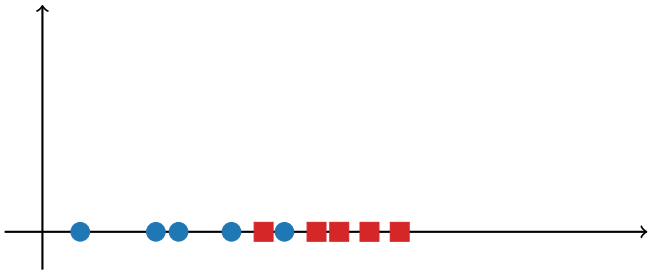
Universität Salzburg — Florian Graf

Machine Learning

Übungsblatt **10** — 30 Punkte

Aufgabe 1. 3 P.

Die nachfolgende Grafik zeigt ein 1-dimensionales binäres Klassifizierungsproblem mit Labels $c \in \{●, ■\}$. in 1D. Wir wählen die Gaußsche Diskriminanzanalyse (GDA) als Ansatz.



(a) Zeichnen Sie grob die Dichten der bedingten Verteilungen $p(y = c|x)$ ein.

(b) Wie wird ein neuer Punkt $x' \in \mathbb{R}$ einer Klasse zugewiesen?

(c) Wo liegt die Entscheidungsgrenze? Markieren Sie diese.

Aufgabe 2. 2 P.

Sowohl bei logistischer Regression, als auch bei klassischer linearer Regression haben wir ein Modell für die jeweilige Zielgröße y, gegeben den Einflussgrößen $\mathbf{x}$ und den Modellparametern $\boldsymbol{\theta}$.

(a) Schreiben Sie die jeweiligen Modelle an, also $p(y|\mathbf{x}, \boldsymbol{\theta})$.

(b) Wie schätzen wir üblicherweise die Parameter $\boldsymbol{\theta}$ der jeweiligen Modelle bei gegebenem Datensatz $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$?

Aufgabe 3. 3 P.

Wie unterscheidet sich Maximum Likelihood Schätzung von Maximum A-Posteriori Schätzung? Erklären Sie die grundlegenden Ideen und deren Unterschied an einem Beispiel, z.B. im Kontext linearer Regression.

Aufgabe 4. 3 P.

Gegeben seien Beobachtungen $x_1, \ldots, x_n \in \mathbb{R}$. Wir nehmen an, dass die Daten eine unabhängige Stichprobe einer Gleichverteilung $\mathcal{U}_{[c-r,c+r]}$ auf dem Interval $[c - r, c + r]$ sind.

(a) Bestimmen Sie die Parameter $r > 0$ und $c \in \mathbb{R}$ mittels der Maximum-Likelihood Methode.

(b) Bestimmen Sie die Parameter $r > 0$ und $c \in \mathbb{R}$ mittels MAP Schätzung unter der Annahme eines Laplace Priors $p(r) = \text{Lap}(0, \frac{1}{\lambda})$ mit $\lambda > 0$.

Aufgabe 5. 2 P.

Betrachten Sie folgenden Ausschnitt eines neuronalen Netzes mit $\mathbf{x} \in \mathbb{R}^d$ als Input und kompatiblen Gewichtsmatrizen:

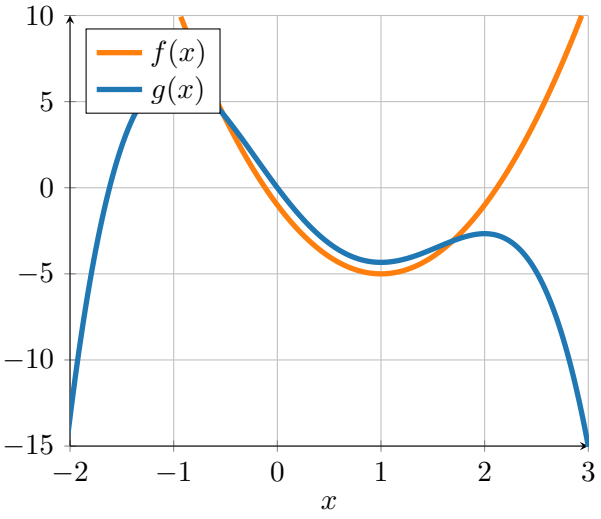
$$\begin{aligned} \mathbf{z} &= \mathbf{W}_1\mathbf{x} + \mathbf{b}_1 \\ \mathbf{h} &= \text{ReLU}(\mathbf{z}) \\ \mathbf{y} &= \mathbf{x} + \mathbf{W}_2\mathbf{h} \end{aligned}$$

(a) Wie könnten Sie die Gewichte in diesem Abschnitt des Netzes setzen, um die Identitätsfunktion $\mathbf{x} \mapsto \mathbf{x}$ zu realisieren?

(b) In der Praxis setzen wir die Gewichtsmatrizen nicht händisch, sondern mittels Minimierung einer Fehlerfunktion. Ist es trotzdem möglich zu erreichen, dass $\mathbf{y} \approx \mathbf{x}$. Falls ja, wie genau; falls nein, wieso nicht?

Aufgabe 6. 2 P.

Wir betrachten die Funktionen $f(x) = (x-1)^2$ und $g(x) = -x^4 + \frac{8x^3}{3} + 2x^2 - 8x$, die in der folgenden Abbildung dargestellt ist. Wie verhält sich das Gradientenabstieg-Verfahren mit konstanter Schrittweite und



Startwert 0 in Abhängigkeit des Schrittweite Parameters.

Aufgabe 7. 3 P.

Nehmen Sie an, Sie haben eine Funktion $F : \mathbb{R}^4 \to \mathbb{R}^2$, $\mathbf{x} \mapsto C(B(A(\mathbf{x})))$, also eine Komposition der Funktionen A, B, C. Zur Einfachheit nehmen wir an, dass $A, B : \mathbb{R}^4 \to \mathbb{R}^4$ und $C : \mathbb{R}^4 \to \mathbb{R}^2$ und nennen die Zwischenergebnisse $\mathbf{c} = C(\mathbf{b})$, $\mathbf{b} = B(\mathbf{a})$ und $\mathbf{a} = A(\mathbf{x})$.

(a) Schreiben Sie die Definition der Jacobimatrix $\mathbf{J}$ der Funktion F an.

(b) Schreiben Sie $\mathbf{J}$ über die Kettenregel an.

(c) Wenn wir $\mathbf{J}$ zeilenweise berechnen wollten, müssten wir nacheinander mit welchen Vektoren $\mathbf{v}$ multiplizieren? Wie sieht dies aus, wenn $\mathbf{J}$ spaltenweise berechnet werden soll?

(d) Wie nennt man diese beiden Strategien? Welche der beiden Strategien ist in unserem Fall günstiger?

Aufgabe 8. 2 P.

Gegeben seien reellwertige Daten $\{(x_1, y_1), \ldots, (x_n, y_n)\} \subset \mathbb{R} \times \mathbb{R}$. Es gilt, dass $y_i = p(x_i) + \epsilon_i$, wobei $\epsilon_i \sim \mathcal{N}(0, 1)$ und p ein Polynom vom Grad 5 ist. Dies ist uns aber unbekannt. Wir haben insgesamt 300 Datenpaare zur Verfügung, von denen wir die zufällig N Paare zum fitten eines linearen Regressionsmodel mit polynomiellen Merkmalen verwenden, und die übrigen $300 - N$ zur Validierung verwenden.

(a) Wie werden sich (im Mittel) der Trainings- und Validierungsfehler verhalten, wenn wir N erhöhen?

(b) Da wir den Grad des Polynoms p nicht kennen, fitten wir 3 Regressionsmodelle, mit polynomen Merkmalen der Ordnung 2, 4 und 6. Hierbei bedeutetder Ordnung k, dass Merkmale $(1, x^1, \ldots, x^k)$ verwendet werden. Welches der Regressionsmodelle wird für $N = 150$ voraussichtlich den kleinsten Validierungsfehler haben.

Aufgabe 9. 10 P.

Präsentieren Sie die Lösung einer Aufgabe im Proseminar. Sie können sich als Gruppe für eine Aufgabe melden. Welche Person innerhalb der Gruppe präsentiert wird von mir ausgewählt.