Veranstaltungen Data Science & Decision Science im SoSe2021

Was Wann Wo

- Hochaktuelle Fragestellungen aus dem Bereich Data Science und Decision Science

- Viele Themen sind fachlich bereits im ersten Studienabschnitt bearbeitbar
- Neben der Bearbeitung fachlicher Inhalte steht auch die freie Rede und Präsentation eines Themas im Fokus

- Bei Wunsch sind für Bachelor geeignete Themen auch als Cluster-Seminar (Finance bzw. Operations) ausgestaltbar

- Leistung kann im Wesentlichen auch bereits in der vorlesungsfreien Zeit erbracht werden

- Veranstaltungen finden wegen Corona **virtuell durch Präsentation per Video statt**

- Vorträge im Zweierteam

Themenübersicht

1. Clustering – k-means, DBSCAN & Co.
2. Statistische Signifikanz: Sinn und Unsinn
3. Visualisierung von Algorithmen
4. Artificial Intelligence: Anwendungen von Multi-Armed-Bandits in Planungsproblemen
5. Das Local-Level-Modell mit einer erklärenden Variable
6. Text Mining: Spam-Klassifikation von Textnachrichten
7. Regularisierung und Resampling Methoden
8. Data Visualization und Data Storytelling
9. Führt Deep Learning in eine Sackgasse?
10. Modellierung von Tail-Events mit Hilfe von Copulas
11. Das „Pendler-Paradoxon“
12. Threshold-Autoregressive (TAR)-Modelle

Einbringbarkeiten (Bachelor)

Projektstudium Data Science (WIW-0366) und (Cases in) Decision Science (WIW-0365)

- BWL und VWL (PO 2015): General Management & Economics, Major Finance & Information, Major Operations & Information Management
- GBM (PO 2013, PO 2015): Global Business and Economics
- WIN (PO 2015): Spezialisierung - Oper., Finanz- & Inf.management
- WING (PO 2011): Finance, Operations & Information Management
- WIMA (PO 2015): Gruppe E: WiWi-Grundlagen, Gruppe G: Wahlbereich/Wahlpflichtbereich
- Mathematik-Bachelor: N-WiWi: Nebenfach Wirtschaftswissenschaften

Bewerbungsverfahren

- Bewerbung ab sofort bis spätestens 28.02.2021 möglich
- Die Auswahl und Themenvergabe erfolgt nach Leistungskriterien: Jede Bewerberin/Jeder Bewerber **muss daher neben seiner Online-Bewerbung während der Bewerbungsfrist einen aktuellen, vollständigen Studis-Auszug** per Mail an karin.wuensch@wiwi.uni-augsburg.de schicken.
- Bewerbungsformular auf der Website des LS Okhrin (diese wird ab Beginn der Reading-Week verfügbar sein)
  - Hierbei muss jedes Team eine Präferenzordnung von bis zu 5 Themen angeben
  - Somit werden wir versuchen, den Nutzen zu maximieren: Teams, welche ausgewählt wurden und nicht ihre Erstpräferenz bekommen, werden möglichst ihre nächste Präferenz(en) angeboten bekommen, falls sie ausgewählt wurden.
- Bewerbungen **als Zweierteam werden präferiert behandelt**: Hierbei sind als Team nur Gruppenzusammensetzungen Bachelor+Bachelor bzw. Master + Master möglich
- Jedes Thema wird **höchstens einmal vergeben**
- Themen- und Ortzusage **bereits Anfang März 2021**
- **Ab Anfang März 2021 kann (aber muß nicht!) bereits mit der Themenbearbeitung begonnen werden und ein Großteil des Arbeitsaufwandes kann somit bereits in der Vorlesungsfreien Zeit erbracht werden und entlastet das Semester**

Hinweise zur Prüfungsleistung

Die Prüfungsleistung besteht aus einem **als Video zu erbringenden** Vortrag im Zweier-Team, gestützt auf eine Präsentation + anschließend eine kurze Live-Fragerunde (per Zoom)

Zu den Vorträgen:

- Vorträge im Zweierteam, frei gehalten
- Zieldauer 50 Minuten (gleichmäßig auf beide Vortragende aufgeteilt)
- Für das gesamte Publikum, nicht nur die Dozenten aufzubereiten
- Die Präsentation ist als **wissenschaftliche Leistung in Präsentationsform** zu betrachten!
- Abgabe der Präsentation & der Präsentationsvideos **in digitaler Form** bis (inklusive) 18.05.2021, 12:00 Uhr

1. Clustering – k-means, DBSCAN & Co.

Universität Augsburg University

In der Praxis versucht man häufig Beobachtungen zu gruppieren. Dabei sollen die Objekte, die zu einer Gruppe bzw. einem Cluster gehören, möglichst ähnlich sein und die Beobachtungen die zu verschiedenen Clustern gehören möglichst unterschiedlich sein. Zum Beispiel, kann ein Unternehmen mit solchen Verfahren die Kunden gruppieren oder ein Investmentfond kann die einzelnen Firmen anhand von deren Charakteristiken zu homogen Klassen clustern. Zusätzlich zu den klassischen Verfahren wie z.B. Hierarchisches Clustering gib es auch viele Ansätze, die spezielle Eigenschaften des Datensatzes ansprechen. Im Rahmen der Arbeit soll ein Verfahren, zum Beispiel DBSCAN, „model-based clustering“, k-means und „soft k-means“ erklärt und anhand von reellen Daten angewendet werden.

Aufgabenstellung

- Messung der Ähnlichkeiten und Distanzen
- Einführung ins Clustering und Vorstellung des Ausgewählten Verfahrens
- Wahl der Clusterzahl und Visualisierung der Ergebnisse
- Anwendung des Verfahrens an reelle Daten

Ausgangsliteratur

- SCHUBERT, SANDER, ESTER, KRIEGEL, XU: „DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN“, ACM Transactions on Database Systems, Vol. 42, No. 3, 2017 mit dem R-package: DBSCAN
- Friedman , Tibshiriani, Hastie, The Elements of Statistical Learning
- Eigene Recherche

Betreuer: Yarema Okhrin

2. Statistische Signifikanz: Sinn und Unsinn

Seit Mitte des 20. Jahrhunderts hat sich das Konzept der statistischen Signifikanz als Standard zur Beurteilung empirischer Ergebnisse etabliert. Fast ebenso alt ist die Kritik an fehlerhaften Praktiken und Schlussfolgerungen, die mit statistischen Signifikanzaussagen verbunden sind. Eine spezielle Quelle solcher Fehlinterpretationen ist der häufig als „Irrtumswahrscheinlichkeit“ bezeichnete p-Wert. So wird beispielsweise bei hinreichend großem Stichprobenumfang jeder noch so kleine Effekt „statistisch signifikant“, gleichgültig, ob dieser Effekt auch relevant ist.

Aufgabenstellung

- Gewähren eines Überblickes über die mit statistischen Signifikanztests verbundenen Probleme und potenziellen Irrtümer.
- Diskussion insbesondere auch der Problematiken des Publication Bias und des multiplen Testens.
- Erstellung eigener numerischer Beispiele zur Veranschaulichung der behandelten Phänomene.

Ausgangsliteratur

- Krämer W (2011): The cult of statistical significance: What economists should and should not do to make their data talk. Schmollers Jahrbuch 131:455–468
- Wasserstein RL, Lazar NA (2016): The ASA statement on p-values: Context, process, and purpose. The American Statistician 70:129–133
- Ziliak ST, McCloskey DN (2008): The cult of statistical significance. University of Michigan Press, Ann Arbor
- Eigene Recherche

Betreuer: Michael Krapp
Hinweis: Dieses Thema richtet sich primär an Bachelor-Studierende.

3. Visualisierung von Algorithmen

Universität Augsburg University

Um fachfremden Personen in Hinblick auf Algorithmen ein diesbezügliches Methodenverständnis in kurzer Zeit zu ermöglichen, ist die grafische Visualisierung (Animation) der Algorithmen gegenüber der ausschließlichen Verwendung mathematischer Notation oftmals die effizientere Wahl. Im Rahmen der Arbeit sollen mittels Python (+ vorzugsweise Blender) zwei ausgewählte Algorithmen der Optimierung (parallel zur mathematischen Notation) grafisch veranschaulicht werden. Die ausgewählten Algorithmen sind methodisch vergleichsweise einfach, weshalb sich die Bearbeitung des Themas auch für Studierende im ersten Studienabschnitt eignet. Die Visualisierung steht bei der Themenbearbeitung im Vordergrund.

Aufgabenstellung

- Selbstständige Einarbeitung in die Visualisierungsmöglichkeiten [Python (+ vorzugsweise Blender)]
- Vorstellung der Visualisierungsmöglichkeiten
- Erläuterung der beiden ausgewählten Algorithmen auf Basis der:
  - Grafischen Visualisierung
  - Mathematischen Notation
- Anfertigen jeweils eines (animierten) Kurzvideos je Algorithmus

Ausgangsliteratur

- Eigene Recherche (nach Absprache mit dem Betreuer)

Betreuer: Deniz Preil

4. Artificial Intelligence: Anwendungen von Multi-Armed-Bandits in Planungsproblemen

Multi-armed Bandits (zu deutsch „mehrarmige Banditen“) sind in Bezug auf deren Namensgebung zwar angelehnt an die Glücksspielwelt. Ihre Anwendungsgebiete sind von Glückspielautomaten in Casinos jedoch weit entfernt und finden sich in zahlreichen Bereichen des täglichen Lebens. Multi-armed Bandits sind dem Reinforcement Learning, einem Teilbereich der Artificial Intelligence, zugeordnet. Die wesentliche Problemstellung umfasst einen Entscheidungsträger, welcher die Wahlzwischen zahlreichen Aktionen bzw. „Armen“ hat. Jede Aktion generiert einen sogenannten Reward. Die Rewards sind i.d.R. jedoch stochastisch, weshalb die Identifikation der besten Aktion nicht trivial ist. Im Rahmen der Themenbearbeitung sollen das ausgewählte Multi-armed Bandit Algorithmen auf eine ökonomische Problemstellung aus dem Bereich des Operations Research angewendet werden.

Aufgabenstellung

- Einführung und Erläuterung der ökonomischen Problemstellung
- Vorstellung und Erläuterung der angewendeten Multi-armed Bandit Algorithmen.
- Anwendung der Verfahren auf die Problemstellung.

Ausgangsliteratur

- Bouneffouf D, Rish I (2019): A survey on practical applications of multi-armed and contextual bandits. arXiv preprint arXiv:1904.10040
- Sutton RS, Barto AG (2018): Reinforcement learning: An introduction. MIT Press
- Eigene Recherche

Betreuer: Deniz Preil

5. Das Local-Level-Modell mit einer erklärenden Variable

Universität Augsburg University

Die State-Space-Modellierung ist eine beliebte Methode zur Beschreibung und Vorhersage von Zeitreihendaten. Sie stellt eine einheitliche Methodik zur Behandlung einer Vielzahl von Problemen in der Zeitreihenanalyse dar. Grundlage bildet das Komponentenmodell, welches die Zeitreihe in saisonale Schwankungen, Trendschwankungen und Zufallsschwankungen zerlegt. Ferner nimmt sie an, dass sich die zeitliche Entwicklung dieser Komponenten durch unbeobachtete parametrischen Prozess beschreiben lässt. Durch die Hinzunahme erklärender Variablen kann die Modellierung verbessert werden. Eine konkrete Implementierung der State-Space-Modellierung stellt das Local-Level-Modell mit einer erklärenden Variable dar.

Aufgabenstellung

- Stellen Sie die Idee der State-Space-Modelle vor und erläutern Sie das Local-Level-Modell mit einer erklärenden Variable.
- Erläutern Sie das Verfahren zur Schätzung der Modellparameter.
- Wenden Sie das Local-Level-Modell an einem selbst gewählten Datensatz mit der Statistiksprache R an und interpretieren Sie die Ergebnisse.

Ausgangsliteratur

- Commandeur, J. J., & Koopman, S. J. (2007). An introduction to state space time series analysis. Oxford University Press.
- Durbin, J., & Koopman, S. J. (2012). Time series analysis by state space methods. Oxford university press.
- Eigene Recherche.

Betreuerin: Ellena Nachbar

6. Text Mining: Spam-Klassifikation von Textnachrichten

Spam-Nachrichten können einen erheblichen Schaden verursachen. Dieser entsteht durch den zusätzlichen Aufwand für die Identifikation und Bearbeitung von Spam-Nachrichten, sowie durch mögliche Schadsoftware, die über diesen Weg verbreitet werden kann. Sogenannte „Spam-Filter“ sollen Spam-Nachrichten automatisch identifizieren und deren Zustellung verhindern oder sie entsprechend markieren. Grundlage dieser Verfahren sind statistische Methoden, wie zum Beispiel das logistische Regressionsmodell.

Aufgabenstellung

- Stellen Sie Text-Mining Verfahren vor, mit denen Textdaten numerisch repräsentiert werden können.
- Präsentieren Sie das logistische Regressionsmodell, ein Machine Learning Verfahren zur Klassifikation von Daten.
- Klassifizieren Sie SMS-Nachrichten in Spam und Nicht-Spam und evaluieren Sie die Ergebnisse. Der Datensatz wird bereitgestellt.

Ausgangsliteratur

- Aggarwal, C. C., & Zhai, C. (Eds.). (2012). Mining text data. Springer Science & Business Media.
- Fahrmeir, L., Kneib, T., Lang, S., & Marx, B. (2007). Regression. Springer-Verlag Berlin Heidelberg.
- Salton, G., & McGill, M. J. (1983). Introduction to modern information retrieval. McGraw-Hill.
- Eigene Recherche.

Betreuerin: Ellena Nachbar

Hinweis: Dieses Thema wird nur für Bachelor-Studierende angeboten.

7. Regularisierung und Resampling Methoden

Universität Augsburg University

Im Machine-Learning (ML) bezeichnet der Bias-Variance Trade-off (deutsch:Verzerrung-Varianz Dilemma) das Phänomen, dass sich Verzerrung und Empfindlichkeit gegenüber Rauschen in den Trainingsdaten in einem Trade-off Verhältnis befinden, wenn es zur Minimierung dieser Fehlerquellen kommt. Während unflexible ML Modelle oftmals zu hohen Bias aber geringer Varianz neigen, ist bei komplexen ML Modellen das Gegenteil der Fall. Hier kann es zu einer Überanpassung an die Trainingsdaten kommen, was allgemeinals „Overfitting“ bzw. „Model Overfit“ bezeichnet wird.Um dies zu verhindern, stehen dem ML Anwender verschiedene bewährte Methoden zur Verfügung.

Aufgabenstellung

- Anschauliche Darlegung der Probleme, welche sich aus Overfitting und dem Bias-Variance Trade-off ergeben.
- Einführung in die Regularisierung und Vorstellung von Resampling Methoden
- Beispielhafte Anwendung dieser Methoden in R oder Python

Ausgangsliteratur

- Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2014. An Introduction to Statistical Learning: with Applications in R. Springer Publishing Company, Incorporated.
- Eigene Recherche.

Betreuer: Jonathan Pfahler

8. Data Visualization und Data Storytelling

Universität Augsburg University

In der Aufbereitung und Auswertung von Daten ist es grundlegend wichtig sich ein Bild von dem zugrundeliegenden Datensatz und den späteren Resultaten zu machen. So kann u. a. die Struktur der Daten erkannt werden und auf mögliche Muster eingegangen werden. Je nach Bedarf gibt es unterschiedliche Visualisierungsmethoden (Histogramme, Boxplotts, Streudiagramme,…), deren einfache Implementierung mit den Python Libraries Seaborn und Matplotlib möglich ist. Zudem ist es wichtig, dass die Interpretation der Visualisierungen verständlich ausgelegt werden (Data Storytelling).

Aufgabenstellung

- Verwenden sie den Boston House Prices Datensatz zur Erstellung geeigneter Seaborn/ Matplotlib Plots.
- Erklären Sie die verwendete Methode, vergleichen Sie diese mit anderen Methoden und illustrieren sie den individuellen Vorteil der Methode.
- Interpretieren Sie Ihre Visualisierungen und achten Sie auf eine kohärente „Story“.

Ausgangsliteratur

- Waskom, M. (2012) Seaborn: Statistical Data Visualisation. Abgerufen22.01.2020. https://seaborn.pydata.org
- Hunter, J., Dale, D., Firing, E., DroettboomM. (2012). Matplotlib. Abgerufen22.01.2020. https://matplotlib.org
- Eigene Recherche.

Betreuer: Jonathan Pfahler

Hinweis: Dieses Thema wird nur für Bachelorstudenten angeboten.

9. Führt Deep Learning in eine Sackgasse?

Viele Anwendungen der KI Methoden funktionieren außerordentlich gut, aber sehr häufig nur unter Laborbedingungen. Bei den Anwendungen in der realen Welt häufen sich gravierende Fälle wo diese Methoden versagt haben, wie z.B. rassistische Chatbots, Facial Recognition oder Recruitment Tools bis hin zu Autopilot, das durch kleine Stickers auf der Fahrbahn, getäuscht werden kann. Als potenziellen Grund wird von den Autoren des Papers sogenannte Underspecification gennant. Zusätzlich zu den theoretischen Überlegungen schildern Autoren die Problematik anhand von vielen Beispielen.

Aufgabenstellung

- Theoretische Erklärung der Problematik von Underspecification.
- Eigene Ausarbeitung eines einfachen Beispiels in R.
- Vorstellung der Ergebnisse aus Paper.

Ausgangsliteratur

- D'Amour, A. et el (2020) "Underspecification Presents Challenges for Credibility in Modern Machine Learning", https://arxiv.org/pdf/2011.03395.pdf
- Eigene Recherche

Betreuer: Eugen Heine

10. Modellierung von Tail-Events mit Hilfe von Copulas

Gemeinsame Preissenkungen von Finanzinstrumenten treten häufiger und ausgeprägter auf (insbesondere in der Krisenzeiten) als gemeinsame Preissteigerungen. Sogenannte „*Exceedance*“ Korrelationen erlauben die Asymmetrie der Abhängigkeiten zu messen. Mit Hilfe von Copulas ist es möglich, die Zusammenhänge zwischen zwei oder mehreren verschiedenen Faktoren zu modellieren.

Aufgabenstellung

- Der Begriff der Abhängigkeit soll mit Hilfe von Copulas definiert werden, darunter auch Tail-Abhängigkeit
- Verschiedene Familien von Copulas sollen präsentiert und deren Eigenschaften diskutiert werden
- Die Renditen sind auf Asymmetrie zu prüfen und passende Copulas sollen geschätzt werden

Ausgangsliteratur

- Genest, C., Favre, A.-C., 2007. Everything you always wanted to know about copula modeling but were afraid to ask. Journal of Hydrologic Engineering, 12, 347-368
- Nelsen, R. ,1999. An introduction to copulas. Springer
- R-paket copula
- Weitere selbst recherchierte Literatur und R-Pakete etc.

Betreuer: Eugen Heine

Hinweis: Dieses Thema wird nur für Bachelorstudenten angeboten.

11. Das „Pendler-Paradoxon“

Während der Corona-Krise wurde vielen arbeitenden Personen offenbar, dass der Wegfall der Pendelzeit zur Arbeit zu positivem Nutzen führt. Dennoch ist zu erwarten, dass Pendeln weiter Teil vieler Arbeitsverhältnisse sein wird. In einem Paper von 2008 stellen Stutzer und Frey eine Gleichgewichtshypothese auf: Ein rationales Individuum wird den Aufwand des Pendelns nur auf sich nehmen, wenn es diesen Nutzenverlust durch anders gearteten Zusatznutzen (mehr Gehalt, besserer Job etc.) ausgleichen kann. Dennoch stellen die Autoren in ihrer Studie einen negativen Zusammenhang zwischen der fürs Pendeln aufgewendeten Zeit und der berichteten „Zufriedenheit“ der befragten Personen fest: Die befragten Personen nehmen somit einen „negativen Netto-Nutzen“ des Pendelns war, welcher die Vorteile nicht aufwiegt. Dies widerspricht ökonomischen Gleichgewichtsüberlegungen. Da die Befragten dennoch den Zustand nicht verändern, benennen Stutzer und Frey den beschriebenen Widerspruch als „Pendler-Paradoxon“.

Aufgabenstellung

- Erläuterung der Modelle und eingesetzten Daten, mit welchen der Zusammenhang zwischen Pendeln und Zufriedenheit untersucht wird
- Vorstellung verschiedener (neuerer) Publikationen und ihrer Ergebnisse, welche neue Aspekte des Themas beleuchten

Ausgangsliteratur

- Frey, B.S., Stutzer., A. (2008): Stress that doesn’t pay: The Commuting Paradox. Scandavian Journal of Economics 110 (2), S. 339-366.
- Eigene Recherche

Betreuer: Sebastian Heiden

Hinweis: Das Thema wird nur für Bachelor-Studierende vergeben

12. Threshold-Autoregressive (TAR)-Modelle

Universität Augsburg University

Für viele Beispiele von Zeitreihen lassen sich Nichtlinearitäten und asymmetrisches Verhalten beobachten. Die Klasse von TAR Modellen ermöglicht durch verschiedene Zustände (Regimes), die durch lineare Prozesse (z.B. Autoregressive AR-Modelle) beschrieben werden, Nichtlinearität in partielle Linearität zu überführen. Hierdurch ist es möglich, zyklisches Verhalten von beobachteten Daten zu modellieren und valide Prognosen zu erstellen.

Aufgabenstellung

- Erklärung der Problematik an geeigneten Beispielen. Wie löst das TAR Modell diese Probleme? Was für Erweiterungen gibt es?
- Anpassung von TAR Modellen mit der Statistiksprache R an selbst gewählte Datensätze. Wie wird das Modell an die Daten angepasst? Wie bestimmt man die Regimes? (z.B. die Methode von Tsay).
- Interpretation der Ergebnisse und der Möglichkeit des Modells.

Ausgangsliteratur

- Tong, H.(2011). "Threshold models in time series analysis —30 years on (with discussions by P.Whittle, M.Rosenblatt, B.E.Hansen, P.Brockwell, N.I.Samia & F.Battaglia)". Statistics & Its Interface, 4, 107-136
- Tong, H.& Lim, K. S. (1980) "Threshold Autoregression, Limit Cycles and Cyclical Data (with discussion)", Journal of the Royal Statistical Society, Series B, 42, 245-292
- Eigene Recherche

Betreuer: Sebastian Heiden