Universität Salzburg — Florian Graf

Machine Learning

Übungsblatt **5** — 24 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

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

d.h., die Richtungsableitung ist die Ableitung von f entlang der Kurve γ im Punkt $\mathbf{p}$.

(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.

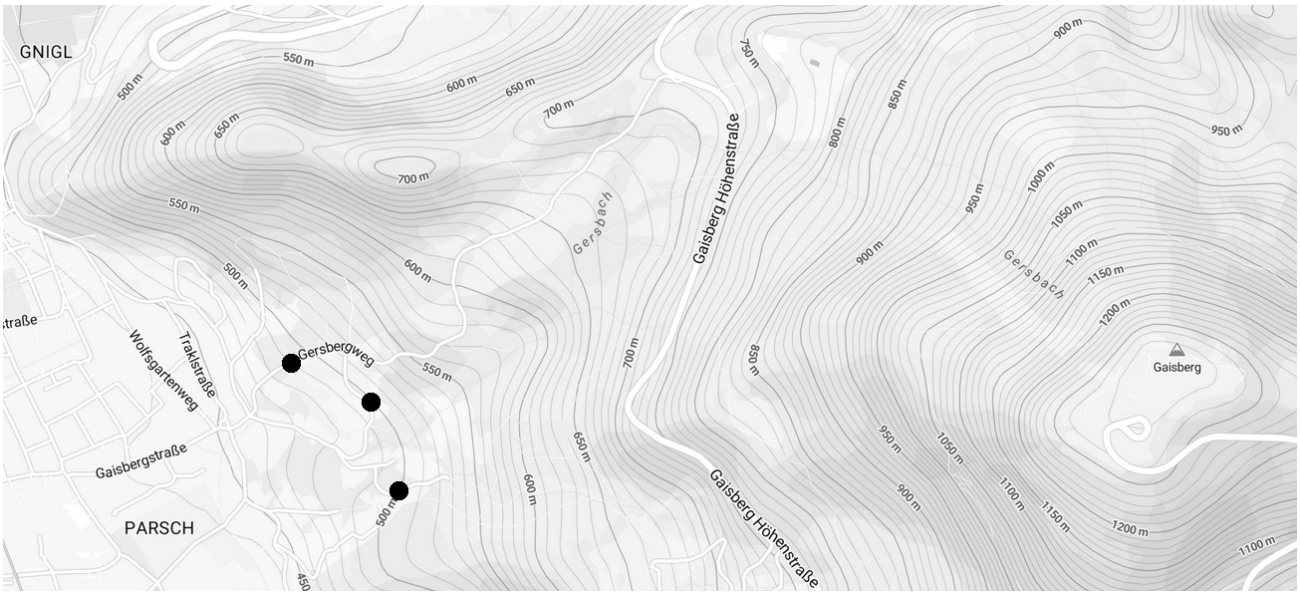
(b) Die Tangentialvektoren im Punkt $\mathbf{p}$ einer Menge $M \subset \mathbb{R}^d$ sind die Vektoren $\mathbf{v} \in \mathbb{R}^d$, die sich als $\mathbf{v} = \gamma'(0)$ schreiben lassen, wobei $\gamma : (-\epsilon, \epsilon) \to M$ eine Kurve ist die innerhalb von M verläuft und $\gamma(0) = \mathbf{p}$.

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) Die nachfolgende Abbildung zeigt das Höhenprofil in der Nähe des Gaisberg. Zeichnen Sie den ungefähren Verlauf des Gradientenabstiegsverfahrens (mit infinitesimaler Schrittweite) zur Maximierung der Höhe (also zur Minimierung der negativen Höhe) ausgehend von den drei schwarz markierten Punkten in die Karte ein.



Aufgabe 2. *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:

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

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) Wir betrachten nun ein QDA Modell mit $k = 2$ Klassen. Drücken Sie erneut $p(y = 1|\mathbf{x}, \boldsymbol{\theta})$ mithilfe der σ-Funktion aus.

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

Wir fitten die folgenden binären Klassifizierungsmodelle M mittels der Maximum Likelihood Methode an die Daten $(x_1, y_1), \ldots, (x_n, y_n)$, wobei $x_i \in \mathbb{R}^d$ und $y_i \in \{0, 1\}$.

- **GaußI:** Ein generatives Klasifizierungsmodell, wobei die bedingte Klassenverteilungen isotrop Gaußsch, d.h. $p(\mathbf{x}|y = c, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_c, \mathbf{I})$. Außerdem sei $p(y)$ gleichverteilt.
- **GaußX:** Wie GaußI, aber die Kovarianzmatrizen sind lernbar, also $p(\mathbf{x}|y = c, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_c, \mathbf{\Sigma}_c)$.
- **LinLog:** Ein logistisches Regressionsmodell mit linearen Features.
- **QuadLog:** Ein logistisches Regressionsmodell mit linearen und quadratischen Features.

Wir evaluieren die gefitteten Modelle $M(\hat{\boldsymbol{\theta}})$ indem wir die folgende Größe auf den (Trainings-)daten berechnen:

$$L(M) = \frac{1}{n} \sum_{i=1}^{n} \log p(y_i|\mathbf{x}_i, M(\hat{\boldsymbol{\theta}}))$$

(a) Geben Sie für jedes der folgenden Modellpaare an, ob $L(M) \leq L(M')$, $L(M) \geq L(M')$, oder ob keine solche Aussage getroffen werden kann (letzteres heißt, dass M je nach Auswahl der Trainingsdaten besser oder schlechter als M' sein kann). Begründen Sie ihre Antwort.

(i) GaussI, LinLog
(ii) GaussX, QuadLog
(iii) LinLog, QuadLog
(iv) GaussI, QuadLog

(b) Anstelle von L nutzen wir nun die Klassifizierungsgenauigkeit

$$R(M) = \frac{1}{n} \sum_{i=1}^{n} 1_{y_i \neq \hat{y}(\mathbf{x}_i)} \ ,$$

um die Modelle zu vergleichen. Hierbei ist $\hat{y}(x_i)$ die durch das Modell vorhergesagte Klasse für x_i.

Gilt im Falle von $L(M) > L(M')$ auch dass $R(M) < R(M')$?

(c) Wir ergänzen nun die Trainingsdaten um Beobachtungen $(x_{n+1}, y_{n+1}), \ldots, (x_{n+m}, y_{n+m})$, die alle einer zusätzlichen dritten Klasse $y_{n+1}, \ldots, y_{n+m} = 3$ angehören. Welche Schritte sind notwendig, um die Modelle GaussI, GaussX, LinLog und QuadLog auf die neuen Daten zu erweitern?