Wien, am 9. Dezember 2020

260014 (Uni Wien), 142.340 (TU Wien), “Statistische Methoden der Datenanalyse”

Alternative Prüfungsmodalitäten Wintersemester 2020/21

Aufgrund der Pandemie ist es dieses Semester optional [!] gestattet, anstatt die Klausurprüfung abzulegen, ein längeres Anwendungsbeispiel abzugeben. Das Anwendungsbeispiel ist inklusive Programmcode einzureichen. Ein circa 3-6 seitiges Dokument ist hinzuzufügen, das die Aufgabenstellung beschreibt, den Code dokumentiert und die Ergebnisse inklusive graphischer Illustrationen detailliert präsentiert. Das Programm muss Bezug nehmen auf den Vorlesungsstoff (oder darüber hinausgehen), dieser Bezug muss aus der Dokumentation klar hervorgehen. Die Programmiersprache muss nicht Python sein. Assembly, Basic, Visual Basic, Fortran und Cobol sind jedoch explizit nicht erlaubt.

Der Code darf explizit innerhalb einer Bachelor-, Masterarbeit oder Dissertation zur Verwendung kommen, darf allerdings nicht Inhalt einer anderen Vorlesung oder Übung sein. Die Arbeit muss signifikant den Komplexitätsgrad einzelner Übungsbeispiele übersteigen. Falls erwünscht, gebe ich gerne Feedback zu einer Arbeit. Eine Überarbeitung der Arbeit nach der Feedbackrunde ist erlaubt und explizit erwünscht. Instruktive Visualisierungen der Ergebnisse werden besonders positiv gewertet. Zusätzlich muß die Arbeit in einem kurzen Video (z.b. per ZOOM aufgenommen) Pflicht. Gebt bitte im Video euren Namen, Matrikelnummer und Universität bekannt. Arbeiten zu zweit sind erlaubt; in diesem Falle muss jedoch der Aufwand signifikant (50%) höher sein als im Falle von Arbeiten allein.

Spätestmögliches Abgabedatum: 15. Februar 2021.

Ich hoffe, mit dieser alternativen Prüfungsmodalität manchen von Ihnen ein attraktives Angebot zu machen,

MfG

Wolfgang Waltenberger

Themenvorschläge für die Prüfungsarbeit, “Statistische Methoden der Datenanalyse” (SV 142.340, VU 260014)

Thema 1: CIA World Factbook (Deskriptive Statistik, Hauptkomponentenanalyse, Regression)
==========

- Geht auf https://www.cia.gov/library/publications/the-world-factbook/
- Extrahiert die Daten für alle (circa 200) Länder
- Definiert einige interessante Merkmale die euch interessieren, z.B. Kaufkraftparität, GINI Index, Anzahl der Einwohner, Breitengrad, Lebenserwartung.
- Fasst die Daten mit den üblichen Lagemaßen zusammen. Gibt es Korrelationen, wenn ja welche und wie groß? Kann man eine Hauptkomponentenanalyse anwenden?
- Gibt es eine sinnvolle Regression zweier oder mehrerer Parameter, die anwendbar ist?
- Könnt ihr einen sinnvollen Chi-Quadrat Test für manche der Größen formulieren?

Thema 2: US Election Turnout Rates (Regression, Chi-Quadrat Test)
==========

- Geht auf https://www.kaggle.com/imoore/2020-us-general-election-turnout-rates
- Ladet den Datensatz herunter
- Macht explorative Datenanalyse, plottet alle möglichen Merkmale gegen andere?
- Gibt es Größen die die Turnout Rate offenbar nicht beeinflussen? Formuliert den Hypothesentest aud diese Aussage, z.B. über den Chiquadrat Test.
- Gibt es Größen, die linear oder polynomial einzugehen scheinen? Regrediert die Turnout Rate gegen diese Größen.

Thema 3: What makes us happy? (Freies Thema)
==========

- Geht auf https://www.kaggle.com/unsdsn/world-happiness
- Findet heraus, was Menschen glücklich macht.

Thema 4: kaggle datasets (allgemeine Aufgabenstellung)
==========

- Geht auf https://www.kaggle.com
- Wählt einen Datensatz aus, der euch interessiert
- Diskutiert ihn, ähnlich wie obige Aufgabenstellungen

Thema 5: Der Konservatismus der LHC PhysikerInnen (likelihoods, p-Werte)
==========

Am LHC wird nach neuer Physik gefahndet, indem nach Proton-Proton Kollisionen (“Events”) mit bestimmten Eigenschaften gesucht wird.
Die Ergebnisse der Suchen nach neuer Physik werden euch gegeben als:
Anzahl der Events die wir von der uns bekannten Physik *erwarten*
Der normalverteilte Fehler auf diese Anzahl der Events
Die tatsächlich beobachtete Anzahl der Events

Berechnet für hunderte von Suchen die p-Werte. Modelliert dafür die likelihood als Produkt von Gauss * Poissonverteilung. Plottet diese p-Werte. Ihr werdet erkennen, dass es zu wenige “extreme” Ergebnisse gibt.
Mit welchem Faktor ist der Fehler zu multiplizieren, um realistischere Werte zu bekommen?
Daten auf Anfrage. Dieses Beispiel verlangt ein wenig tieferes Verständnis der Physik, deswegen bin ich (WW) gerne bereit hier etwas mehr Hilfestellung zu leisten.

Thema 6: Visualisierung und Beschreibung von Events in Dunkle Materie Detektoren
==========

- Herausforderungen: Verstehe und verwende die PCA und t-SNE Funktionen der Scikit-Learn Library korrekt und interpretiere die Resultate. Visualisiere Ergebnisse interaktiv mit der Plotly Library.
- Ein kleines, gelabeltes Datenset der Parameter getriggerter Events aus Dunkle Materie Detektoren wird im csv Format bereitgestellt
- Es soll darauf PCA und t-SNE angewandt werden (https://en.wikipedia.org/wiki/T-distributed_stochastic_neighbor_embedding). Es können Funktionen der sklearn Library verwendet werden, die Algorithmen müssen nicht selbst geschrieben werden.
- Qualitative Beschreibung der Ergebnisse: Welche Klassen von Events clustern, welche nicht? Können Gründe dafür identifiziert werden? Interaktive Visualisierung mit der Plotly Library
- (Optional) Sampling und Analyse von Fake Event-Parameter durch Kernel Density Estimation

Thema 7: Programmiere einen Decision Tree zur Klassifizierung von Daten
==========

- Herausforderungen: Implementieren eines fortgeschrittenen Algorithmus in Python, bzw Verstehen und Adaptieren von bestehendem Code (keiner Library Funktion) auf ein anderes Datenset.

- Für den Algorithmus eines einfachen Decision Tree gibt es viele Tutorials online, e.g. https://machinelearningmastery.com/implement-decision-tree-algorithm-scratch-python/
- Schreibe einen einfachen Decision Tree Algorithmus in Python, ohne eine vorgefertigte Library Funktion dafür zu verwenden (e.g. kein Scikit-Learn). Mathematik-Libraries, e.g. Numpy, dürfen selbstverständlich verwendet werden.
- Vorschlag für Dataset: Identifizierung von Minen vs. Gestein durch Echolot Signalen. https://archive.ics.uci.edu/ml/datasets/Connectionist+Bench+(Sonar,+Mines+vs.+Rocks) Es kann auch ein anderes Datenset verwendet werden, allerdings nicht das “Banknote Dataset” aus dem Tutorial ;)+

Thema 8: Klassifizierung von handgeschriebenen Ziffern mit Gaussian Mixture Models
==========

- Herausforderungen: Verstehe und verwende die Gaussian Mixture Models Funktion der Sklearn Library an einem Standard-Datenset.
- MNIST Dataset: http://yann.lecun.com/exdb/mnist/
- Für Gaussian Mixture Models gibt es viele Tutorials online, e.g. https://www.python-course.eu/expectation_maximization_and_gaussian_mixture_models.php
- Klassifiziere die handgeschriebenen Ziffern des MNIST-Dataset mit Gaussian Mixture Models. Wie hoch ist dein Accuracy Score? Vergleiche den Score mit anderen Ergebnissen, die du online findest, und versuche zu erklären warum dein Score besser/schlechter ist als die anderer Algorithmen.

Thema 9: “Upper Limit” für Dunkle Materie mit Maximum Gap-Methode
==========

- Herausforderungen: Selbstständig einen Algorithmus aus einem Paper extrahieren und in Python implementieren. Vorwissen über Teilchenphysik und fortgeschrittene Programmierkenntnisse empfohlen, aber nicht unbedingt nötig.
- Die Maximum Gap Methode nach Yellin ist eine Standardmethode um eine Obere Schranke für die Masse und den Streuquerschnitt von hypothetischen Dunkle Materie Teilchen aus dem Energiespektrum eines Experiments zu bestimmen. Eine der beiden Methoden soll in Python implementiert werden.
- Ein Energiespektrum wird im *.txt Format bereitgestellt.
- Paper: https://arxiv.org/abs/physics/0203002
- Mit etwas Googeln sind Implementierungen in GitHub zu finden. Insofern mit den Urheberrechtlichen Bestimmungen der Quellen vereinbar, darf Code aus dieses Quellen verwendet werden.

Thema 10: Allgemeines Data Science Projekt
==========

- Auf https://data-flair.training/blogs/machine-learning-datasets/ gibt es eine Liste von Datasets und vorgeschlagenen Projekten - finde ein Dataset und Projekt und setze es um!

- Falls die vorgeschlagenen Projekt zu Machine Learning-lastig sind, können beliebige Adaptierungen der Aufgabenstellungen vorgenommen werden. Bei einem interessanten Datenset ist auch die ausführliche quantitative und qualitative Beschreibung der Daten und Korrelation, inkl. Visualisierungen, durchaus spannend.