Large Individual Project

Vladimir J N Bykov

Agenda

- Målet med projektet
- Vad är det som krävs
- EXTRA utmaningar om ni vågar

Stora Individuella Projektet

Vad ni skall göra?

Demonstrera att ni kan hantera alla steg i machine learning:

- **Dataförberedelse:** Hantera saknade värden, kategoriska variabler och skapa en dataset redo för analys.
- **EDA, statistik:** Utforska datasetet genom visualiseringar och statistiska metoder för att identifiera mönster och viktiga variabler.
- **Modellering:** Identifiera problemet som kan lösas med hjälp av ML algoritmer, välj och träna lämpliga algoritmer för att lösa problemet. Finns det målvariabel? Om inte kanske ni kan skapa den med hjälp av unsupervised learning.
- **Utvärdering:** Använd lämpliga metoder som Confusion Matrix, Classification Report, AUC-ROC och precision-recall för att mäta prestandan.

Mål med Projektet

Demonstrera den bästa och den mest komplexa tekniken som ni kan (vågar använda) i machine learning:

- **Detta projekt ni kommer att använda på era jobb intervjuer för att briljera i era kunskaper och färdigheter i machine learning.**

Vad ni skall göra?

- Visa att ni kan förstå data eller generera hypotes med hjälp av Unsupervised learning.
- Ni kan använda datareduktion (PCA, UMAP, etc) om det behövs i modell utveckling. Det kan vara viktig om dator har svårt att hantera nödvändiga beräkningar.
- Visa att ni kan hantera olika algoritmer i **Supervised Learning** inkluderar **Deep Learning**. Ni kan bygga neurala nätverk som kan hantera både strukturerad data och ostrukturerad data (som bilder, text, ljud, etc).
- **OBS! Ni kan testa flera olika algoritmer, men inte enbart Linjär eller Multiple Linjär Regression.**

Data source: example

Kaggle

UCI Machine Learning Repository

En av de mest populära och omfattande källorna för maskininlärningsdatamängder. Här hittar du dataset för alla typer av maskininlärningsproblem, inklusive klassificering, regression, och klustring.

Google Dataset Search

Google Dataset Search är en sökmotor för dataset som gör det enkelt att hitta dataset från olika källor på webben. Den samlar information från många öppna datakällor.

AWS Open Data Registry

Amazon Web Services (AWS) erbjuder en samling offentliga dataset inom en mängd olika ämnen, såsom geospatial data, genetik, ekonomi och mer. Det är en bra resurs för stora dataset som kan användas direkt på molnbaserade plattformar.

Data.gov

Data.gov är den amerikanska regeringens portal för offentliga dataset. Du hittar dataset relaterade till olika ämnen som klimat, hälsa, utbildning och ekonomi.

Quandl

Quandl är en plattform som tillhandahåller finansiella, ekonomiska och sociala dataset. Den är särskilt användbar för användare som vill arbeta med ekonomiska modeller eller tidsseriedata.

Google BigQuery Public Datasets

Google Cloud erbjuder en samling offentliga dataset som är lagrade i BigQuery. Det finns dataset inom många områden, inklusive ekonomi, hälsa, och sport. BigQuery-plattformen gör det möjligt att hantera och analysera mycket stora dataset snabbt.

OpenML

OpenML är en plattform för delning och analys av maskininlärningsdata. Den erbjuder tusentals dataset för olika maskininlärningsuppgifter och inkluderar även verktyg för att analysera algoritmernas prestanda på dessa dataset.

DrivenData

DrivenData erbjuder datavetenskapstävlingar med ett fokus på samhällsproblem, vilket gör den till en intressant plattform för att arbeta med dataprojekt som har ett syfte. Här hittar du dataset för både klassificering och regression.

Zenodo

Zenodo är en forskningsdataplattform som tillhandahåller dataset från många olika forskningsprojekt. Det är en bra resurs för att hitta nischade dataset inom vetenskap och forskning.

University Repositories and Public Data Portals

Många universitet har egna datalager som tillhandahåller forskningsdataset. Exempel på detta är MIT, Harvard, och Stanford.

Figure Eight (tidigare CrowdFlower)

Figure Eight har en rad annoterade dataset som har använts för maskininlärningsprojekt, särskilt inom textanalys och bildigenkänning.

EXTRA

- Det finns inga gräns för kreativitet
- Använd interaktiva visualiseringar, dashboards
- Ni kan gå hela vägen upp till model deployment/ distribution med verktyg som Streamlit, Dash, React.js/ Flask

Leka med data, njut av spelet

Tack för er tid