den folgenden Methoden: (a) Least-Squares, (b) Ridge Regression mit Parameter λ_2 und (c) Lasso Regression mit Parameter λ_1.

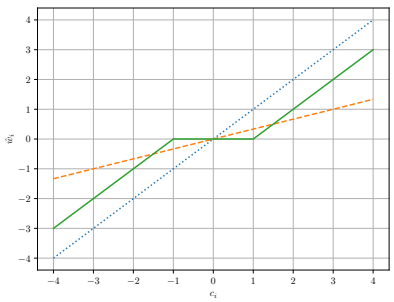

Abbildung 1: $\hat{w}_i$ vs. $c_i = \mathbf{x}_{:,i}^\top \mathbf{y}$.

(i) Ordnen Sie den Kurven in Abbildung 1 die Regressionsmodelle zu. Begründen Sie Ihre Wahl.

(ii) Bestimmen Sie die zugehörigen Werte der Regularisierungsstärken λ_1 und λ_2.

Aufgabe 2. *Erwartungswert und Varianz der Regressionskoeffizienten* 12 P.

Wir betrachten Daten $(\mathbf{x}_i, y_i)_{i=1}^n$ mit $\mathbf{x}_i \in \mathbb{R}^d$ und $y_i \in \{-1, 1\}$, die tatsächlich durch ein lineares Modell generiert werden, d.h. $y_i = \mathbf{w}^\top \mathbf{x}_i + \varepsilon_i$. Hierbei sind die $\mathbf{x}_i$ deterministisch und die $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ unabhängig normalverteilt.

Es kann als bekannt vorrausgesetzt werden, dass der OLS Schätzer für $\mathbf{w}$ durch die Formel $\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$ gegeben ist, wobei $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^\top$ und $\mathbf{y} = [y_1, \ldots, y_n]^\top$.

Wir betrachten nun $\hat{\mathbf{w}}$ als Zufallsvariable (die durch das Rauschen ε_i beeinflusst wird).

(a) Berechnen Sie den Bias des OLS Schätzers, also dass $\mathbb{E}[\hat{\mathbf{w}}] - \mathbf{w}$.

(b) Bestimmen Sie die Kovarianzmatrix $\Sigma(\hat{\mathbf{w}}) = \mathbb{E}[(\hat{\mathbf{w}} - \mathbb{E}[\hat{\mathbf{w}}])(\hat{\mathbf{w}} - \mathbb{E}[\hat{\mathbf{w}}])^\top]$ des OLS Schätzers $\hat{\mathbf{w}}$. Geben Sie außerdem die Spur der Kovarianzmatrix in Abhängigkeit der Eigenwerte von $\mathbf{X}^\top \mathbf{X}$ an.

(c) Wiederholen Sie die Aufgaben (a) und (b) für den Ridge Schätzer.

(d) Interpretieren Sie die Ergebnisse aus den Aufgaben (a) bis (c). Gehen Sie insbesondere auf die Rolle der Regularisierung ein.