Universität Salzburg		Florian Graf
	Machine Learning	
Übungsblatt **4**		20 Punkte

Aufgabe 1. *Logistic Regression vs. LDA/QDA* 10 P.

Gegeben sind die folgenden binären Klassifizierungsmodelle, die durch das Maximieren der jeweiligen Likelihoodfunktion trainiert wurden.

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

Nach Trainingsende berechnen wir die Leistung jedes Modells M auf der Trainingsmenge wie folgt:

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

(Dies ist die *bedingte* log Likelihood $p(y|\mathbf{x}, \hat{\boldsymbol{\theta}})$ und nicht die gemeinsame log Likelihood $\log p(y, \mathbf{x}|\hat{\boldsymbol{\theta}})$.)
Vergleichen Sie die Leistung der Modelle untereinander mithilfe von L. Nutzen Sie hierfür die Notation $L(M) \le L(M')$, wenn das Modell M niedrigere (oder gleiche) log Likelihood als M' auf den Trainingsdaten hat (für jede beleibige Trainingsmenge).
Geben Sie für jedes der folgenden Modellpaare an, ob $L(M) \le L(M')$, $L(M) \ge L(M')$, oder ob keine solche Aussage getroffen werden kann (d.h. M kann je nach Trainingsdaten besser oder schlechter als M' sein.) und begründen Sie ihre Antwort.

(a) GaussI, LinLog

(b) GaussX, QuadLog

(c) LinLog, QuadLog

(d) GaussI, QuadLog

(e) Anstatt mit der log Likelihood, können wir auch die Leistung mittels der Klassifizierungsgenauigkeit vergleichen:

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

wobei $\hat{y}(x_i)$ die durch das Modell vorhergesagte Klasse für x_i ist. Stimmt es, dass $L(M) > L(M')$ implizert dass $R(M) < R(M')$. Begründen Sie warum, bzw. warum nicht.

Aufgabe 2. *Grundlagen – Mehrdimensionale Diffentialrechnung* 10 P.

Es sei $f : \mathbb{R}^d \to \mathbb{R}$ eine (total) differenzierbare Funktion.

(a) Definieren Sie die i-te partielle Ableitung $\frac{\partial f}{\partial x_i}(p)$ von f an der Stelle $p \in \mathbb{R}^d$ und den Gradienten $\nabla f(p)$.

Zusätzlich sei $\gamma : (-\epsilon, \epsilon) \to U \subset \mathbb{R}^d$ eine differenzierbare Funktion mit $\gamma(0) = p$. Die Ableitung $\gamma'(0) =: v \in \mathbb{R}^d$ wird der Tangentialvektor von γ im Punkt p genannt. Die Richtungsableitung von f an der Stelle p in Richtung v ist definiert als $D_v f(p) := \frac{d}{dt} f(\gamma(t))|_{t=0}$.

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

(c) Zeigen Sie mithilfe der mehrdimensionalen Kettenregel, dass

$$D_v f(p) = \nabla f(p) \cdot v$$

und somit, dass die Richtungsableitung unabhängig von der Kurve γ ist (solange $\gamma(0) = p$ und $\gamma'(0) = v$)

(d) Folgern sie aus (c), dass die partielle Ableitung $\frac{\partial f}{\partial x_i}$ die Richtungsableitung von f in Richtung des i-ten standard Basisvektors $\mathbf{e}_i$ ist.

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

$$\nabla f(p) = \arg\max_{v \in \mathbb{R}^d} \frac{D_v f(p)}{\|v\|} \quad .$$

(f) Erklären Sie mithilfe von (e) die Update Regel des Gradient Descent Verfahrens.