Universität Salzburg		Florian Graf
	Machine Learning	
Übungsblatt **9**		25 Punkte

Aufgabe 1. *Ridge- und Lasso Regression* 15 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) Lasso Regression entspricht dem Minimierungsproblem $\hat{\mathbf{w}}_{\text{Lasso}} = \operatorname{argmin}_{\mathbf{w}} f(\mathbf{w})$, wobei $f(\mathbf{w}) = \frac{1}{2}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 + \lambda\|\mathbf{w}\|_1$ und $\lambda > 0$.

  (i) Bestimmen Sie den Gradienten an den Stellen wo f differenzierbar ist und drücken Sie ihn in Abhängigkeit von $\hat{\mathbf{w}}_{\text{OLS}}$ aus.

  (ii) Vereinfachen Sie die Formel für den Gradienten unter der Annahme, dass die Merkmale $\mathbf{x}_{:,j} \in \mathbb{R}^d$ orthonormal sind.

  (iii) Skizzieren Sie die partielle Ableitung $\frac{\partial}{\partial w_1} f(w)$ handschriftlich in einem Koordinatensystem.

  (iv) Der Einfachheit halber nehmen wir zusätzlich an, dass unsere Beobachtungen nur ein Merkmal haben und daher $f : \mathbb{R} \to \mathbb{R}$ (mit Offset 0). Schließen Sie von der Ableitung auf die Form des Funktionsgraphen von f. Beschreiben Sie dessen Form im Allgemeinen und skizzieren Sie ihn in den folgenden Fällen. Achten Sie dabei besonders auf die kritischen Punkte von f.

    (A) $\hat{w}_{\text{OLS}} \leq -\lambda$

    (B) $\hat{w}_{\text{OLS}} \in (-\lambda, \lambda)$

    (C) $\hat{w}_{\text{OLS}} \geq \lambda$

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

  (v) Bestimmen Sie $\hat{w}_{\text{Lasso}} = \arg\min f(w)$ unter den bisherigen Annahmen (Offset 0, ein normalisiertes Merkmal). Wie vermuten Sie generalisiert die Formel im Fall von d orthonormalen Merkmalen?

(b) 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 $\hat{\mathbf{w}}_{\text{Ridge}}$ in Abhängigkeit des ordinary-least-squares Schätzer $\hat{\mathbf{w}}_{\text{OLS}}$.

(c) Wir vergleichen nun OLS-, Ridge- und Lasso-Regression. Wir nehmen dazu an, dass die Merkmale orthonormal sind. In Abbildung 1 ist die i-te Koordinate $\hat{w}_i$ des jeweiligen Schätzers $\hat{\mathbf{w}}$ gegen $c_i = \mathbf{x}_{:,i}^\top \mathbf{y}$ dargestellt.

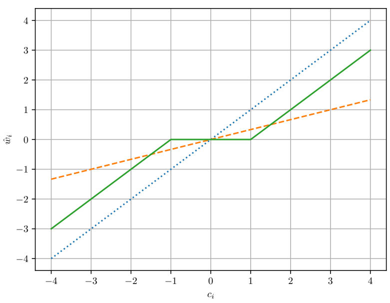

Abbildung 1: $\hat{w}_i$ vs. $c_i = \mathbf{x}_{:,i}^\top \mathbf{y}$.

  (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* 10 P.

Wir betrachten Daten $(\mathbf{x}_i, y_i)_{i=1}^n$ mit $\mathbf{x}_i \in \mathbb{R}^d$ und $y_i \in \mathbb{R}$, die tatsächlich durch ein lineares Modell generiert werden, d.h. $y_i = \mathcal{N}(\mathbf{w}^\top \mathbf{x}_i, \sigma^2)$. Hierbei sind die $\mathbf{x}_i$ deterministisch und die y_i voneinander unabhängig.
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 Erwartungswert $\mathbb{E}[\hat{\mathbf{w}}]$ des OLS Schätzers. Nutzen Sie dazu den Zufallsvektor $\mathbf{y}$.

(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.