DAV
Deutsche
Aktuarvereinigung e.V.

Schriftliche Prüfung im Grundwissen

Data Science & Künstliche Intelligenz

gemäß Prüfungsordnung 6
der Deutschen Aktuarvereinigung e. V.

Musterklausur

Hinweise:

- Als Hilfsmittel ist ein Taschenrechner zugelassen.
- Die Gesamtpunktzahl beträgt 180 Punkte. Die Klausur ist bestanden, wenn mindestens 90 Punkte erreicht werden.
- Bitte prüfen Sie die Ihnen vorliegende Prüfungsklausur auf Vollständigkeit. Die Klausur besteht aus x Seiten.
- Alle Antworten sind zu begründen und bei Rechenaufgaben muss der Lösungsweg ersichtlich sein.

Mitglieder der Prüfungskommission:

Dr. Zoran Nikolić, Dr. Stefan Nörtemann, Prof. Dr. Anja Bettina Schmiedt, Prof. Dr. Fabian Transchel, Dr. Wiltrud Weidner

Aufgabe 1 [Gesellschaftliches Umfeld und Regulatorik] [45 Punkte]

Die (fiktive) international tätige *Modern Times Insurance Ltd.* (MTI) mit Sitz in Zürich (in der Schweiz) plant, ausschließlich für den deutschen Markt, die Einführung eines neuen innovativen Produktes in der Sparte „Leben".

Die Produktentwicklungsabteilung hat dazu einen Vorschlag für ein „pay-per-use-Produkt" erarbeitet. Nach der Registrierung über die Homepage der MTI kann der Kunde für sich über eine App beliebig viele jeweils 24 Stunden laufende Unfallversicherungen kaufen.

Die gesamte Abwicklung des Geschäfts, insbesondere Vertrieb und Bestandsverwaltung, soll über eine cloud-Lösung direkt in der Firmenzentrale in Zürich erfolgen.

(a) *[10 Punkte]* Erläutern Sie die Rechtssystematik hinsichtlich des Datenschutzes, die in diesem Fallbespiel anzuwenden ist und benennen Sie die Rechtsquellen, die für dieses Fallbeispiel einschlägig sind.

(b) *[8 Punkte]* Ist eine Verarbeitung personenbezogener Daten in diesem Fall erlaubt? Begründen Sie Ihre Antwort und belegen Sie sie mit einer Rechtsquelle.

(c) *[7 Punkte]* Nennen Sie sieben (von neun) zentrale Datenschutzgrundsätze, die in diesem Fallbeispiel anzuwenden sind.

Nach erfolgreicher Einführung möchte das Aktuariat ein Profittesting durchführen, um zu prüfen, ob das neue Angebot profitabel ist. Für die Analyse sollen die erhobenen personenbezogenen Daten der Kunden verwendet werden.

(d) *[8 Punkte]* Nehmen Sie Stellung zur Rechtmäßigkeit des Vorhabens mit Blick auf den Datenschutz.

In der KI-Abteilung der MIT wurde kürzlich eine KI-Tool für das Profittesting entwickelt. Leider wissen Sie nicht, welche Art Lernverfahren dahintersteckt.

(e) *[12 Punkte]* Wägen Sie sorgsam ab und geben Sie eine Einschätzung über die regulatorischen Anforderungen für den Einsatz gemäß der KI-Verordnung.

Lösungsvorschlag:

Aufgabe 1

(a) Die Schweiz ist nicht Mitglied der EU und in der Schweiz gilt die Datenschutzgrundverordnung DSGVO nicht. Dennoch unterliegt die Datenverarbeitung (obwohl in der Schweiz durchgeführt) der DSGVO, da hier Daten von Bürgern der EU verarbeitet werden. („Marktortprinzip"!) Damit ist auch das Bundesdatenschutzgesetz zu beachten. Daneben sind die Datenschutzbestimmungen der Schweiz zu beachten.

(b) Gemäß Artikel 6 DSGVO ist die Verarbeitung personenbezogener Daten, die zur Erfüllung eines Vertrages oder zur Durchführung vorvertraglichen Maßnahmen erforderlich sind, erlaubt. Daneben kann die MIT die explizite Einwilligung der Verarbeitung von der betroffenen Person erheben.

(c) In diesem Fallbeispiel sind die sieben Grundsätze für die Verarbeitung personenbezogener Daten gemäß Artikel 5 DSGVO zu beachten. Im Einzelnen: „Rechtmäßigkeit, Verarbeitung nach Treu und Glauben", „Transparenz", „Zweckbindung", „Datenminimierung", „Richtigkeit", „Speicherbegrenzung", „Integrität und Vertraulichkeit" sowie „Rechenschaftspflicht".

(d) Gemäß Artikel 6 f DSGVO ist die Verarbeitung personenbezogener Daten auch dann, wenn die Verarbeitung ist zur Wahrung der berechtigten Interessen des Verantwortlichen oder eines Dritten erforderlich ist. Die Verarbeitung der Daten für eine interne Analyse wie das Profittesting kann man als berechtigtes Interesse der MTI ansehen. Zum Schutz der betroffenen Personen sind geeignete Maßnahmen, wie Anonymisierung oder Pseudonymisierung der Daten in Betracht zu ziehen.

(e) Da die Produkte in Deutschland verkauft werden, ist auch hier die KI-Verordnung anzuwenden. Welches Lernverfahren hinter dem KI-basierten Profittesting steckt, ist für die Einschätzung der regulatorischen Anforderungen zunächst nicht relevant. Die Anwendung erfüllt keines der Kriterien des Artikel 5 der KI-Verordnung und ist deshalb nicht verboten. Auch handelt es sich nicht um eine Hochrisikoanwendung, da keines der Kriterien in Anhang III erfüllt ist. Möglicherweise handelt es sich um eine KI-Anwendung mit allgemeinem Verwendungszweck für die dann besondere Transparenzanforderungen gelten würden. Dies wäre beispielsweise dann der Fall, wenn das KI-basierte Profittesting auf GenAI basieren würde. Das wäre zu klären. Schließlich stellt sich die Frage, ob es bei der MIT eine Selbstverpflichtung im Sinne des Artikel 95 der KI-Verordnung gibt. Ein solcher Verhaltenskodex müsste dann beachtet werden. Andernfalls gibt es keine regulatorischen Anforderungen.

Aufgabe 2 [Data Science – Von den Daten zum Modell zur Anwendung & KI machen] [45 Punkte]

Sie sind Mitarbeiter eines mittelständischen Versicherungsunternehmens (VU). Dies ist Ihre erste Stelle nach dem Studium. Ihre Chefin spricht Sie auf das Thema Actuarial Data Science an.

a) *[6 Punkte]* Welche Datenquellen sind für einen Actuarial Data Scientist im aktuariellen Umfeld relevant? Nennen Sie drei Datenquellen. Nennen Sie zudem drei mögliche Anwendungen im aktuariellen Kontext.

b) *[4 Punkte]* Erläutern Sie die Begriffe strukturierte, semistrukturierte und unstrukturierte Daten in jeweils einem Satz. Welcher der genannten Begriffe ist Ihrer Ansicht nach im Versicherungsumfeld besonders relevant? Begründen Sie Ihre Wahl.

c) [9 Punkte] Nennen Sie drei Geschäftsprozesse von Versicherungs-unternehmen, aus denen Sie für die Neuauflage eines bestehenden Lebensversicherungsproduktes, erstmalig unter Nutzung von Machine Learning-Verfahren, noch weitere Daten für die Kalkulation der Prämien gewinnen könnten.

d) [6 Punkte] Nennen und beschreiben Sie ein Ihnen bekanntes Prozessmodell für Data Mining.

e) [10 Punkte] Erläutern Sie die wesentlichen Vor- und Nachteile von Data Science- und KI-Entwicklung *on premise* gegenüber der Entwicklung in der *Cloud*.

f) [10 Punkte] In Ihrer Abteilung soll ein neuer KI-Stack auf Basis von Python mit der Anaconda-Umgebung eingeführt werden. Erläutern Sie die wesentliche Funktionsweise von virtuellen Environments und nennen Sie vorhandene Einschränkungen, auf die Sie sich im Umgang mit der IT-Bereitstellung einstellen müssen.

Lösungsvorschlag:

Aufgabe 2

a) Datenquellen:

- Informationen aus den Bestandsführungssystemen, DWH-Systemen, …
- Informationen aus Wearables, Daten aus der Nutzung von Apps, Homepage, …
- Kundendaten (neben den klassischen Bestandsdaten auch aus der Korrespondenz mit dem Kunden), z.B. Infos über Familienstand und Familienmitglieder, …
- Wetterdaten
- Unfalldaten, Daten zu Einbrüchen, …
- Daten aus Unfallmeldesystemen beim Auto
- …

Anwendungen

- pay as you drive-Tarife
- bessere Auswertung der eigenen Tarife z.B. Rückschlüsse zu Zusammenhängen aus der Risikoprüfung zu späteren Schadenfällen
- bessere Vorhersagen von Unwettern, Erdbeben, …
- weitere Lösungen möglich

b) **strukturierte Daten**

- haben ein vorgegebenes Format, in das sich alle Informationen einordnen lassen: normalisierte Form
- innerhalb einer relationalen Datenbank haben sie eine Zeilen- und Spaltenposition
- sind leicht zu finden und zu bearbeiten
- man kann sie z.B. mit SQL auslesen

semistrukturierte Daten

- haben keine allgemeine Struktur, tragen aber einen Teil der Strukturinformationen in sich
- sind nicht vollständig strukturiert und nicht vollständig unstrukturiert
- Beispiel: E-Mail: Grundstruktur mit Absender, Empfänger, Betreff und weitere Informationen des Nachrichtenkopfes. Der eigentliche Inhalt ist ein strukturloser Text.

unstrukturierte Daten:

- Informationen, die in einer nicht identifizierbaren und nicht normalisierten Datenstruktur vorliegen
- kein Datenmodell vorhanden
- schwer zu analysieren und zu verarbeiten
- Beispiele: Texte, Bilder, Audio- und Videodateien, Präsentationen

c) Mögliche **Geschäftsprozesse** sind:

- Gesundheitsprüfung: hier könnten Daten von Zahnarztbesuchen, Arztberichten, Röntgenbilder verwendet werden.
- Schadensmanagement: hier könnten Daten aus den eingereichte Zahnarztrechnungen ausgelesen und verwendet werden.
- Kundenmanagement: hier könnten Daten über das Anrufverhalten, bzw. Kontaktverhalten ausgewertet werden.
- Vertriebssteuerung: hier könnten Daten über die vergangenen Abschlüsse pro Region o.ä. ausgewertet werden.
- ...

d) Ein häufig verwendetes Modell ist das sog. **CRISP-DM**, lang: Cross-Industry Standard Process for Data Mining. Dies ist in sechs Phasen gegliedert.

Business Understanding: Aufbau des Verständnisses über die zu lösende Herausforderung und die vorherrschenden Rahmenbedingungen intern wie extern.

Data Understanding: Aufbau des Verständnisses über die zur Verfügung stehenden Daten, Kosten, Struktur, Hindernisse und Interfaces. Insbesondere muss die Hypothese belegt werden, dass die Daten überhaupt zur Lösung der Aufgabe in Frage kommen, nicht aber *wie* dies geschehen kann.

Data Processing: Aufbereitungsphase, in der die vorhandenen bzw. erworbenen Daten vorverarbeitet und umformatiert werden, damit sie der geplanten Modellierung zugänglich werden.

Modelling: Anwendung von Algorithmen des Maschinellen Lernens bzw. der stat. Inferenz, um aufgestellte Hypothesen zu prüfen, bzw. Vorhersagetools zu trainieren bzw. Muster in den Daten zu finden.

Evaluation: Betrachtung von Kennzahlen und Heuristiken zur Beantwortung der Frage, ob die erzeugten Modelle und Ergebnisse aussagekräftig und ursächlich nützlich für die im Business Understanding definierten Ziele sind. [Dies ist nicht gleichbedeutend mit einer guten Minimierung der Zielfunktion, sondern nimmt ausdrücklich Bezug auf die Ursprungsausgabe, nicht nur die synthetische Performance des Modells.]

Implementation: Anwendung der gewonnenen Erkenntnisse und Modelle auf die Geschäftsprozesse; mithin Umsetzung der Lösungsstrategien zur Verbesserung des Geschäfts oder Erfüllung des gesetzten Ziels.

e) Bei der Betrachtung von Entwicklungsumgebungen sind folgende Aspekte zu beachten:

- Berechnungsressourcen: Während in der Cloud durch *Hyperscaling* praktisch beliebige *Workloads* umgesetzt werden können, ist bei lokaler Entwicklung das Limit durch die *Anzahl und Ausstattung der Hardware* relevant. Umgekehrt neigen Cloud-Umgebungen durch genau diesen Umstand dazu, *unnötige Kosten durch Berechnungen* zu erzeugen, die nicht notwendig wären oder nicht ausreichend optimiert laufen. Auch ist on premise sicherzustellen, dass KI-Ressourcen nicht mit dem *Mainframe bzw. Data Warehouse* geteilt werden müssen, um bspw. die klassischen Batch-Prozesse nicht zu beeinträchtigen.
- Laufzeitenvironment: Beide Umgebungen erlauben prinzipiell die Verwendung beliebiger Software-Versionen. Während in der Cloud *managed solutions* die z.T. aufwendige *Aktualisierungsverwaltung* abnehmen können, kann es dadurch zu Inkompatibilitäten kommen. Produktivanwendungen sollten daher immer in *Containern* (Docker o.ä.) gekapselt sein.

- Sicherheit und Datenschutz: Bei externer Lösung muss ein Auftragsdatenverarbeitungsvertrag mit dem Cloud-Anbieter geschlossen werden und über den eigenen Code hinaus Privacy-by-Design und Security-by-Design umgesetzt werden. Je nach Anbieter kann dies ein Ausschlusskriterium sein. Diese Aspekte müssen lokal auch beachtet werden, können aber immerhin selbst gesteuert werden.
- Nachhaltigkeitsaspekte: Rechenzentrum ist nicht gleich Rechenzentrum und Workloads erzeugen neben CO2-Auswirkungen möglicherweise weitere biosphärenadverse Effekte: Kühlwasser-Management, Nachhaltigkeitsaspekte von Komponenten und Softwareffizienz sind alles Aspekte, die sich negativ auf die Reputation des einsetzenden Unternehmens auswirken können.
- Support: Während Cloud-Anbieter typischerweise auch Enterprise-Support für ihre Kunden anbieten, muss bei einer on premise-Lösung diese Expertise entweder in der eigenen IT oder bei einem Partner zur Verfügung stehen. Grundsätzlich kann letzteres günstiger sein, erfordert aber aufwendigeres Management.
- Weitere Antwortmöglichkeiten sind möglich.

f) **Virtuelle Environments** mit Anaconda:

- Virtuelle Environments erlauben das Entkoppeln der Python-Bibliotheken von der Laufzeitumgebung des Betriebsystems. Dadurch können Kombinationen von Bibliotheken für spezifische Zwecke angepasst bzw. zusammengestellt werden. Conda und pip können kombiniert werden: Sobald ein conda-Environment erstellt wurde, kann mit der dort installierten Kopie von pip auch dort Drittquellenpakete verwendet werden.
- Das flexible Management erlaubt es daher, einzelne Projekte zur gleichen Zeit mit unterschiedlichen Softwareversionen zu entwickeln. Die meisten IDEs wie VSCode oder PyCharm können mit verschiedenen Environments umgehen und diese z.T. Auch begrenzt managen.
- Das Conda-Paketmanagement erlaubt die Pflege von nicht-Python-spezifischen Abhängigkeiten wie dem CUDA-Laufzeitpaket, das zur GPU/TPU-Beschleunigung vieler KI-Anwendungen erforderlich ist.
- Im Gegensatz zu zum Python-Paketmanagement pip bietet die Anaconda-Umgebung grafische Tools zur Pflege der virtuellen Environments.
- Weitere Antwortmöglichkeiten sind möglich.

Einschränkungen:

- Python und Pakete sind größtenteils freie Software. Das bedeutet, dass diese generell ohne Support, Garantien oder Verantwortlichkeiten bereitgestellt werden und sich Abhängigkeiten und APIs von einer Version zur nächsten verändern können. Die besondere Herausforderung, die sich daraus ergibt, besteht also im Lifecycle-Management.
- Weitere Antwortmöglichkeiten sind möglich.

Aufgabe 3 [Überwachtes und unüberwachtes maschinelles Lernen] [45 Punkte]

a) [22 Punkte] Überwachtes Lernen

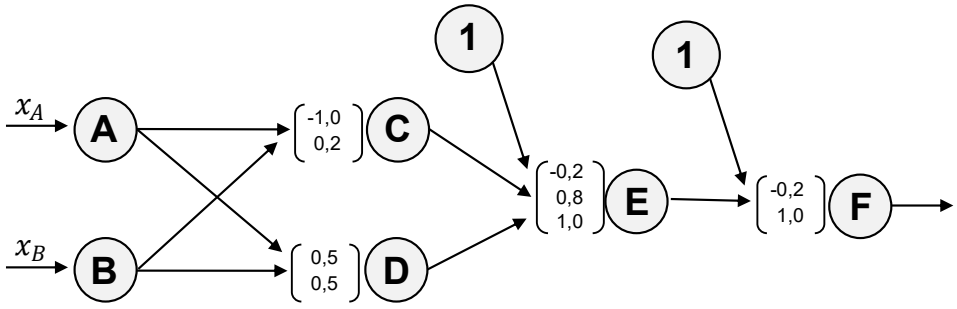
1. [2 Punkte] Geben Sie für ein Regressions- und Klassifikationsproblem jeweils ein Beispiel aus der Versicherungsbranche an.

2. [6 Punkte] Beschreiben Sie formal das Greedy-Verfahren zum Wachstum eines binären Entscheidungsbaumes. Gehen Sie dabei insbesondere auf die Rolle von Splitvariable, Schwellenwert, Splitkriterium und Stoppkriterium ein.

3. [1 Punkt] Die Tiefe eines Entscheidungsbaumes lässt sich über Stoppkriterien steuern. Nennen Sie zwei Beispiele für solche Stoppkriterien.

4. [6 Punkte] Beschreiben Sie die Grundidee von Ensemble-Methoden am Beispiel von Random Forests. Gehen Sie weiterhin darauf ein, durch welche beiden Techniken unterschiedliche Modellinstanzen erzeugt werden und beschreiben Sie beiden Techniken in jeweils zwei Sätzen.

5. [7 Punkte] Berechnen Sie für das abgebildete neuronale Netz einen Forwardpass für den Inputvektor $(x_A; x_B) = (0{,}4; 1{,}0)$, wenn als Aktivierungsfunktion *ReLu* (mit Schwellenwert Null) gewählt wird. Berechnen Sie dazu die Outputs der Units C, D, E und F.



b) [23 Punkte] Unüberwachtes Lernen

1. [2 Punkte] Skizzieren Sie zwei Anwendungsbeispiele der Clusteranalyse aus der Versicherungsbranche in jeweils einem Satz.

2. [1 Punkt] Wie lautet die Formel zu Berechnung von Klassendistanzen $D(C_1, C_2)$ zwischen je zwei Clustern C_1, C_2 nach dem Average-Linkage-Verfahren?

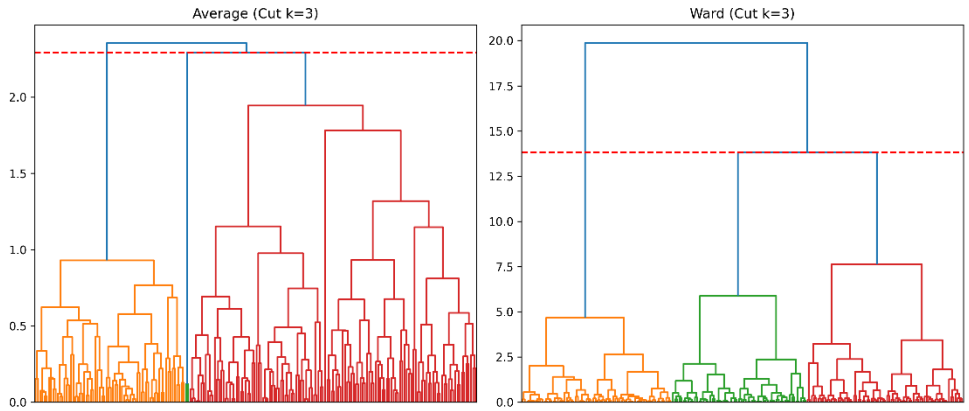
3. [1 Punkt] Das Average-Linkage-Verfahren gehört zu den agglomerativen Clusterverfahren. Beschreiben Sie in einem Satz das Prinzip der agglomerativen Verfahren der Clusteranalyse.

4. [9 Punkte] Gegeben sei eine Menge von Objekten $O = \{1,2,3,4\}$ mit folgender Objektdistanzmatrix:

$$\begin{pmatrix} 0 & 2 & 6 & 5 \\ 2 & 0 & 4 & 7 \\ 6 & 4 & 0 & 3 \\ 5 & 7 & 3 & 0 \end{pmatrix}$$

   Führen Sie ein agglomeratives Clusterverfahren mit Average-Linkage durch. Geben Sie dabei in jedem Iterationsschritt das Clustering an.

5. [3 Punkte] Für einen Datensatz mit 200 Objekten aus drei in etwa gleich großen Clustern wurden das Average-Linkage- und das Ward-Verfahren durchgeführt. Die Ergebnisse beider Verfahren sind durch die abgebildeten Dendrogramme mit einem Cut für $k = 3$ Cluster visualisiert. Welches der beiden Verfahren liefert das bessere Clustering? Begründen Sie.



6. [2 Punkte] Beschreiben Sie in zwei Sätzen das Konstruktionsprinzip eines Dendrogramms.

7. [2 Punkte] Beschreiben Sie in jeweils einem Satz die Ausgangssituation und die Fragestellung, die partitionierenden Verfahren der Clusteranalyse zugrunde liegen.

8. [3 Punkte] Wie lautet das Varianzkriterium der Clusteranalyse beim k-means Clustering? Geben Sie die Formel an und interpretieren Sie diese in zwei Sätzen.

Lösungsvorschlag Aufgabe 3:

a) [22 Punkte] Überwachtes Lernen

1. [2 Punkte]

   - Beispiel für ein Regressionsproblem mit numerischer Zielgröße: Höhe eines Schadensbetrags (z.B. Kfz-, Hausrat-, Gebäudeversicherung); Lebensdauer oder Restlaufzeit einer Police; …

   - Beispiel für ein Klassifikationsproblem mit kategorieller Zielgröße: Kündigt der Kunde die Police? (Ja / Nein); Ist ein gemeldeter Schaden betrugsverdächtig? (Ja / Nein); …

2. [6 Punkte] Greedy-Verfahren zur Wachstum eines binären Entscheidungsbaumes:

   - Zu Beginn wird der gesamte Kovariatenraum $\mathcal{X} \subseteq \mathbb{R}^p$ als eine Region betrachtet, d.h. $\mathcal{X} = R$.

   - In einem Knoten mit Region R wird eine Splitvariable X_{j^*} und ein Schwellenwert t^* gesucht, so dass die Aufteilung

     $R'(j^*, t^*) = \{\boldsymbol{X}\colon X_{j^*} \leq t^*\}$ und $R''(j^*, t^*) = \{\boldsymbol{X}\colon X_{j^*} > t^*\}$, $\boldsymbol{X} = (X_1, \dots, X_p)^T$,

     die größte Verringerung der Verlustfunktion L bewirkt bzw. das sogenannte Splitkriterium minimiert, d.h.

     $$(j^*, t^*) = \underset{j,t}{\operatorname{argmin}}\left[L\left(R'(j,t)\right) + L\left(R''(j^*, t^*)\right)\right]$$

   - Nach Bestimmung des besten Splits (j^*, t^*) werden die Daten in die Regionen R' und R'' aufgeteilt.

   - Der Prozess wird für jede neue Region rekursiv wiederholt, bis ein Stoppkriterium erfüllt ist.

3. [1 Punkt] Beispiele für Stoppkriterien:

   - Maximale Tiefe

   - Maximale Anzahl an Blättern

   - …

4. [6 Punkte]

   - Grundidee: Ein einfaches Lernverfahren (hier, ein Binärbaum) wird in mehreren Modellinstanzen (z.B. $b = 1, \ \ldots, B$) angewendet und anschließend die mittlere Prognose

     $$\hat{\mathrm{f}}_{\mathrm{Ensemble}}(\boldsymbol{x}) = \frac{1}{B}\sum_{b=1}^{B} \hat{\mathrm{f}}_b(\boldsymbol{x}), \ \ \boldsymbol{x} \in \mathcal{X},$$

     als finales Modell verwendet, wobei $\hat{\mathrm{f}}_b(\boldsymbol{x})$ die Prognosen der einzelnen, auf den Trainingsdaten entwickelten, Modellinstanzen bezeichnet.

   - Unterschiedliche Modellinstanzen können durch Bagging (Bootstrap Aggregating) und Random Subspace (Feature Bagging) erzeugt werden.

   - Bagging: Aus den Trainingsdaten $\mathcal{X}_{\mathrm{Train}}$ werden Bootstrap-Samples erzeugt (durch unabhängiges Ziehen mit Zurücklegen, bis ein neuer Datensatz derselben Größe entsteht). Für jedes zufällige Bootstrap-Sample wird ein Modell $\hat{f}_b$ trainiert.

   - Random Subspace: Zusätzlich wird aus den Trainingsdaten $\mathcal{X}_{\mathrm{Train}} \subseteq \mathbb{R}^p$ mit p Merkmalen für jeden Split eines Baumes eine zufällige Teilmenge von $q < p$ der Merkmale ausgewählt. Es wird also eine künstlichen Variabilität in den Merkmalen erzeugt, um zu verhindern, dass dominante Merkmale alle Bäume bestimmen.

5. [7 Punkte] Forwardpass für den Inputvektor $(x_A;\ x_B) = (0{,}4;\ 1{,}0)$:

   - Output von Unit C: $ReLU(0{,}4 \cdot (-1{,}0) + 1{,}0 \cdot 0{,}2) = ReLU\ (-0{,}2) = 0$
   - Output von Unit D: $ReLU(0{,}4 \cdot 0{,}5 + \ 1{,}0 \cdot 0{,}5) = ReLU\ (0{,}7) = 0{,}7$
   - Output von Unit E: $ReLU(-0{,}2 + 0 \cdot 0{,}8 + 0{,}7 \ \cdot 1{,}0) = ReLU\ (0{,}5) = 0{,}5$
   - Output von Unit F: $ReLU(-0{,}2 + 0{,}5 \cdot 1{,}0) = ReLU\ (0{,}3) = 0{,}3$

b) [23 Punkte] Unüberwachtes Lernen

1. [2 Punkte]

   Anwendungsbeispiele der Clusteranalyse aus der Versicherungsbranche:

   - Portfoliokonsolidierung, z.B. für die Migration von Versicherungsbeständen oder für Projektionsrechnungen
   - Risikofaktorenanalyse, z.B. für die Berufsklassifizierung in der Berufsunfähigkeitsversicherung oder für die Bildung von Typen- und Regionsklassen in der Kfz-Versicherung

- …

2. [1 Punkt] Gegeben der Objektdistanzen $d_{ij}, i,j \in O$, ist die Berechnung von Klassendistanzen $D(C_1, C_2)$ zwischen je zwei Clustern C_1, C_2 im Average-Linkage-Verfahren wie folgt definiert:

   $D(C_1, C_2) = \frac{1}{|C_1||C_2|} \sum_{i \in C_1} \sum_{j \in C_2} d_{ij}$

3. [1 Punkt] Prinzip der agglomerativen Verfahren der Clusteranalyse:

   Eine Folge von Clusterings $C^{(1)}, \ldots, C^{(n)}$ von n Objekten $O = \{1, \ldots, n\}$ wird so konstruiert, dass für jedes $k \in \{1, \ldots, n\}$ genau ein Clustering $C^{(k)}$ bestehend aus k Clustern erzeugt wird.

4. [9 Punkte]

   - Initiales Clustering: $C^{(4)} = \{\{1\}, \{2\}, \{3\}, \{4\}\}$

     Zugehörige Klassendistanzmatrix (= Objektdistanzmatrix):

     $$\begin{pmatrix} 0 & 2 & 6 & 5 \\ 2 & 0 & 4 & 7 \\ 6 & 4 & 0 & 3 \\ 5 & 7 & 3 & 0 \end{pmatrix}$$

     Kleinstes Element ist $D(\{1\}, \{2\}) = 2$.

   - Fasse daher die beiden Cluster $\{1\}$ und $\{2\}$ zu einem Cluster zusammen und erhalte das Clustering $C^{(3)} = \{\{1, 2\}, \{3\}, \{4\}\}$ mit den (neuen) Klassendistanzen nach Average-Linkage:

     - $D(\{1,\ 2\}, \{3\}) = \frac{6+4}{2} = 5$
     - $D(\{1,\ 2\}, \{4\}) = \frac{5+7}{2} = 6$
     - $D(\{3\}, \{4\}) = 3$

     Zugehörige Klassendistanzmatrix:

     $$\begin{pmatrix} 0 & 5 & 6 \\ 5 & 0 & 3 \\ 6 & 3 & 0 \end{pmatrix}$$

     Kleinstes Element ist $D(\{3\}, \{4\}) = 3$.

- Fasse daher die beiden Cluster $\{3\}$ und $\{4\}$ zu einem Cluster zusammen und erhalte das Clustering $C^{(2)} = \{\{1, 2\}, \{3, 4\}\}$ mit den (neuen) Klassendistanzen nach Average-Linkage:

  ➢ $D(\{1, 2\}, \{3, 4\}) = \frac{6+5+4+7}{4} = 5{,}5$

  Zugehörige Klassendistanzmatrix:

$$\begin{pmatrix} 0 & 5{,}5 \\ 5{,}5 & 0 \end{pmatrix}$$

- Fasse die beiden verbleibenden Cluster zusammen und erhalte das Clustering $C^{(1)} = \{\{1, 2, 3, 4\}\}$.

5. [3 Punkte]

   - Da die zu klassierenden Objekte aus drei in etwa gleichgroßen Clustern stammen, liefert das Ward-Verfahren das bessere Ergebnis, da das Dendrogramm drei deutliche, etwa gleich große Cluster zeigt.

   - Bei dem Average-Linkage-Verfahren zeigt das Dendrogramm nur zwei deutliche Cluster bzw. das dritte Cluster enthält nur wenige Elemente.

6. [2 Punkte] Konstruktionsprinzip eines Dendrogramms:

   - Horizontal werden die Objekte so angeordnet, dass sich die Äste nicht überlappen.

   - Vertikal werden die Klassendistanzen als Verzweigungspunkte dargestellt.

7. [2 Punkte] Partitionierende Verfahren der Clusteranalyse:

   - Ausgangssituation: Gegeben ist eine Menge von n Objekten, die durch metrische Merkmalsvektoren $x_1, \ldots, x_n \in \mathbb{R}^p$ repräsentiert sind, und für die eine Partition in k Cluster $C_1, \ldots, C_k$ vorliegt.

   - Fragestellung: Kann die gegebene Partition – bezüglich eines Optimalitätskriteriums – bei fester Clusteranzahl k zu einer Partition $C_1^*, \ldots, C_k^*$ optimiert werden?

8. [3 Punkte] Varianzkriterium der Clusteranalyse beim k-means Clustering:

   - Formel: $\sum_{i=1}^{k} \sum_{j \in C_i} \left\| x_j - \overline{x}_{C_i} \right\|^2 \rightarrow \min_{C_1, \ldots, C_k}$

   - Interpretation: $\sum_{i=1}^{k} \sum_{j \in C_i} \left\| x - \overline{x}_{C_i} \right\|^2$ beschreibt die sogenannte „total within-cluster variation“ des Clustering $C = \{C_1, \ldots, C_k\}$. Gesucht wird also ein Clustering $\{C_1^*, \ldots, C_k^*\}$ mit hoher Homogenität innerhalb der Cluster.

Aufgabe 4 [Generative Künstliche Intelligenz] [45 Punkte]

a) [6 Punkte] Beantworten Sie die folgenden Fragen – eine Begründung ist nicht erforderlich. Pro Teilaufgabe können **mehrere** Aussagen zutreffen. Bitte schreiben Sie jeweils **alle** zutreffenden Antworten auf Ihren Lösungsbogen. Bei einer fehlenden oder falschen Nennung pro Teilaufgabe werden keine Punkte vergeben.

1. [2 Punkte] Welche der folgenden Punkte sind Nachteile von Embeddings im Kontext der Sprachverarbeitung?
   - i. Fehlende Erklärbarkeit neuronaler Netze
   - ii. Übernahme des Bias aus Trainingsdaten
   - iii. Gefahr fehlender Verarbeitbarkeit unbekannter Wörter
   - iv. Fehlendes Verständnis von Analogien
2. [2 Punkte] Welche der folgenden Bestandteile sind Teil des Decoder-only Transformers (z. B. GPT)?
   - i. Embedding
   - ii. Language Modeling Head
   - iii. Retrieval
   - iv. Sampling der nächsten Tokens
3. [2 Punkte] Was trifft auf die Top-Down-Tokenisierung zu?
   - i. Vokabular ist vordefiniert
   - ii. Segmentierung erfolgt regelbasiert
   - iii. Vokabular wird gelernt
   - iv. Training eines tiefen neuronalen Netzes

b) [6 Punkte] Wofür stehen die drei Buchstaben in der Abkürzung RAG? Erläutern Sie die drei Bestandteile in jeweils bis zu zwei Sätzen.

c) [5 Punkte] Beantworten Sie die folgenden Fragen zu Embeddings:

1. [2 Punkte] Was sind Embeddings und wozu werden sie genutzt?

2. [3 Punkte] Fertigen Sie eine beispielhafte zweidimensionale Skizze für sinnvolle Embeddings der Begriffe „Mann“, „Frau“, „Aktuar“, „Aktuarin“ an. Formulieren Sie die Vektoroperation, die das Wort „Aktuar“ aus den anderen drei Begriffen ableitet, d. h. beschreiben Sie die Funktion

$$\text{Aktuar} \approx F(\text{Mann}, \text{Frau}, \text{Aktuarin}).$$

d) [8 Punkte] Es sei das folgende Beispiel-Korpus gegeben:

*Wir möchten Ihr **Auto** kaufen!*

*Sie sind es leid, Ihr altes **Auto** fahren zu müssen? Sie möchten lieber ein modernes, umweltfreundliches **Auto** fahren? Dann sind Sie bei unserem **Auto** Ankauf genau richtig. Wir möchten Ihr **Auto** kaufen und machen Ihnen ein attraktives Angebot. Bevor wir Ihnen dieses unterbreiten, werden wir Ihr **Auto** gewissenhaft prüfen, damit auch der nächste Besitzer noch lange mit dem **Auto** fahren kann. Kaufen Sie Ihr neues **Auto** doch auch gleich bei uns!*

1. [5 Punkte] Ermitteln Sie die Wahrscheinlichkeiten, die sich unter Verwendung des Bi-Gram-Modells für das nächste Wort des vom User vorgegebenen Satzes „*Ich würde ausgesprochen gern ein altes Auto*“ auf Basis des Beispiel-Korpus ergeben. Nennen Sie alle Wörter, welche eine von Null verschiedene Wahrscheinlichkeit haben. Stellen Sie tabellarisch die Wahrscheinlichkeiten für diese Wörter dar.
2. [1 Punkt] Welches Wort würde folgen, wenn Sie eine Greedy-Auswahl anwenden?
3. [2 Punkte] Welches nächste Wort würden Sie bei einer Greedy-Auswahl für das Tri-Gram-Modell erhalten? Begründen Sie Ihre Antwort kurz.

e) [8 Punkte] Beim Top-k Sampling-Verfahren werden die k Token mit den höchsten Wahrscheinlichkeiten

$$p_i := \frac{\exp(u_i/T)}{\sum_j \exp(u_j/T)} \quad , i = 1 \dots |V|,$$

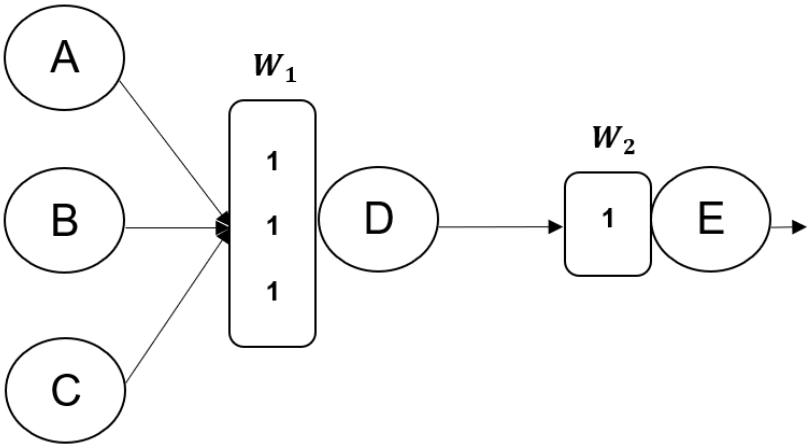
gewählt, wobei V das gegebene Vokabular und u das „Unembedding“ bezeichne. Anschließend wird die Verteilung renormalisiert als

$$p_i' := \frac{p_i}{\sum_{m \in Top-k} p_m} \qquad \forall i \in \textit{Top-k.}$$

Wir nehmen an, dass $u_i > u_{i+1}$ für alle $i = 1 \dots |V| - 1$.

Welche Wahrscheinlichkeiten p_i' ergeben sich für $T \to 0$ (mit Herleitung)? Welcher Verteilung entspricht dies? Wie heißt das entsprechende Grenzverfahren?

f) [12 Punkte] Es sei das untenstehende neuronale Netz gegeben. Als Aktivierungsfunktionen nutzt das Netz im Hidden-Layer die Leaky ReLU-Funktion (mit Schwellenwert 0 und Parameter $\alpha = 0{,}1$) und im Output-Layer die Identitätsfunktion. Es gibt keine Bias-Neuronen.



Für das neuronale Netz mit den angegebenen Gewichten wurde bereits ein Forward-Pass berechnet. Die hierzu verwendeten Input-Werte und Ergebnisse sind der folgenden Tabelle zu entnehmen:

Index i	Feature x_i	Prediction $\hat{y}_i$	Target y_i
1	(0,2; 0,3; 0,4)	0,9	1
2	(-0,1; 0,5; -0,2)	0,2	2
3	(-1; 1,5; -0,2)	0,3	1

(1) [2 Punkte] Zur Berechnung des Training-Losses soll der Mean-Squared-Error (MSE) verwendet werden. Geben Sie die Definition des MSE an. Berechnen Sie anschließend den Trainings-Loss, der sich nach dem aktuellen Forward-Pass aus der gegebenen Tabelle ergibt. Runden Sie Ihr Ergebnis auf **2 Nachkommastellen**.

(2) [4 Punkte] Berechnen Sie das neue Gewicht $\boldsymbol{W_2^{neu}}$, das sich bei einem Back-Propagation Schritt mit der MSE-Verlustfunktion und der Lernrate 0,3 ergibt.

(3) [6 Punkte] Berechnen Sie nun auch das neue Gewicht $\boldsymbol{W_1^{neu}}$, das sich bei einem Back-Propagation Schritt mit der MSE-Verlustfunktion und der Lernrate 0,3 ergibt.

Lösungsvorschlag:

Aufgabe 4

a)

(1) ii), iii)

(2) i), ii), iv)

(3) i), ii)

b) **Retrieval:** Abgleich der Query mit verfügbaren Dokumenten und Rückgabe eines Relevanz-Rankings. So werden die relevantesten Informationen aus einer externen Wissensquelle identifiziert.

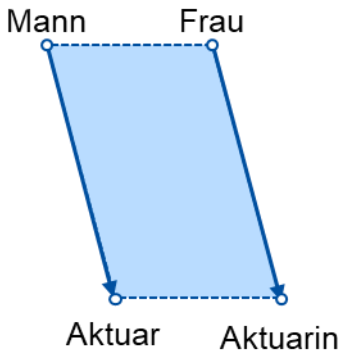
Augmented: Hier wird die Query angepasst (Prompt Engineering), um auf spezifische Dokumente/Chunks zu verweisen.

Generation: Hier erstellt das Large Language Model die Antwort. Es kombiniert das interne Sprachverständnis mit den externen Informationen, um eine kohärente, kontextreiche Ausgabe zu erzeugen.

c)

(1) Embeddings sind Vektorrepräsentationen von Wörtern oder Tokens, die deren Bedeutung in einem mathematischen Raum abbilden. Durch Berechnung der Ähnlichkeit zwischen zwei Vektoren/Embeddings (mittels des Kosinus bzw. des normierten Skalarprodukts) kann semantische Ähnlichkeit mathematisch erfasst werden. Embeddings werden beispielsweise für Transformer-Modelle genutzt.

(2) Eine Beispielskizze könnte wie folgt aussehen:



Vektoroperation:

$$\text{Aktuar} \approx \text{Aktuarin} + (\text{Mann} - \text{Frau})$$

d)

(1) Das letzte Wort des Satzanfangs ist „Auto", sodass die folgenden Bi-Gram-Wahrscheinlichkeiten ermittelt werden müssen:

$$\hat{P}(kaufen|Auto) = \frac{Count(Auto,\ kaufen)}{Count(Auto)} = \frac{2}{8} = 0{,}25$$

$$\hat{P}(fahren|Auto) = \frac{Count(Auto,\ fahren)}{Count(Auto)} = \frac{3}{8} = 0{,}375$$

$$\hat{P}(Ankauf|Auto) = \hat{P}(gewissenhaft|Auto) = \hat{P}(doch|Auto) = 0{,}125$$

Für die Tabelle ergibt sich somit:

Wort	**Wahrscheinlichkeit**
kaufen	0,25
fahren	0,375
Ankauf	0,125
gewissenhaft	0,125
doch	0,125

Hinweis: Die Tabelle ist als Antwort ausreichend, der obige Rechenweg ist zur Beantwortung der Frage nicht zwingend nötig.

(2) Bei einer Greedy-Auswahl wird das Wort mit der höchsten Wahrscheinlichkeit gewählt, hier also „fahren“.

(3) Beim Tri-Gram-Modell müssen die beiden letzten Wörter des Satzanfangs berücksichtigt werden, hier also die Sequenz (altes, Auto). Diese tritt im Beispielkorpus nur einmal auf. Als Wahrscheinlichkeit ergibt sich also 1 für das darauffolgende Wort „fahren“. Bei einer Greedy-Auswahl würde auch hier das nächste Wort „fahren“ lauten.

e) Es gilt

$$p_i' = \frac{p_i}{\sum_{m\in\text{Top-}k} p_m} = \frac{\frac{\exp(u_i/T)}{\sum_j \exp(u_j/T)}}{\sum_{m\in\text{Top-}k} \frac{\exp(u_m/T)}{\sum_j \exp(u_j/T)}} = \frac{\exp(u_i/T)}{\sum_{m\in\text{Top-}k} \exp(u_m/T)}$$

$$= \frac{\exp(u_i/T)}{\exp(u_i/T) + \sum_{m\in\text{Top-}k, m\neq i} \exp(u_m/T)} = \frac{1}{1 + \sum_{m\in\text{Top-}k, m\neq i} \exp\big((u_m - u_i)/T\big)}$$

Eine Fallunterscheidung ergibt für $u_m - u_i > 0$, dass $\exp\big((u_m - u_i)/T\big) \longrightarrow \infty$ für $T \to 0$, während für $u_m - u_i < 0$ gilt, dass $\exp\big((u_m - u_i)/T\big) \longrightarrow 0$ für $T \to 0$. Somit konvergiert p_1' gegen 1, da $u_1 > u_m$ für alle $m \neq 1$, andernfalls, d. h. für $i > 1$, konvergiert p_i' gegen 0.

Die Verteilung ist also im Grenzwert degeneriert, ihre gesamte Masse wird in einem einzigen Punkt konzentriert. Es wird also nur aus einem Token gesampled. Dieses Token ist das Token mit der höchsten Wahrscheinlichkeit p_1, denn aus $u_1 > u_m$ für alle $m \neq 1$ folgt $\exp(u_1/T) > \exp(u_m/T)$ für alle $m \neq 1$ und nach Definition somit auch $p_1 > p_m$ für alle $m \neq 1$. Im Grenzwert handelt es sich also um das Greedy-Verfahren.

f)

(1) Der Mean-Squared-Error berechnet sich als

$$\text{MSE}(\hat{Y}, Y) = \frac{1}{n}\|\hat{Y} - Y\|_2^2 = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

Für die gegebenen Werte gilt $n = 3$ und

$$\text{MSE}(\hat{Y}, Y) = \frac{1}{3}(0{,}1^2 + 1{,}8^2 + 0{,}7^2) = \frac{187}{150} \approx 1{,}25$$

(2) Das neue Gewicht berechnet sich (in der Notation der Folien) als

$$W_2^{neu} = W_2 - \varepsilon \cdot \frac{\partial \mathcal{L}}{\partial W_2}$$

mit

$$\frac{\partial \mathcal{L}}{\partial W_2} = \frac{\partial \mathcal{L}}{\partial \hat{Y}} \cdot \frac{\partial \hat{Y}}{\partial W_2} = A_1^T \cdot \frac{2}{n}(\hat{Y} - Y),$$

da als Verlustfunktion der MSE gewählt wurde und $\hat{Y} = id(Z_2) = A_1 W_1$. Für die Aufgabe gilt $n = 3$, $\varepsilon = 0{,}3$,

$$(\hat{Y} - Y) = \begin{pmatrix} 0{,}9 - 1 \\ 0{,}2 - 2 \\ 0{,}3 - 1 \end{pmatrix} = \begin{pmatrix} -0{,}1 \\ -1{,}8 \\ -0{,}7 \end{pmatrix} \quad \text{und} \quad W_2 = 1.$$

Weiterhin können wir aus den Formeln $Z_2 = A_1 W_2$ mit $W_2 = 1$ und $\hat{Y} = id(Z_2)$ schließen, dass $A_1 = \hat{Y}$, also

$$A_1 = \begin{pmatrix} 0{,}9 \\ 0{,}2 \\ 0{,}3 \end{pmatrix}.$$

Einsetzen liefert

$$\frac{\partial \mathcal{L}}{\partial W_2} = A_1^T \cdot \frac{2}{n} \cdot (\hat{Y} - Y) = \frac{2}{3} \cdot (0{,}9 \;\; 0{,}2 \;\; 0{,}3) \cdot \begin{pmatrix} -0{,}1 \\ -1{,}8 \\ -0{,}7 \end{pmatrix} = -0{,}44$$

und

$$W_2^{neu} = 1 - 0{,}3 \cdot (-0{,}44) = 1{,}132.$$

(3) Das neue Gewicht berechnet sich (wiederum in der Notation der Folien) als

$$W_1^{neu} = W_1 - \varepsilon \cdot \frac{\partial \mathcal{L}}{\partial W_1},$$

wobei mit $Z_1 = XW_1$ folgt

$$\frac{\partial \mathcal{L}}{\partial W_1} = X^T \cdot \frac{\partial \mathcal{L}}{\partial Z_1}.$$

Weiter ist

$$\frac{\partial \mathcal{L}}{\partial Z_1} = \frac{\partial \mathcal{L}}{\partial A_1} \cdot f'(Z_1)$$

wobei f die Aktivierungsfunktion bezeichne mit $A_1 = f(Z_1)$. Hier ist f die Leaky-ReLU Funktion, d.h. $f(z) \coloneqq \max(\alpha z, z)$. Es gilt also $f'(z) = 1$ für $z > 0$ und $f'(z) = \alpha$ für $z < 0$. Außerdem ist

$$\frac{\partial \mathcal{L}}{\partial A_1} = \frac{\partial \mathcal{L}}{\partial \hat{Y}} \cdot \frac{\partial \hat{Y}}{\partial A_1} = \frac{2}{n}\left(\hat{Y} - Y\right) \cdot W_2^T,$$

da als Verlustfunktion der MSE gewählt wurde und $\hat{Y} = A_1 W_2$. Insgesamt erhalten wir

$$W_1^{neu} = W_1 - \varepsilon \cdot X^T \cdot \frac{2}{n}\left(\hat{Y} - Y\right) \cdot W_2^T \cdot f'(Z_1),$$

Aus der Aufgabenstellung lässt sich ablesen, dass

$$W_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \quad \text{und} \quad X = \begin{pmatrix} 0{,}2 & 0{,}3 & 0{,}4 \\ -0{,}1 & 0{,}5 & -0{,}2 \\ -1 & 1{,}5 & -0{,}2 \end{pmatrix}$$

und aus

$$A_1 = f(Z_1) = \begin{pmatrix} 0{,}9 \\ 0{,}2 \\ 0{,}3 \end{pmatrix}$$

lässt sich schließen, dass bereits alle Einträge von Z_1 größer Null sein müssen, sodass für die Ableitung $f'(Z_1) = 1$ gilt.

Einsetzen liefert somit

$$W_1^{neu} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} - 0{,}3 \cdot \begin{pmatrix} 0{,}2 & -0{,}1 & -1 \\ 0{,}3 & 0{,}5 & 1{,}5 \\ 0{,}4 & -0{,}2 & -0{,}2 \end{pmatrix} \cdot \frac{2}{3} \cdot \begin{pmatrix} -0{,}1 \\ -1{,}8 \\ -0{,}7 \end{pmatrix} \cdot 1 \cdot 1$$

$$= \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} - 0{,}2 \cdot \begin{pmatrix} 0{,}86 \\ -1{,}98 \\ 0{,}46 \end{pmatrix} = \begin{pmatrix} 0{,}828 \\ 1{,}396 \\ 0{,}908 \end{pmatrix}.$$