EINFÜHRUNG IN DAS MASCHINELLE LERNEN – L4

PRINCIPAL COMPONENT ANALYSIS UND CLUSTERING

Bekommen in L4

Abgabe in L5

Benutze das gereinigte Ames Housing Dataset, mit passendem encoding für die kategorischen Features und skaliere den Datensatz mit min-max scaling. Siehe auch link, link

Benutze PCA, um den Datensatz auf 98% der Varianz zu reduzieren. Wie viel von der Varianz des Datensatzes erklärt jede Principal component (siehe hier - explained_variance_ratio)? Mit wie vielen Features bleibt man? Wie viele Principal Components sind nötig wenn man 95%, 90%, 80% der Varianz behalten will? Zeichne die kumulative Varianz als Funktion der Anzahl von Principal Components.

Mit dem auf mehrere möglichen Schwellen der Varianz reduzierten Datensatz trainiere je ein k-Means Clustering Modell. Benutze die Ellenbogenmethode um die optimale Anzahl der Cluster zu finden. Zeichne das Plot. Zeichne Silhouettenplots für mehrere Anzahlen von Clusters. Erstelle auch ein scatter plot, wo die ersten 2 principal components als Punkte sichtbar sind. Färbe die Punkte unterschiedlich, je nach Cluster. Gibt es Unterschiede? Erkläre.

Versuche den Einfluss der Features auf die wichtigsten zwei principal components zu sehen. Benutze `loadings = pca.components_.T * np.sqrt(pca.explained_variance_)`. Repräsentiere das grafisch.

Erkläre in eigenen Worten, welche Eigenschaften jedes Cluster hat. Was für Häuser findet man tendenziell in Cluster x? Benutze, z.B. Visualisierungen aus L1.

Gib dem Modell einige ihm unbekannte Datenpunkte. Erfinde selbst ein paar plausible Häuser. Welchem Cluster werden sie zugeordnet? Ist es, intuitiv, richtig oder falsch? Erkläre. Hinweis: vor der Inferenz muss man den Datenpunkt mithilfe vom PCA transformieren – warum?

Bonus:

a. Benutze für die Skalierung der Daten sowohl min-max-Skalierung, als auch Normalisierung. Macht das Logarithmieren von Features (wo sinnvoll) einen Unterschied? Wie unterscheiden sich die Ergebnisse? Versuche, eine **Erklärung** zu finden.
b. Zeichne **interaktive** 3D Scatter Plots (zoomen und Rotieren) mit den ersten 3 principal components, verschieden gefärbt, für jedes Clustering. Zeichne, mit einem anderen Symbol, auch die jeweiligen Zentroide.