3.Sitzung

09.07.2021

Imbalanced-Data

Welches ist die passende Metrik?

1. Too good to be true -> imbalanced data bedarf ein paar spezieller Überlegungen
2. Wiederholung der wichtigsten Metriken
3. AUC

Eine kurze Überlegung zum Test-Set

4. imbalanced data sollte kein zufälliges Test-Set haben

Strategien für imbalanced data

5. Oversample Minority-class: SMOTE, etc
6. Undersample Majority-class
7. do both: oversample Minority-class and undersample Majority-class
8. gewichte Minority-class stärker
9. behandle Minority-class wie outlier

Kernel Methods

Support-Vektor-Machine

10. Verstehe die charakteristische Gleichung: Warum skalieren SVMs schlecht?

Kernel-Trick

11. Unterschiedliche Klassen sind in höheren Dimensionen trennbar
12. Wirkliche Projektion in höhere Dimensionen
13. Implizite Projektion in höhere Dimensionen

Kernel als Similarity-Measure

14. Verstehe die beiden Sichtweisen: non-lineares Ähnlichkeits-Mass vs. implizite Projektion und euklidsche Distanz
15. einfache Kernel-Regression mit Kernels zur Bestimmung der neighborhood
16. Dot-Product als Ähnlichkeits-Mass: cosine-similarity

Auto-ML

Hyper-Parameter-Tuning

17. Nachteile von Algorithmen mit vielen Parametern verstehen
18. Bayesian Optimization: surrogate function, acquisition function, objective function
19. Auto-Sklearn / SMAC als ein Beispiel für Hyper-Parameter-Tuning
20. Verstehe die wichtigsten Parameter von Auto-Sklearn

Brute-Force

21. autogluon.tabular als brute-force approach
22. Wiederhole Stacking, Bagging, Catboost

Genetic Programming

23. TPOT als ein Vertreter von Genetic Programming

Ungefähre Abschätzung der Möglichkeiten der Ansätze

24. Verschiedene Ansätze von AutoML bewerten können

Clustering

Ungenauigkeit von Cluster-Lösungen

25. Optimale Anzahl von Clustern unbekannt
26. Distanz-Basierende Verfahren sind anfällig auf unterschiedliche Skalierung
27. viele verschieden Cluster-Lösungen

Spectral-Clustering als Graph-Ansatz

28. Eigenwert-Ansatz mit Hilfe der Laplacian
29. Tf-Idf wiederholen; normierte Vektoren
30. cosine-similarity ist dot-product ist Korrelationskoeffizient r
31. approximate-nearest-neighbors zur Konstruktion des symmetrischen Graphen

DB-Scan als Dichte-basiertes Verfahren

32. Ungefähres Verständnis dieses Ansatzes