Skitser til forløb om modellering

1 Avanceret dataanalyse

1.1 Hvad bliver gennemgået

Dataanalyse bruges i dag i stort set alle sammenhænge, både i erhvervslivet og i forskningsverden. Med dette forløb vil eleverne få kendskab til og hands-on-erfaring med nogle af de allermest brugte analysemetoder.

Emnerne for forløbet vil være:

Beregning af usikkerhed

Ved hjælp af Monte Carlo-simulationer vil eleverne lære, hvordan man kan beregne usikkerheden på et resultat, hvis målingerne, der indgår i beregningen, ikke er bestemt helt nøjagtigt. Et simpelt eksempel er at beregne volumen af en kasse, hvor siderne er målt ved hjælp af en lineal, og som derfor ikke er helt nøjagtige. Hvor stor vil usikkerheden/unøjagtigheden på det beregnede volumen være?

Minimeringsteknikker

Minimering handler om at finde minimum af en funktion, når det ikke er muligt at finde en analytisk løsning. Minimering bruges i mange sammenhænge, f.eks. til at finde tendenslinjer, hvor man minimerer afstanden mellem tendenslinjen og datapunkterne ved at ændre værdien af variablene i funktionsforskriften. Eleverne vil stifte bekendtskab med nogle minimeringsteknikker, som kan bruges i mange andre sammenhænge end bare til at finde tendenslinjer.

Korrelationer

Eleverne vil lære at bestemme korrelationer mellem data, dvs. om to parametre afhænger – eller ser ud til at afhænge – af hinanden. Det kan eksempelvis bruges til at undersøge, om mængden af kriminalitet følger arbejdsløsheden, tid på året osv. Eleverne vil også lære at fortolke korrelationer og se, hvornår det er muligt at finde tilsyneladende stærke korrelationer, uden at de faktisk betyder noget.

1.2 Skitse af forløb

Forløbet strækker sig over tre uger og er tænkt til at kombinere et it-fag med et andet fag. Den mest oplagte kombination vil være med et naturvidenskabeligt eller samfundsvidenskabeligt fag, da det skal være muligt at kunne finde data at arbejde med.

Forløbet afsluttes med et mindre tværfagligt projekt, hvor eleverne diskuterer de lærte metoder ud fra et it-mæssigt synspunkt, samtidig med at metoderne anvendes på dataene fra kombinationsfaget, og resultaterne diskuteres i fagets kontekst.

Uge 1

It-fag: Begynde at bruge python. Eleverne skal lære de mest basale ting såsom variable, løkker, if-else, skrive simple funktioner, samt lave basale plots.
Kombinationsfag: Gennemgå typiske dataanalysemetoder, f.eks. tendenslinjer i Excel, gennemsnit, standardafvigelser osv. Det præcise emne, der vil undersøges, fastlægges.

Uge 2

It-fag: Rejseholdet kommer og giver en lektion i avanceret dataanalyse. Se emnerne ovenfor.
Kombinationsfag: Data indhentes eller optages, hvis der laves et forsøg. Analyse af dataene ved hjælp af de gennemgåede metoder påbegyndes.

Uge 3

It- og kombinationsfag: Dataanalyse færdiggøres og projektrapporten udarbejdes.

2 Machine learning

2.1 Hvad bliver gennemgået

Eleverne vil få en introduktion til machine learning, som er en betegnelse for programmer, der kan træne sig selv til at blive bedre til opgave på baggrund af tidligere erfaring – dvs. en form for kunstig intelligens. Et typisk eksempel på machine learning er et effektivt spam-filter. Her har et program selv analyseret en masse tidligere spam-mails, og kan på den baggrund klassificere en ny mail som spam/ikke-spam, også selvom den nye mail er meget forskellig dem, programmet har set før.

Emnerne for forløbet vil være:

Regression

Regression handler om at finde en sammenhæng mellem datapunkter og på den baggrund forudsige værdier for nye punkter. Når man finder tendenslinjer i f.eks. Excel, bruger man faktisk regression, men man skal selv specificere, hvilken funktion, man vil tilpasse. Med machine learning finder computeren selv en god funktion – eller holder sig til at lave forudsigelser udelukkende på baggrund af dataene, uden at antage nogen form for matematisk sammenhæng. Eleverne vil prøve at arbejde med begge typer regressionsmetoder.

Klassifikation

I klassifikation bestemmer man ikke talværdier, men prøver i stedet at bestemme klasser eller kategorier for data. Et eksempel er spam-filteret, der klassificerer ny mail som enten "spam" eller "ikke-spam". Et andet eksempel er at bestemme en diagnose for en patient på baggrund af patientens temperatur, blodtryk, køn osv. Eleverne vil stifte bekendtskab med en letforståelig, men enormt effektiv, metode til klassifikation.

Klyngeanalyse

Nogle gange findes der klynger eller grupper af datapunkter. Det kan f.eks. være data om kunder i en bestemt virksomhed, hvor grupperne vil være kunder, der har samme interesser eller indkøbsmønstre. Eleverne vil prøve at arbejde med en metode til at finde klynger, som også kan forstås intuitivt.

Dimentionalitetsreduktion

Hver gang man tilføjer en måling til et datapunkt, tilføjer man også en dimension i sit "datarum". Har man målt f.eks. vægt og højde af en række personer, har man et todimensionalt datarum. Tilføjer man også alderen har man et tredimensionalt datarum. Med machine learning kan man arbejde med så mange dimensioner, man vil – i f.eks. tekstanalyse arbejder man nogle gange med millioner af dimensioner. Desværre bliver metoderne ofte mindre præcise, når man tilføjer flere dimensioner. Med dimensionalitetsreduktion kan man reducere antallet af dimensioner ved at finde nogle få kombinationer af målinger, som indeholder størstedelen af den oprindelige information.

Dimensionalitetsreduktion er nyttigt både som hjælp til andre machine learning-metoder, men også når man skal visualisere data med mange dimensioner. Eleverne vil få mulighed for at arbejde med en af de mest anvendte metoder til dimensionalitetsreduktion.

Feature selection

Forskellige typer af målinger, f.eks. højde, vægt, alder osv., kaldes i machine learning-jargon for "features". Med feature selection kan man finde de features, der har største betydning for den opgave, men vil have løst. Det kunne f.eks. være hvilke målinger (vægt, alder, blodtryk osv.), der giver mest information om hvorvidt en patient er ved at udvikle kræft. Eleverne vil prøve at arbejde med den pt. bedste metode til feature selection.

2.2 Skitse af forløb

Forløbet strækker sig over tre uger og er tænkt til at kombinere et it-fag med et andet fag. Den mest oplagte kombination vil være med et naturvidenskabeligt eller samfundsvidenskabeligt fag, da det skal være muligt at kunne finde data at arbejde med. Data kan enten opsamles som en del af et forsøg, eller man kan finde nogle interessante datasæt på nettet (f.eks. på `http://archive.ics.uci.edu/ml/` eller `http://mldata.org/`).

Forløbet afsluttes med et mindre tværfagligt projekt, hvor eleverne diskuterer de lærte metoder ud fra et it-mæssigt synspunkt, samtidig med at metoderne anvendes på dataene fra kombinationsfaget, og resultaterne diskuteres i fagets kontekst.

Uge 1

It-fag: Begynde at bruge python. Eleverne skal lære de mest basale ting såsom variable, løkker, if-else, skrive simple funktioner, samt lave basale plots.
Kombinationsfag: Det præcise emne, der vil undersøges, fastlægges og der findes passende datasæt.

Uge 2

It-fag: Rejseholdet kommer og giver en lektion i machine learning. Se emnerne ovenfor.
Kombinationsfag: Analyse af dataene ved hjælp af de gennemgåede metoder påbegyndes.

Uge 3

It- og kombinationsfag: Dataanalyse færdiggøres og projektrapporten udarbejdes.