ABSCHNITT 4

DIE MACHINE LEARNING BASICS

GRUNDLEGENDE ML ALGORITHMEN KENNENLERNEN

Alina Herderich

IN DER NÄCHSTEN STUNDE

Bitte findet euch in Gruppen von zwei bis drei Personen zusammen und versucht, den euch zugeteilten Machine Learning Algorithmus zu erarbeiten, indem ihr online recherchiert und ihn für die anderen in einer kurzen Präsentation aufbereitet.

FOLGENDE ORIENTIERUNGSFRAGEN GEBEN EUCH HILFESTELLUNG (1/2):

Ist das Verfahren supervised oder unsupervised?

Für welchen Task wird das Modell angewendet?

Welche Form haben die Input Daten?

Welche Form hat der Output des Modells?

Wie funktioniert der Algorithmus generell?

Was wird optimiert (cost function)?

Welche Annahmen trifft der Algorithmus?

FOLGENDE ORIENTIERUNGSFRAGEN GEBEN EUCH HILFESTELLUNG (2/2):

Wie kann man die Güte der Vorhersagen messen (evaluation metric)?

Wann könnte der Algorithmus nicht gut funktionieren?

Welche Lösungen gibt es dafür?

Welche Anwendungsfälle gibt es?

Könnt ihr euch Beispiele für die Psychologie ausdenken?

Welche packages stehen in Python für den Algorithmus zur Verfügung?

Welche Parameter können da gesetzt werden und was steckt dahinter?

UNSERE FÜNF ALGORITHMEN

- Logistic Regression
- Support Vector Machines
- Decision Trees
- k-means
- Principal Component Analysis

1 // LOGISTIC REGRESSION

Ein Klassifikationsverfahren

Welche Arten logistischer Regression gibt es?

In welchem Zusammenhang steht logistische mit linearer Regression?

Welche Unterschiede gibt es?

Was ist die Sigmoid Funktion? Wie heißt sie noch?

Was ist das Maximum Likelihood Estimation Verfahren?

2 // SUPPORT VECTOR MACHINES (SVM)

Ein Klassifikationsverfahren

Was ist ein support vector?

Was ist eine Hyperebene (hyperplane)?

Was bedeutet linear teilbare (linearly separable) vs. nicht linear teilbare (non-linearly separable) Daten?

Was ist ein kernel?

Was ist ein Regularisierungsparameter?

3 // DECISION TREES

Ein Klassifikationsverfahren

Was sind Knoten (nodes), Pfade (edges) und Blätter (leafs)?

Was ist der Gini coefficient, entropy, information gain?

Für stehen CART und ID3?

Was ist early stopping, pruning?

Fortgeschrittene Decision Tree Algorithmen: Was ist ein random forest, bagging, boosting?

4 // K-MEANS

Ein Clusteringverfahren

Wofür steht k-means?

Was ist ein centroid?

Was ist inter-cluster vs. intra-cluster Distanz?

Was ist der Euklidische Abstand und welche anderen Distanz- oder Ähnlichkeitsmaße gibt es noch?

Was ist die Ellenbogen Methode, der silhouette coefficient?

5 // PRINCIPAL COMPONENT ANALYSIS (PCA)

Ein Dimensionsreduktionsverfahren

Was versteht man unter dem Fluch der Dimensionalität?

Was ist eine Kovarianzmatrix?

Was ist ein Eigenvektor, eine Hauptkomponente?

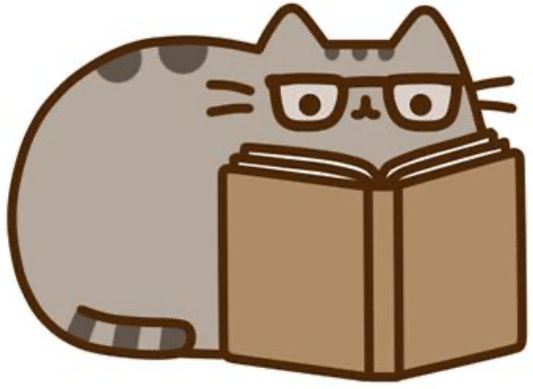
Was ist ein Eigenwert und wie verhält er sich zur Varianz in den Daten?

Was ist ein Scree Plot?

ES GILT:

1. Startet mit meinen Empfehlungen oder beginnt eure eigene Reise.
2. Die allgemeinen Fragen beziehen sich auf alle Algorithmen - immer.
3. Die spezifischen Fragen geben euch Hinweise auf Schlagworte, auf die ihr bei eurer Recherche achten solltet.
4. **Ihr müsst nicht alle Fragen beantworten!** Erkundet, was euch am meisten interessiert oder recherchiert einfach gerade heraus.
5. Ressourcen in Textform sind toll, aber Videos sind manchmal nützliche Visualisierungen, um zu verstehen, was passiert.

ALLE INFORMATIONEN FINDET IHR IM DAZUGEHÖRIGEN GOOGLE DOC (LINK IM GITHUB REPOSITORY).



VIEL ERFOLG!