Moderne Ansätze des Machine Learning: Von neuronalen Netzen bis zur Bayesianischen Statistik (Seminar)
Vorschläge für Datensätze

Prof. Dr. Michael Merz
Nha-Nghi de la Cruz
Marie Hielscher
Max Lüdecke
Jan Rabenseifner

March 6, 2025

Dataset Descriptions

1. **London Crime Dataset**
   Description: This dataset contains crime data in London, broken down by borough, major category, minor category, and year/month. It provides insights into crime trends across different areas of London over time.
   Task: Time series analysis (crime trends over time) or classification (predicting crime categories).
   Link: `https://www.kaggle.com/datasets/jboysen/london-crime`

2. **San Francisco Crime Classification (SF Crime)**
   Description: This dataset contains crime incidents in San Francisco, including details like category, location, and time. The goal is to predict the category of crime based on the given features.
   Task: Classification (predicting crime categories).
   Link: `https://www.kaggle.com/c/sf-crime`

3. **Santander Customer Transaction Prediction**
   Description: This dataset contains anonymized banking data, where the goal is to predict whether a customer will make a transaction in the future.
   Task: Binary classification (predicting if a transaction will occur).
   Link: `https://www.kaggle.com/c/santander-customer-transaction-prediction/overview`

4. **Santander Product Recommendation**
   Description: This dataset contains customer banking data, and the goal

is to recommend products to customers based on their transaction history and behavior.
Task: Classification (recommending products).
Link: https://www.kaggle.com/c/santander-product-recommendation/data

5. **Amazon and Best Buy Electronics**
Description: This dataset contains product data from Amazon and Best Buy, including product names, prices, and reviews. It can be used for price comparison, sentiment analysis, or product recommendation.
Task: Regression (price prediction), classification (sentiment analysis), or recommendation systems.
Link: https://www.kaggle.com/datasets/datafiniti/amazon-and-best-buy-electronics

6. **Favorita Grocery Sales Forecasting**
Description: This dataset contains historical sales data from a grocery store chain in Ecuador. The goal is to forecast future sales based on historical trends, promotions, and other factors.
Task: Time series forecasting (sales prediction).
Link: https://www.kaggle.com/competitions/favorita-grocery-sales-forecasting/data

7. **Zillow Economics Data (ZECON)**
Description: This dataset contains housing market data from Zillow, including home values, rent prices, and economic indicators. It can be used to analyze trends in the housing market.
Task: Regression (predicting home prices or rent) or time series analysis (housing market trends).
Link: https://www.kaggle.com/datasets/zillow/zecon

8. **Inventory Product Demand Forecasting**
Description: This dataset contains product demand data, which can be used to forecast future demand for inventory management.
Task: Time series forecasting (demand prediction).
Link: https://github.com/bkkinfo/Inventory-Product-Demand-Forecasting-ML-

9. **Adult Income Dataset (UCI)**
Description: This dataset contains demographic data, including age, education, occupation, and income. The goal is to predict whether an individual earns more than $50,000 per year.
Task: Binary classification (income prediction).
Link: https://archive.ics.uci.edu/dataset/2/adult

10. **Lending Club Dataset**
    Description: This dataset contains loan data from Lending Club, including loan amounts, interest rates, and borrower information. It can be used to predict loan defaults or analyze lending trends.
    Task: Classification (predicting loan defaults) or regression (predicting interest rates).
    Link: `https://www.kaggle.com/datasets/wordsforthewise/lending-club/data`

11. **PolData Repository**
    Description: This repository contains datasets related to political science and economics, including election results, policy data, and various political indicators. It serves as a source for alternative datasets. Before retrieving a dataset from this platform, please consult with your assigned supervisor.
    Task: Time series analysis (political/economic trends) or regression (predicting political outcomes).
    Link: `https://github.com/erikgahner/PolData`