**Folkesundhed**

Screeningsprogrammer for sygdomme baseret på

genetiske markører med brug af AI.

Ideen til dette undervisningsforløb stammer fra den vedhæftede artikel om screening for tarmkræft ved brug af fem genetiske markører, som samlet set har vist sig at være tilstrækkeligt forskellige fra syge til raske personer til at kunne danne baggrund for et screeningsprogram.

Artiklen er meget teknisk, men den præsenterer også nogle relevante overvejelser omkring screeningsprogrammer generelt. De behøver ikke være ufejlbarlige på individniveau for at forbedre folkesundheden, men skal fange mange syge tidligt i forløbet, så behandling er mulig. De skal også være enkle og helst ikke invasive, så mange vælger screeningen til.

Vi har ikke tid og ressourcer til at etablere et datagrundløb fra en stor klinisk undersøgelse af mange mennesker, så vi vil kunstigt og med brug af matematik generere et tilsvarende datamateriale, som vi kan bruge som eksempel. Vi vil i den forbindelse bruge de indbyrdes forhold mellem syge og raske for de fem udvalgte genetiske markører fra artiklens tabel 2. Det giver forhåbentlig et realistisk bud på, hvor store forskelle mellem raske og syge, man kan forvente at finde ved genundersøgelser i forbindelse med andre alvorlige sygdomme.

Fra artiklens har vi følgende forhold syge/raske for de fem udvalgte genetiske markører:

Gen1: 0,43

Gen2: 0,42 Bemærk, at nogle i gennemsnit er højere for de syge

Gen3: 1,34 og nogle er lavere (fra 42% til 134%).

Gen4: 1,29

Gen5: 0,42

Vi vil generere talværdier for genmarkørerne ved at sample en normalfordeling, hvor vi har valgt en middelværdi som den gennemsnitlige værdi for den pågældende genmarkør og en spredning som et mål for den biologiske variation.

**Aktivitet 1** (Afvikles af læreren)

Repeter lidt deskriptiv statistik fra ugrupperede observationer til grupperede observationer og histogrammer. Brug elevernes højder som eksempel.

Bemærk, at histogrammer ofte har en klokkeform, og inddrag flere eksempler blandt andet med afsæt i biologisk variation.

Præsenter tæthedsfunktioner fra normalfordelingen som en model for klokkeformen, som har vist sig at passe godt i mange anvendelser. Sandsynligheder er arealer under grafen.

Vis hvordan man kan sample data med normalfordelingen. Brug f.eks. højder af værnepligtige med middelværdi cm og spredning cm.

For raske har vi valgt at bruge middelværdien og spredningen altså normalfordelingen .

For syge ændrer vi middelværdien afhængig af det angivne forhold for den pågældende genmarkør, men beholder en spredning på 5.

For gen1 bruges .

For gen2 bruges .

For gen3 bruges .

For gen4 bruges .

For gen5 bruges .

Vi kan nu generere vores kunstige kliniske data ved at sample disse normalfordelinger.

Det kan gøres på følgende måde i Maple:

Først definerer vi de nødvendige normalfordelinger:

Et billede, der indeholder tekst, Font/skrifttype, hvid, typografi

Automatisk genereret beskrivelse

Vi genererer nu de kliniske data for 10 raske og 10 syge personer. Under preferencer og precision har jeg sat outputtet i Maple til to decimaler. Tallene i hver søjle svarer til talværdierne for gen1 til gen5 for en person:

Et billede, der indeholder tekst, Font/skrifttype, skærmbillede, nummer/tal

Automatisk genereret beskrivelse

Resultaterne bliver naturligvis forskellige hver gang man sampler normalfordelingerne. Herunder er et eksempel:

Et billede, der indeholder tekst, skærmbillede, Font/skrifttype

Automatisk genereret beskrivelse

Vi har nu de kliniske data fra vores 20 fiktive personer. F.eks. kan vi aflæse, at for den syge person nummer tre er tallene for genmarkørerne:

Et billede, der indeholder tekst, Font/skrifttype, nummer/tal, skærmbillede

Automatisk genereret beskrivelse

**Aktivitet 2** (Eleverne genererer fiktive kliniske data - skal bruges senere)

Generer kunstige kliniske data for raske og for syge ved at sample normalfordelingerne ovenfor.

Prøv også generere data, hvor spredningen er sat ned fra 5 til 2. Det bør give en større forskel på de kliniske data for syge og for raske.

Prøv tilsvarende at generere data, hvor spredningen er sat op fra 5 til 10. Det bør give en mindre forskel på de kliniske data for syge og for raske.

Bemærk, at vi har valgt at bruge normalfordelingerne som udgangspunkt. Det er blot for at illustrere metoden. Det har altså ikke afsæt i kendte biologiske parametre. At sætte middelværdien til 10 svarer blot til at vælge en enhed, hvor talværdien bliver 10, så det er ret uproblematisk. At sætte spredningen til 5 kan være mere problematisk, da det svarer til at angive information om den biologiske variation for den pågældende genmarkør.

I en virkelig anvendelse med brug af rigtige kliniske data for raske og for syge ville man sikkert se, at normalfordelinger er en god model for de kliniske data, men at middelværdi og spredning er forskellig for hver af de fem genmarkører.

**Aktivitet 3** (Afvikles af læreren)

Giv en kort introduktion til træning af neurale net. Gerne med brug af aalborg-intelligence.ai

Vi vil nu illustrere, hvordan de kliniske data kan bruges til at træne et neuralt net, som efterfølgende kan bruges i vores screeningsprogram for den pågældende sygdom. Vi lader de første syv raske og syv syge personer udgøre vores træningsdata.

Først sørger vi for, at Maple kan vise større matricer på skærmen:

Et billede, der indeholder Font/skrifttype, typografi, hvid, kalligrafi

Automatisk genereret beskrivelse

Vi indlæser også Maples DeepLearning-pakke:

Et billede, der indeholder Font/skrifttype, typografi, kalligrafi, tekst

Automatisk genereret beskrivelse

Herefter indtaster vi vores træningsdata idet vi tildeler de raske resultatet 1 og de syge resultatet 0:

Et billede, der indeholder tekst, Font/skrifttype, hvid, algebra

Automatisk genereret beskrivelse

Det giver følgende tabel

Et billede, der indeholder tekst, skærmbillede, Font/skrifttype, nummer/tal

Automatisk genereret beskrivelse

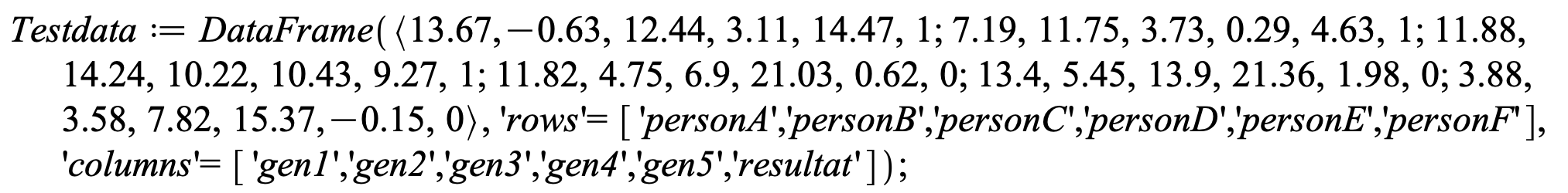
Vi kan nu definere og træne et neuralt net med vores kliniske data:

Et billede, der indeholder tekst, Font/skrifttype, hvid, algebra

Automatisk genereret beskrivelse

Vi har nu et neuralt net, som er trænet og klart til brug.

Før vi kan bruge det i et screeningsprogram, skal vi bruge de resterende seks personer fra vores fiktive kliniske data til at teste, om det virker efter hensigten. Vi indtaster:



og det giver

Et billede, der indeholder tekst, Font/skrifttype, skærmbillede, Elektrisk blå

Automatisk genereret beskrivelse

Vi tester så det neurale net:

Et billede, der indeholder tekst, skærmbillede, Font/skrifttype, linje/række

Automatisk genereret beskrivelse

Vi ser, at modellen stort set ser ud til at virke, men dog ikke fanger, at personF er syg. PersonD og personE ville imidlertid være blevet opdaget ved en screening.

Kvaliteten af en AI anvendelse afhænger voldsomt af kvaliteten og mængden af træningsdata.

**Aktivitet 4** (Eleverne bruger egne fiktive kliniske data til træning af et neuralt net og efterfølgende

test af at screeningen med AI virker)

Brug nogle af jeres egne fiktive kliniske data fra aktivitet 2 som træningsdata og andre af jeres kliniske data som testdata.

Definer og træn et neuralt net med jeres træningsdata.

Afprøv derefter med jeres testdata om screeningen for sygdommen med AI virker.

Et billede, der indeholder tekst, skærmbillede, Font/skrifttype, Publikation/tidsskrift/artikel

Automatisk genereret beskrivelse

Et billede, der indeholder tekst, skærmbillede, design

Automatisk genereret beskrivelse

Et billede, der indeholder tekst, Publikation/tidsskrift/artikel, avis, papir

Automatisk genereret beskrivelse

Et billede, der indeholder tekst, avis, papir, sort-hvid

Automatisk genereret beskrivelse

Et billede, der indeholder tekst, dokument, sort-hvid, papir

Automatisk genereret beskrivelse

Et billede, der indeholder tekst, papir, Publikation/tidsskrift/artikel, avis

Automatisk genereret beskrivelse