Wahrscheinlichkeitstheorie und Statistik

Anwendungsbeispiele und Prüfungsvorbereitung

Aufgabe 13-1 (Batching) Betrachten Sie den Machine-Learning Algorithmus zum Erkennen handgeschriebener Ziffern in [1] und beantworten Sie folgende Fragen:

(a) Wie lautet die Verlustfunktion?

(b) Welcher Optimierer wird zum Minimieren der Verlustfunktion eingesetzt?

(c) Welche Gradienten verwendet das Optimierungsverfahren und sind diese zufällig oder deterministisch?

Lösung:

(a) Die Verlustfunktion L lautet `sparse_categorical_crossentropy`. Dies ist die negative Log-Likelihood für die Verteilung von Zufallsvariablen mit Werten in einer endlichen Menge $\{1, \dots, m\}$.

(b) Der Optimierer heisst ADAM; das Kürzel ist von Adaptive Moment Estimation abgeleitet. Es handelt sich um eine stochastisches Gradientabstiegsverfahren mit Normalisierung des ersten und zweiten Moments der stochastischen Gradienten.

(c) Die Gradienten lauten $\sum_{i=1}^{32} \nabla L(f_\theta(X_i), Y_i)$, wobei f_θ das zu trainierende neuronale Netz ist und $(X_1, Y_1), \dots, (X_{32}, Y_{32})$ eine Stichprobe aus dem gegebenen Datensatz. Diese Stichproben werden Batches genannt. Durch zufällige Wahl der Stichprobe erhält man zufällige Gradienten. In der Praxis durchlaufen die Batches sequentiell den gesamten Datensatz.

MC 13-2 (Universalität) Sei F eine universelle Menge von neuronalen Netzen $f : \mathbb{R}^d \to \mathbb{R}$ und $g : \mathbb{R}^d \to \mathbb{R}$ eine stetige Funktion. Welche Aussagen gelten für alle $\epsilon > 0$? (Mehrere richtige Antworten sind möglich.)

(a) $\forall n \in \mathbb{N}, \forall x_1, \dots, x_n \in \mathbb{R}^d, \exists f \in F, \forall i \in \{1, \dots, n\} : |f(x_i) - g(x_i)| < \epsilon$.

(b) $\exists f \in F, \forall n \in \mathbb{N}, \forall x_1, \dots, x_n \in \mathbb{R}^d, \forall i \in \{1, \dots, n\} : |f(x_i) - g(x_i)| < \epsilon$.

(c) $\forall n \in \mathbb{N}, \exists f \in F, \forall x \in [-n, n]^d : |f(x) - g(x)| < \epsilon$.

(d) $\exists f \in F, \forall n \in \mathbb{N}, \forall x \in [-n, n]^d : |f(x) - g(x)| < \epsilon$.

Lösung: Nur (a) und (c) sind richtig.

(a) Richtig. Jede endliche Menge $\{x_1, \dots, x_n\}$ ist kompakt, und g kann auf kompakten Mengen uniform durch Netze in F approximiert werden.

(b) Falsch. Gegenbeispiel: F ist die Menge der flachen ReLU Netzwerke und $g(x) = x^2$. Dann ist kein Netz in F uniform nahe an g, weil die Netze in F alle stückweise linear sind, g hingegen quadratisch, und somit $|f(x) - g(x)| \to \infty$ für $x \to \infty$.

(c) Richtig. Die Menge $[-n, n]^d$ ist kompakt, und g kann auf kompakten Mengen uniform durch Netze in F approximiert werden.

(d) Falsch, mit demselben Gegenbeispiel wie in (b).

MC 13-3 (Approximationsfehler). Sei F eine universelle Menge von neuronalen Netzen $f : \mathbb{R}^d \to \mathbb{R}$. Welche Aussagen gelten für den empirischen Verlust

$$L(f) = \sum_{i=1}^{n} \big(f(X_i(\omega)) - Y_i(\omega)\big)^2$$

bezüglich eines Datensatzes von unabhängigen, identisch verteilten Beobachtungen (X_i, Y_i), $i \in \{1, \dots, n\}$? (Mehrere richtige Antworten sind möglich.)

(a) $\inf_{f\in F} L(f) = \inf_{f\in C(\mathbb{R}^d,\mathbb{R})} L(f)$

(b) $\arg\min_{f\in F} L(f) \subseteq \arg\min_{f\in C(\mathbb{R}^d,\mathbb{R})} L(f)$

(c) $\inf_{f\in F} L(f) = \inf_{f\in F} \mathbb{E}(f(X_1) - Y_1)^2)$

(a) Richtig. Für jeden Fall ω betrachten wir auf $C(\mathbb{R}^d, \mathbb{R})$ die Topologie der uniformen Konvergenz auf $\{X_1(\omega), \dots, X_n(\omega)\}$: Bezüglich dieser Topologie ist L stetig in f und F dicht in $C(\mathbb{R}^d, \mathbb{R})$.

(b) Richtig. Sei $f^* \in \arg\min_{f\in F} L(f)$. Dann gilt $L(f^*) = \inf_{f\in F} L(f) = \inf_{f\in C(\mathbb{R}^d,\mathbb{R})} L(f)$ dank (a), also $f^* \in \arg\min_{f\in C(\mathbb{R}^d,\mathbb{R})} L(f)$.

(c) Falsch. Gegenbeispiel: Sei F die Menge der flachen ReLU Netze in Dimension $d = 1$. Dann gilt $\inf_{f\in F} L(f) = 0$ per Konstruktion in Aufgabe 12-2. Seien Y_i unabhängig von X_i mit $\mathbb{E}(Y_i) = 0$ und $\mathbb{V}\mathrm{ar}(Y_i) > 0$. Dann gilt $\arg\min_{f\in C(\mathbb{R},\mathbb{R})} \mathbb{E}(L(f)) = \{0\}$ und $\inf_{f\in C(\mathbb{R},\mathbb{R})} L(f) = L(0) = \mathbb{V}\mathrm{ar}(Y_1) > 0$. Also gilt $\inf_{f\in F} L(f) \neq \inf_{f\in C(\mathbb{R},\mathbb{R})} L(f)$.

MC 13-4 (Machine Learning als inverses Problem). Wir betrachten den empirischen Verlust eines ReLU Netzes auf einem Datensatz von Paaren reeller Zahlen (x_i, y_i):

$$L(\theta) = \frac{1}{n}\sum_{i=1}^{n} \big(f_\theta(x_i) - y_i\big)^2, \qquad \theta = (w, a, b) \in \mathbb{R}^m \times \mathbb{R}^m \times \mathbb{R}^m, \qquad f_\theta(x) = \sum_{i=1}^{m} w_i \max\{0, a_i x + b_i\}.$$

Welche Aussagen sind für hinreichend grosses m korrekt? (Mehrere richtige Antworten sind möglich.)

(a) Es gibt genau ein $\theta^* \in \arg\min_\theta L(\theta)$.

(b) Es gibt genau ein $\theta^* \in \arg\min_\theta L(\theta)$ und θ^* hängt stetig von den Daten (x_i, y_i) ab.

(c) $\inf_\theta L(\theta)$ hängt stetig von den Daten (x_i, y_i) ab.

(d) $\inf_\theta L(\theta) = 0$.

Lösung: (a) und (b) sind falsch, (c) und (d) sind richtig.

(d) Richtig. Für $m \geq 3n$ gilt wegen Aufgabe 12-2 $\inf_\theta L(\theta) = 0$ und das Infimum wird angenommen.

(c) Richtig, dank (d).

(a) Falsch. Wie in (d) argumentiert, existiert ein Minimierer $\theta^* = (w^*, a^*, b^*)$. Falls alle (w_i, a_i, b_i) gleich sind, erhält man durch weitere Minimierer durch Anpassen von w_i für fixe a_i und b_i. Andernfalls erhält man weitere Minimierer durch Permutation in i.

(b) Falsch, wegen (a).

MC 13-5 (Lösungsbegriff Bayesianische Optimierung). Auf einem Wahrscheinlichkeitsraum $(\Theta, \mathcal{A}, \pi_0)$ betrachten wir eine messbare Funktion $L : \Theta \to \mathbb{R}$ mit Minimierer $\theta^* \in \arg\min_\theta L(\theta)$ und Minimum $m = L(\theta^*)$. Welche der folgenden Aussagen treffen auf π_n-verteilte Zufallsvariablen θ_n zu, wobei

$$\pi_n(d\theta) \propto \exp(-nL(\theta)\pi_0(d\theta), \qquad n \in \mathbb{N}.$$

(Mehrere richtige Antworten sind möglich.)

(a) $L(\theta_n) \to m$ in Wahrscheinlichkeit.

(b) $L(\theta_n) \to 0$ in Wahrscheinlichkeit.

(c) $\theta_n \to \theta^*$ in Wahrscheinlichkeit.

(d) $\theta_n \to \theta^*$ in Wahrscheinlichkeit, falls L stetig ist.

(a) Richtig, dank Aufgabe 11.1 mit $A := \{\theta \in \Theta : L(\theta) > m + \delta\}$, für beliebiges $\delta > 0$.

(b) Falsch, außer $m = 0$, was im Allgemeinen nicht der Fall ist.

(c) Falsch. Gegenbeispiel: $L(\theta) = (\theta^2 - 1)^2$, $\pi_0(d\theta) = \mathbb{1}_{[-1,1]}/2$, $\arg\min_\theta L(\theta) = \{-1, 1\}$, $\pi_n \to (\delta_{-1} + \delta_1)/2$ aus Symmetrie.

(d) Falsch, selbes Gegenbeispiel.

MC 13-6 (Invarianz unter der Langevin-Dynamik). Sei θ_0 eine $\mathbb{R}^d$-wertige Zufallsvariable mit Dichtefunktion proportional zu $\exp(-L(\theta)/\epsilon)$, für eine messbare Funktion $L : \mathbb{R}^d \to \mathbb{R}$ und $\epsilon > 0$. Ferner sei W unabhängig von θ_0 und standardnormalverteilt. Unter welcher Dynamik ist π infinitesimal invariant im Sinn von

$$\frac{d}{dt}\bigg|_{t=0} \mathbb{E}\Big(f\big(\theta(t)\big)\Big) = 0, \qquad \text{für alle Testfunktionen } f.$$

(Mehrere richtige Antworten sind möglich.)

(a) $\dot\theta(t) = -\nabla L(\theta(t))$, $\theta(0) = \theta_0$

(b) $\dot\theta(t) = -\nabla L(\theta(t)) + \sqrt{2t/\epsilon}W$, $\theta(0) = \theta_0$

(c) $\theta(t) = \theta_0 - t\nabla L(\theta_0)$

(d) $\theta(t) = \theta_0 - t\nabla L(\theta(0)) + \sqrt{2t/\epsilon}W$

(a) Falsch.

- Intuition: Der Gradientenabstieg ist nur an Nullstellen von ∇L konstant, aber π ist nicht nur an Nullstellen von ∇L konzentriert, ausser L ist konstant.
- Formales: Falls L eine nicht-konstante Testfunktion ist, kann man in der Invarianzbedingung $f = L$ wählen und erhält

$$\frac{d}{dt}\Big|_{t=0} \mathbb{E}\Big(L\big(\theta(t)\big)\Big) = \int \big\|\nabla L(\theta)\big\|^2 \exp\big(-L(\theta)/\epsilon\big)\, d\theta > 0.$$

(b) Richtig, wie in der Vorlesung gezeigt.

(c) Falsch, weil die Ableitungen $\dot\theta(0)$ in (a) und (c) übereinstimmen.

(d) Richtig, weil die Ableitungen $\dot\theta(0)$ in (b) und (d) übereinstimmen.

MC 13-7 (Gradientenfluss). Sei $L : \mathbb{R}^d \to \mathbb{R}$ eine glatte Funktion mit einem eindeutigen globalen Minimum θ^* und sei $\theta_0 \in \mathbb{R}^d$. Welche Aussagen sind im Allgemeinen korrekt? (Mehrere richtige Antworten sind möglich.)

(a) Wenn L ein quadratisches Polynom ist, konvergiert der Gradientenfluss gegen θ^*.

(b) Wenn L kein quadratisches Polynom ist, kann der Gradientenfluss divergieren.

(c) Wenn L kein quadratisches Polynom ist, kann der Gradientenfluss gegen $\theta \neq \theta^*$ konvergieren.

(d) Der empirische Verlust eines neuronalen Netzes ist im Allgemeinen kein quadratisches Polynom.

Lösung: Alle Aussagen sind korrekt.

(a) Richtig. Bis auf eine affine Transformation und eine additive Konstante ist die Verlustfunktion von der Form $L(\theta) = \|\theta\|^2/2$ und der Gradientenfluss $\theta(t) = e^{-t}\theta_0$ konvergiert gegen das eindeutige Minimum $\theta^* = 0$.

(b) Richtig, z.B. wenn $L(\theta) = \exp(-\theta)$.

(c) Richtig, z.B. wenn $L(\theta) = (\theta^2 - 1)^2$ und $\theta_0 = 0$.

(d) Richtig, z.B. bei Netzen mit nicht-polynomieller Aktivierungsfunktion und generischen Koeffizienten.

MC 13-8 (Aktivierungsfunktion). Welche der folgenden Funktionen Aktivierungsfunktionen liefern universale neuronale Netze? (Mehrere richtige Antworten sind möglich.)

(a) $f(x) = \max(0, x)$

(b) $f(x) = -\max(0, x)$

(c) $f(x) = \max(0, x^2)$

(d) $f(x) = \max(x/100, x)$

(e) $f(x) = x$

(f) $f(x) = x^2$

Lösung: Alle Aussagen sind korrekt.

(a) Richtig: kein Polynom.

(b) Richtig: kein Polynom.

(c) Richtig: kein Polynom.

(d) Richtig: kein Polynom.

(e) Falsch: ein Polynom.

(f) Falsch: ein Polynom.

Aufgabe 13-9 (Tests für Mittelwert und Standardabweichung). Zwei verschiedene Methoden wurden verwendet, um die Schmelzenthalpie von Wasser zu messen. Methode A lieferte bei 10 Messungen eine durchschnittliche Schmelzenthalpie von 6.0 kJ/mol bei einer Standardabweichung von 0.2 kJ/mol. Methode B lieferte bei 5 Messungen eine durchschnittliche Schmelzenthalpie von 5.6 kJ/mol bei einer Standardabweichung von 0.1 kJ/mol. Die Standardabweichungen wurden ohne Besselkorrektur berechnet.

(a) Testen Sie auf einem Signifikanzniveau von 5 %, ob die beiden Methoden dieselbe Standardabweichung haben.

Hinweis: Der Quotient von zwei unabhängigen χ^2-verteilten Zufallsvariablen ist F-verteilt, und die Quantile der F-Verteilung sind in Tabellen oder Statistik-Software verfügbar.

(b) Testen Sie auf demselben Signifikanzniveau, ob sich die Mittelwerte der beiden Methoden signifikant unterscheiden.

Hinweis: Verwenden Sie den t-Test von Welch [2].

Lösung: Wir bezeichnen die Stichprobengrösssen mit $n_A = 10$ und $n_B = 5$, die Mittelwerte mit $\bar{x}_A = 6.0$ und $\bar{x}_B = 5.6$, sowie die Standardabweichungen mit $s_A = 0.2$ und $s_B = 0.1$. Wir nehmen an, dass die Messergebnisse normalverteilt sind, mit Mittelwerten μ_A bzw. μ_B und Standardabweichungen σ_A bzw. σ_B.

(a) **Hypothesen:** $H_0 : \sigma_A^2 = \sigma_B^2 \qquad H_1 : \sigma_A^2 \neq \sigma_B^2$

Teststatistik: $F = \frac{s_A^2}{s_B^2} = \frac{0.2^2}{0{,}1^2} = \frac{0.004}{0.01} = 4$

Verteilung: Unter H_0 hat die Teststatistik eine F-Verteilung mit Freiheitsgraden $df_1 = n_A - 1 = 9$ und $df_2 = n_B - 1 = 4$.

Kritische Werte: Die 2.5 % und 97.5 % Quantile dieser Verteilung sind 0.212 und 8.905.

Testergebnis: H_0 wird nicht verworfen, da die Teststatistik $F = 4$ im Annahmebereich $[0.212, 8.905]$ liegt. Es gibt keinen signifikanten Unterschied zwischen den Standardabweichungen.

(b) **Hypothesen:** $H_0 : \mu_A = \mu_B \qquad H_1 : \mu_A \neq \mu_B$

Teststatistik: Die Test-Statistik des Welch-Tests lautet

$$t = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}} = \frac{6.0 - 5.6}{\sqrt{\frac{0.2^2}{10} + \frac{0.1^2}{5}}} = \frac{0.4}{\sqrt{0.004 + 0.002}} = \frac{0.4}{\sqrt{0.006}} \approx \frac{0.4}{0.077} \approx 5.19.$$

Freiheitsgrade: Die Freiheitsgrade gemäss Welch-Satterthwaite-Näherung sind

$$df \approx \frac{(s_A^2/n_A + s_B^2/n_B)^2}{\frac{(s_A^2/n_A)^2}{n_A - 1} + \frac{(s_B^2/n_B)^2}{n_B - 1}} \approx 10.$$

Kritischer Wert: Das 2.5 % Quantil der t-Verteilung mit den angegebenen Freiheitsgraden lautet $t_{0.025,10} \approx 2.228$.

Testergebnis: Da $t = 5.19 > 2.228$, wird die Nullhypothese verworfen. Die Mittelwerte unterscheiden sich signifikant auf dem 5 %-Niveau.

Aufgabe 13-10 (Parameterschätzung für Exponentialverteilung). Die Lebensdauer T (in Stunden) einer Glühbirne unter extremen Bedingungen wird als exponentialverteilt mit unbekanntem Parameter $\lambda > 0$ angenommen. Eine Stichprobe von 5 Glühbirnen wurde bis zum Ausfall beobachtet. Die gemessenen Lebensdauern lauten $T_1 = 4, \quad T_2 = 4, \quad T_3 = 5, \quad T_4 = 3$. Beantworten Sie die folgenden Fragen:

(a) Schätzen Sie den Parameter λ und die mittlere Lebensdauer.

(b) Schätzen Sie die Überlebenswahrscheinlichkeit $S(3) = \mathbb{P}(T > 3)$.

(c) Wenn eine Glühbirne bereits 2 Stunden funktioniert hat, wie viele Stunden wird sie voraussichtlich noch halten?

(d) Wenn zwei Glühbirnen unabhängig voneinander betrieben werden, wie groß ist die Wahrscheinlichkeit, dass beide länger als 3 Stunden halten?

Lösung:

(a) Der Mittelwert der beobachteten Lebensdauern ist: $\bar{T} = \frac{4+4+5+3}{4} = \frac{16}{4} = 4$. Der Maximum-Likelihood-Schätzer fürλ ist: $\hat{\lambda} = \frac{1}{\bar{T}} = \frac{1}{4} = 0.25$. Die geschätzte mittlere Lebensdauer beträgt $\frac{1}{\hat{\lambda}} = 4$ Stunden.

(b) Die Überlebensfunktion der Exponentialverteilung ist $S(t) = \mathbb{P}(T > t) = e^{-\lambda t}$. Mit $\hat{\lambda} = 0.25$ ergibt sich $\hat{S}(t) = e^{-0.25t}$ und somit $\hat{S}(3) = e^{-0.25 \cdot 3} = e^{-0.75} \approx 0.472$. Die Wahrscheinlichkeit, dass eine Glühbirne länger als 3 Stunden hält, beträgt also ca. 47.2%.

(c) Da die Exponentialverteilung gedächtnislos ist, gilt: $\mathbb{E}[T - 2 \mid T > 2] = \mathbb{E}[T] = \frac{1}{\hat{\lambda}} = 4$. Die erwartete verbleibende Zeit beträgt somit 8 Stunden.

(d) Da die Glühbirnen unabhängig voneinander betrieben werden, gilt $\mathbb{P}(T_1 > 3 \text{ und } T_2 > 3) \approx \hat{S}(3)^2 = (0.472)^2 \approx 0.223$. Die Wahrscheinlichkeit beträgt also etwa 22.3%.

References

[1] Josef Teichmann. *Introduction to Mathematics of New Technologies in Banking and Finance*. Verfügbar unter: `https://gist.github.com/jteichma/ea9452fc6cc307afcc9309741aeed9b7`. Mai 2025.

[2] Wikipedia. *Welch's t-Test*. Verfügbar unter: `https://en.wikipedia.org/wiki/Welch's_t-test`. Mai 2025.