Universität Salzburg — Florian Graf

Machine Learning

Übungsblatt **6** — 20 Punkte

Aufgabe 1. *Gradientenabstieg* 8 P.

Es sei $f : \mathbb{R}^d \to \mathbb{R}$ eine (total) differenzierbare Funktion. Die Richtungsableitung $D_{\mathbf{v}} f(p)$ der Funktion f an der Stelle $\mathbf{p} \in \mathbb{R}^d$ in Richtung $\mathbf{v} \in \mathbb{R}^d$ ist folgendermaßen definiert:
Es sei $\gamma : (-\epsilon, \epsilon) \to \mathbb{R}^d$ eine differenzierbare Funktion mit $\gamma(0) = \mathbf{p}$ und $\gamma'(0) = \mathbf{v}$. Dann ist

$$D_{\mathbf{v}} f(p) := \frac{d}{dt} f(\gamma(t))|_{t=0} \ .$$

(a) Erklären Sie die Definition der Richtungsableitung, also die Formel $D_{\mathbf{v}} f(p) := \frac{d}{dt} f(\gamma(t))|_{t=0}$.

Aufgrund der mehrdimensionalen Kettenregel gilt

$$\frac{d}{dt} f(\gamma(t))|_{t=0} = \nabla f(\gamma(0)) \cdot \gamma'(0) \ .$$

Die Richtungsableitung hängt daher nicht von der Wahl der Kurve γ ab.

(b) Die Tangentialvektoren im Punkt $\mathbf{p}$ einer Menge $M \subset \mathbb{R}^d$ sind die Vektoren $\mathbf{v} \in \mathbb{R}^d$, die sich als $\mathbf{v} = \gamma'(0)$ schreiben lassen, wobei $\gamma : (-\epsilon, \epsilon) \to M$ eine Kurve ist die innerhalb von M verläuft und $\gamma(0) = \mathbf{p}$.

Zeigen Sie, dass der Gradient $\nabla f(\mathbf{p})$ orthogonal zu den Tangentialvektoren im Punkt $\mathbf{p}$ der Niveaumenge von f mit Niveau $f(\mathbf{p})$ ist.

(c) Zeigen Sie, dass der Gradient $\nabla f(p)$ in Richtung der maximalen Richtungsableitung zeigt, d.h. zeigen Sie dass

$$\frac{\nabla f(\mathbf{p})}{\|\nabla f(\mathbf{p})\|} = \arg\max_{\|\mathbf{v}\|=1} D_{\mathbf{v}} f(\mathbf{p}) \ .$$

(d) Abbildung 1 zeigt ein Höhenprofil. Zeichnen Sie den ungefähren Verlauf des Gradientenaufstiegsverfahrens (mit kleiner Schrittweite) zur Maximierung der Höhe (oder equivalent des Gradientenabstiegsverfahrens zur Minimierung der negativen Höhe) ausgehend von den schwarz markierten Punkten in die Karte ein. Beschreiben Sie Ihr Vorgehen.

Aufgabe 2. *Softmaxfunktion* 4 P.

Es sei $\sigma : \mathbb{R}^d \to \mathbb{R}^d$ die Softmaxfunktion. Zeigen Sie die folgenden Eigenschaften.

(a) Für alle $\mathbf{x} \in \mathbb{R}^d$ ist die Summe der Komponenten $\sum_{i=1}^d [\sigma(\mathbf{x})]_i = 1$.

(b) Falls $[\mathbf{x}]_i > [\mathbf{x}]_j$ (i-te Komponente größer j-te Komponente), dann ist $[\sigma(\mathbf{x})]_i > [\sigma(\mathbf{x})]_j$, und falls $[\mathbf{x}]_i = [\mathbf{x}]_j$, dann ist $[\sigma(\mathbf{x})]_i = [\sigma(\mathbf{x})]_j$.

(c) Folgern Sie, wieso σ Softmaxfunktion genannt wird.

Oftmals wird die Softmaxfunktion um einen sogenannten Temperaturparameter $t > 0$ erweitert. Dann ist $\sigma(\mathbf{x}; t) := \sigma(t\mathbf{x})$.

(d) Untersuchen Sie das Verhalten von $\sigma(\cdot; t)$ in Abhängigkeit von t.

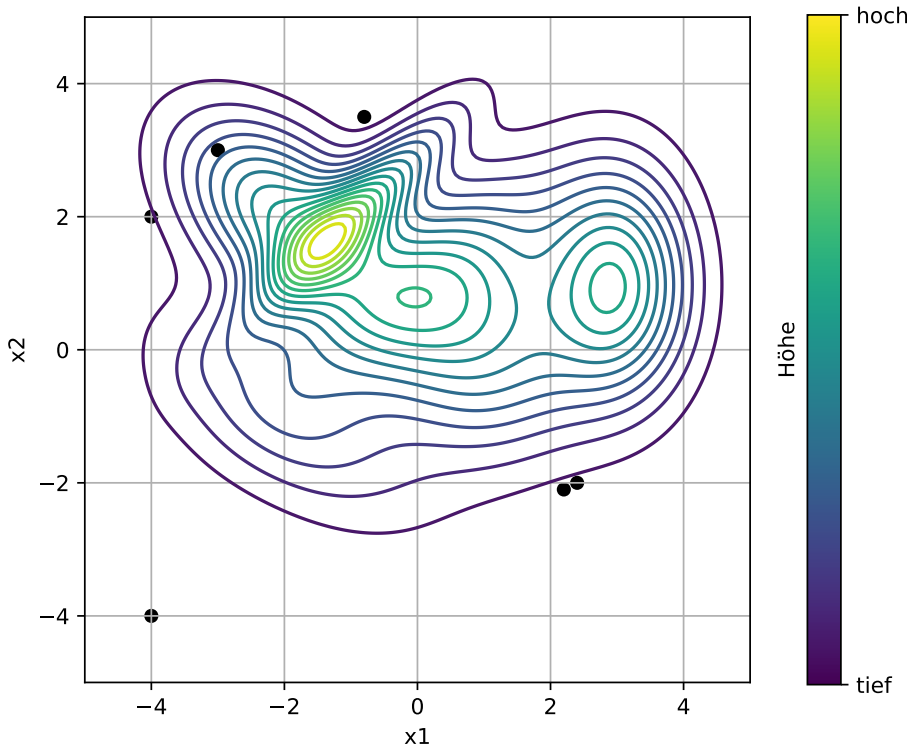

Abbildung 1: Höhenprofil

Aufgabe 3. *Generative und Diskriminative Modelle - I* 8 P.

Wir betrachten ein Gaußsches Diskriminanzanalyse Modell unter der Annahme, dass die Kovarianzmatrizen $\mathbf{\Sigma}_c$ aller Klassen gleich sind, d.h. $\mathbf{\Sigma}_c = \mathbf{\Sigma}\ \forall c$.

(a) Zeigen Sie, dass sich die bedingten Wahrscheinlichkeiten $p(y = c|\mathbf{x}, \boldsymbol{\theta})$ auf die folgende Form bringen lassen:

$$p(y = c|\mathbf{x}, \boldsymbol{\theta}) = a \exp(\mathbf{w}_c^\top \mathbf{x} + b_c) \ ,$$

wobei $a > 0$ nicht von der Klasse c abhängt. Außerdem ist $b_c \in \mathbb{R}$ und $\mathbf{w}_c \in \mathbb{R}^d$.

(b) Folgern Sie, dass $p(y = c|\mathbf{x}, \boldsymbol{\theta})$ sich auch als

$$\begin{pmatrix} p(y = 1|\mathbf{x}, \boldsymbol{\theta}) \\ \vdots \\ p(y = k|\mathbf{x}, \boldsymbol{\theta}) \end{pmatrix} = \text{softmax}(\mathbf{W}\mathbf{x} + \mathbf{b}) \tag{1}$$

schreiben lässt, wobei k die Anzahl der Klassen ist und geben Sie die Matrix $\mathbf{W}$ und den Vektor $\mathbf{b}$ explizit an.

(c) Folgern Sie, dass im Fall von $k = 2$ Klassen die Formel

$$p(y = 1|\mathbf{x}, \boldsymbol{\theta}) = \sigma(\tilde{\mathbf{w}}^\top \mathbf{x} + \tilde{b}) \tag{2}$$

gilt, wobei für $\tilde{\mathbf{w}} \in \mathbb{R}^d$ und $\tilde{b} \in \mathbb{R}$ und σ die Sigmoidfunktion ist.

(d) Vergleichen Sie ausgehend von Gleichung (2) binäre LDA mit binärer logistischer Regression. Gehen Sie dabei auf Gemeinsamkeiten und Unterschiede ein.