"Mathematics of Data science", oder: Mathematische Verfahren für hochdimensionale Daten WS 2023/24[1]

Matthias Birkner[2]

https://www.staff.uni-mainz.de/birkner/MoDS2324/

Mögliche Vortragsthemen

1. **Fluch und Segen der hohen Dimensionen**

   Geometrische Einsichten (und vielleicht Überraschungen) über (Einheits-)Kugel und Würfel im $\mathbb{R}^d$ mit $d \gg 1$, dazu einige (relativ elementare) Konzentrationsungleichungen für reelle Zufallsvariablen

   1 Vortrag oder 2 Vorträge, [BSS, Ch. 2]; falls 2 Vorträge, so sollten auch Konzentrationsungleichungen für zufällige Matrizen aufgegriffen werden, [BSS, Ch. 6], siehe auch [B16, Ch. 4]

2. **Singulärwertzerlegung und Hauptkomponentenanalyse, ggf. auch das "spike problem"**

   Hauptkomponentenanalyse zeigt auf, in welchen (linearen) Richtungen ein gegebener (hochdimensionaler) Datensatz die größte Variabilität enthält und liefert so einen Ansatz zur Dimensionsreduktion; das "spike problem" untersucht die Frage, wie gut damit ein niedrigdimensionales „Signal" in einer großen Zufallsmatrix wiedergefunden werden kann (vgl. auch die Vorstellungsfolien).

   1 Vortrag oder 2 Vorträge; es sollen dabei auch einschlägige Begriffe und Sätze der linearen Algebra (kurz) wiederholt werden, [BSS, Ch. 3], siehe auch [B16, Ch. 1]; falls 2 Vorträge, so soll das spike problem detailliert(er) behandelt werden, ggf. unter Rückgriff auf weitere Literatur

3. **Graphen, Netzwerke und Clusterverfahren**

   Es geht um Verfahren, Datenpunkte nach „Ähnlichkeit" zu gruppieren: k-means-clustering minimiert dazu Abstände zu geeignet zu definierenden „Gruppenmittelpunkten"; spectral clustering nimmt Ideen aus der Graphentheorie zu Hilfe

   1 Vortrag, es werden dabei auch einige Grundbegriffe der Graphentheorie eingeführt, [BSS, Ch. 4], siehe auch [B16, Ch. 3]; das Thema kann ggf. mit dem darauffolgenden (Diffusionsabbildungen) zu einen Tandem verknüpft werden

4. **Diffusionsabbildungen und nicht-lineare Dimensionsreduktion**

   Diffusionsabbildungen ("diffusion maps") sind ein (weiteres) Verfahren zur Gruppierung und Dimensionsreduktion hochdimensionaler Daten, das spectral clustering verallgemeinert: man fasst die Datenpunkte als Knoten eines Graphs und ihre paarweisen Abstände als Kantengewichte auf und nutzt Eigenschaften der Irrfahrt auf diesem Graphen aus

   1 Vortrag, [BSS, Ch. 5], siehe auch [B16, Ch. 2.2–2.3]; das Thema kann ggf. mit dem vorherigen (Clusterverfahren) zu einen Tandem verknüpft werden

5. **Dimensionsreduktion via zufällige Projektionen: Johnson-Lindenstrauss-Lemma und Gordons Theorem**

   Durch gut gewählte Projektionen lässt sich die Dimension von Datenpunkten so reduzieren, dass ihre paarweisen Abstände höchstens um einen Faktor $(1 \pm \varepsilon)$ verzerrt werden.

   1 Vortrag oder 2 Vorträge, [BSS, Ch. 9], siehe auch [B16, Ch. 5]

b.w.

[1] Stand: 19.7.2023

[2] birkner@mathematik.uni-mainz.de

6. **Stochastisches Block-Modell und konvexe Relaxierung**

   In einem zufälligen Netzwerk gibt es zwei (sagen wir, gleich große) Gruppen von Knoten und zufällig gewählte Kanten, wobei Kanten zwischen zwischen Mitgliedern derselben Gruppe wahrscheinlicher sind. Können wir anhand des beobachteten Verbindungsmusters entscheiden, welcher Knoten zu welcher Gruppe gehört? Das Problem stets exakt zu lösen ist sehr schwer, Verfahren der semidefiniten Optimierung finden aber mit hoher Wahrscheinlichkeit das Richtige.

   1 Vortrag oder 2 Vorträge, [BSS, Ch. 8], siehe auch [B16, Ch. 9]; bei 2 Vorträgen soll in mehr Detail auf die Optimierungsverfahren eingegangen werden

7. **Compressive sensing**

   Wie gut kann man „hochdimensionale“ Informationen aus „niederdimensionalen“ (linearen) Messungen rekonstruieren, wenn man zusätzlich weiß (oder annimmt), dass die „Wahrheit“ dünnbesetzt ist (also hauptsächlich 0-Einträge hat)?

   1 Vortrag oder 2 Vorträge, [BSS, Ch. 10], siehe auch [B16, Ch. 6]

8. **Approximation maximaler Schnitte**

   Die Knotenmenge eines Graphen mit gewichteten Kanten soll so in zwei Teile zerlegt werden, dass die Summe der Gewichte der Kanten, die Knoten in unterschiedlichen Teilen verbinden, maximal wird (eine ähnliche Frage kommt beim Thema Clusterverfahren vor). Eine exakte Lösung ist im Allgemeinen sehr schwer zu finden, Verfahren der semidefiniten Optimierung können Approximationen liefern.

   1 Vortrag, [BSS, Ch. 7]

Anmerkungen

- Es kann gelegentlich notwendig sein, über das genannte Buchmanuskript [BSS] hinaus auf weitere Quellen, gegebenenfalls auch auf die zitierte Originalliteratur, zurückzugreifen.
- Wie immer: Wir möchten alle Vorträge verstehen, nicht nur unseren eigenen.
- Rechtzeitig mit der Vorbereitung beginnen und zur Vorbesprechung zu mir kommen!

  Typischerweise (mind.) zwei Termine: erste Besprechung (mind. 6 Wo. vor Vortrag): Text gelesen und durchgearbeitet, etwaige Fragen (die nicht vorher durch eigenes Nachdenken, Literaturrecherche oder Diskussion mit Komilitonen geklärt werden konnten) klar herausdestilliert,

  zweite Besprechung (mind. 2 Wo. vor Vortrag): etwaige Detailklärung(en), plausiblen Entwurf des Vortrags und des Handouts vorlegen

- Manfred Lehn hat einen lesenswerten Text zur Frage „Wie halte ich einen Seminarvortrag?“ verfasst, siehe `https://download.uni-mainz.de/mathematik/Topologie%20und%20Geometrie/Arbeiten/Allgemeine%20Texte/Wie%20halte%20ich%20einen%20Seminarvortrag.pdf`

Literatur

[BSS] Afonso S. Bandeira, Amit Singer, Thomas Strohmer, Mathematics of Data Science, Book in preparation, 2023. `https://people.math.ethz.ch/~abandeira/BandeiraSingerStrohmer-MDS-draft.pdf`

[B16] Afonso S. Bandeira, Ten Lectures and Forty-Two Open Problems in the Mathematics of Data Science. Lecture notes, October 10, 2016. `https://people.math.ethz.ch/~abandeira/TenLecturesFortyTwoProblems.pdf`