Kurs: KI T-InfT-008 und 010 Datenmengen und Embedded Systems

Cândido Vieira

07.11.2024

Balthasar-Neumann-Technikum (BNT)

Inhaltsverzeichnis

1. Einführung in Datenleakage (DL)

- Was ist Datenleakage? (Data Leakage)

- Gefahr: Ein Modell wird zu genau und verliert an Generalisierbarkeit.

2. Bedeutung und Risiken von Datenleakage

- Überbewertung der Modellgenauigkeit: Unklare Trennung zwischen Training und Test.

- Negative Auswirkungen auf Geschäftsentscheidungen.

2. Beispiel für Datenleakage

- Hypothetisches Beispiel: Ein Finanzmodell mit zukünftigen Kursdaten im Training.

- Resultat: Übermäßig hohe Genauigkeit im Training, schlechte Generalisierung.

3. Arten von Datenleakage

- Zugriffsleakage (Access Leakage): Testdaten fließen ins Training.

- Feature-Leakage: Merkmale enthalten zukünftige oder unerlaubte Informationen.

- Ziel-Leakage (Target Leakage): Zielvariable im Feature-Set.

3. Zugriffsleakage erkennen

- Indikatoren: Hohe Trainingsleistung, Testleistung sinkt drastisch.

- Häufig bei schlecht konzipierter Datenaufteilung.

3. Feature-Leakage identifizieren

- Merkmal-Auswahl kritisch prüfen: Kann das Feature im Produktionskontext vorliegen?

- Beispiele: Zeitreihen-Daten oder aggregierte Features.

3. Ziel-Leakage vermeiden

- Beispiel: Inkludieren eines Merkmals, das von der Zielvariable abhängt.

- Lösung: Strenge Prüfung, ob ein Feature kausal unabhängig ist.

4. Identifikation von Datenleakage

- Symptome: Unrealistische Vorhersagen im Test, überdurchschnittliche Performance.

- Metriken: Analyse statistischen Werten zur Validität.

4. Datenleakage erkennen

- Automatisierte Metriken: Kontrolle der Diskrepanz zwischen Trainings- und Testgenauigkeit.

- Feature-Korrelation analysieren: Ist ein Feature zu eng mit dem Ziel verbunden?

5. Methoden zur Vermeidung von Datenleakage

- Ziel: Sicherstellen, dass nur Trainingsdaten im Modelltraining verwendet werden.

- Methoden: Strikte Test-Trennung und Prüfung von Features.

6. Split in Trainings- und Testdaten

- Hold-Out-Methode: Test- und Trainingsdaten fest aufteilen.
- Vorteile: Klare Trennung, wenn Datensatz groß genug ist.
- Praxis-Tipp: Konsistente Testumgebung durch zufälliges Sampling.

7. Cross-Validation und Datenleakage

- K-Fold Cross-Validation: Modelltraining mit verschiedenen Datenpartitionen.

- Vorteil: Leistungsstarke Schätzung ohne Leakage.

7. Bedeutung von Cross-Validation

- Vorteile: Bessere Performance-Messung, erhöhte Genauigkeit.

- Praxis-Tipp: Anwendung bei kleinen Datensätzen oder beschränkten Ressourcen.

8. Datenleakage bei Feature Engineering

- Regeln: Nur Features aus Trainingsdaten erzeugen.

- Typische Fehler: Durchschnitts- oder Zeitreihenaggregation aus Testdaten.

9. Anwendungsfall – Medizin

- Ziel-Leakage Beispiel: Diagnose- und Behandlungsdaten vermischen.

- Lösung: Klar definierte Zeitpunkte und Zeiträume für Trainings- und Testdaten.

9. Anwendungsfall – Finanzen

- Feature-Leakage Beispiel: Marktprognosen mit zukünftigen Kursen.

- Lösung: Nur historische Daten verwenden.

9. Anwendungsfall – Bildanalyse

- Vermeidung: Keine Verwechslung von Trainings- und Testbildern.

- Problem: Datenleakage durch Überlappung von Bildbereichen.

10. Schritt-für-Schritt-Anleitung zur Vermeidung

- Anwendung: Keine Features verwenden, die zukünftige Informationen enthalten.
- Best Practices: Datenaufteilung und Modellvalidierung durchgehend überprüfen.

- **Anleitung**:
  - 1. Split Data.
  - 2. Fit Data Preparation on Training Dataset.
  - 3. Apply Data Preparation to Train and Test Datasets.
  - 4. Evaluate Models.

10. Zusammenfassung der besten Praktiken

- Trennung von Training und Test: Ein Muss in jeder Modellpipeline.
- Strenge Kontrolle beim Feature Engineering.
- Cross-Validation zur soliden Performanzüberprüfung.

11. Ü1: Identifizierung von DL in einem Train-Test-Split

- Jason Brownlee - Data Preparation for Machine Learning - Data Cleaning, Feature Selection, and Data (2020, machine learning mastery)
    - Chapter 4 Data Preparation Without Data Leakage:
        - Listing 4.7: Example of evaluating a model using a train-test split with data leakage.
        - Listing 4.11: Example of evaluating a model using a train-test split without data leakage.

11. Ü2: Identifizierung von DL in einem K-Fold CV-Split

- Jason Brownlee - Data Preparation for Machine Learning - Data Cleaning, Feature Selection, and Data (2020, machine learning mastery)

  - Chapter 4 Data Preparation Without Data Leakage:

    - Listing 4.15: Example of evaluating a model using a cross-validation with data leakage.
    - Listing 4.19: Example of evaluating a model using a cross-validation without data leakage.

11. Praktische Tipps zur Vermeidung von Datenleakage

- Verfahren: Trainings- und Testdaten konsistent trennen.

- Automatisierungstools: Implementierung von Tests und Validierungsmethoden.

- Bild: Liste hilfreicher Tools.

11. Abschluss & Q&A

- Wichtige Erkenntnisse: Datenleakage verstehen, vermeiden und überwachen.

- Q&A: Diskussion und Klärung offener Fragen.

- Bild: Visuelle Zusammenfassung der besprochenen Punkte.

Referenzen

1. Brownlee, J. Data Preparation for Machine Learning, 2020, Machine Learning Mastery.
2. Kazil, J., Jarmul, K. Data Wrangling with Python, O'Reilly Media, 2016.