

URSKILJANDE AV IRISARTER OCH ANDRA EXEMPEL PÅ HUR MATEMATIK/MASKININLÄRNING/AI KAN ANVÄNDAS FÖR ÖVERVAKAD KLASSIFICERING AV OLIKA OBJEKT

Skapare: Christer Andersson, Anne-Maria Ernvall-Hytönen, Timo Koski, Edward Krogius

Skapad hösten 2023.

Det här verket är licenserat med en Creative Commons Erkännande-DelaLika 4.0 Internationell.

Målgrupp: Högstadiе- eller gymnasieelever. Lägstadiеelever i matteklubbar.

Förhandsinformation: Koordinatsystemet, baskunskaper i GeoGebra (eller ett annat likadant system).

Lektionens längd: 45 minuter ->

Inlärningsmål: Efter lektionen vet eleven att olika objekt och växter kan klassificeras med matematiska metoder med stöd av kvantitativa data. Dessa metoder utgör grundvalen i dagens avancerade AI. Hen kan använda GeoGebra med enkla datamängder och göra försök hur t.ex. olika växtarter kan kännas igen.

Övriga mål: Eleven lär sig hur olika växtarter kan matematiskt klassificeras och separeras från varandra och får en liten inblick av hur en inlärningsmaskin fungerar när den försöker att känna igen olika objekt.

Innehåll: matematik, datavetenskap, biologi

Utförande: Uppgifterna under rubriken "Släktet Iris och uppgifter om det" är basuppgifter som passar för alla, och som även kan lösas tillsammans i matteklubbar i lägstadiet. De kräver inte djupa metoder. Man måste förstå talpar och hur dessa kan sättas i ett koordinatsystem. Man måste ha baskunskaper i GeoGebra (eller i ett liknande system eller papper och penna i uppgifter 1–9). Läraren behöver inte förbereda sig – det räcker med att följa instruktionerna. Uppgifterna under rubriken "Fordjupande uppgifter" utgör ett överkursmaterial för gymnasiet.

INTRODUKTION

Data i experimenten nedan kan åskådliggöras i ett tvådimensionellt plan. Således kan det vara möjligt att separera två datamängder, som svarar till exempel mot växter ur två olika arter, från varandra genom att med ögonmått rita en linje mellan dem. Dylikt är inte alltid möjligt, men kan fungera t.ex. i delar av biologisk systematik. I avsnittet med "Fordjupande uppgifter" visas hur en dator kan programmeras för att dra en rät linje för att separera två datamängder. Detta kallas perceptronalgoritmen. Härvid imiterar datorn människans förmåga att dra en linje mellan två datamängder med försök och fel. När denna linje finns på plats, kan den utnyttjas för att klassificera nya datapunkter utan direkt inverkan av en mänsklig expert på biologisk systematik. Dylikt utgör ett exempel på vad artificiell intelligens (AI) kan åstadkomma.

1. LITE OM TERMINOLOGIN I BIOLOGISK SYSTEMATIK

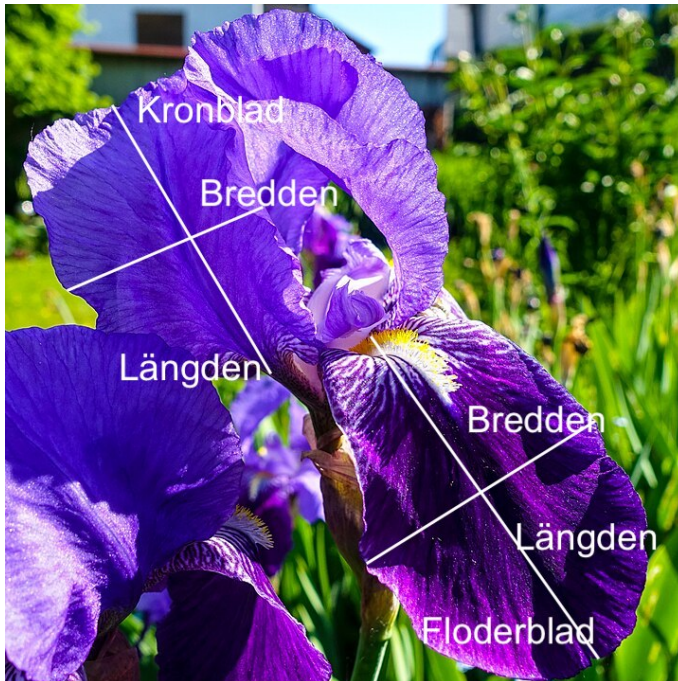
I biologisk systematik är **ett släkte** (synonym: genus) en grupp inom en **familj**. Släkte är alltså ett snävare begrepp än familj. Detta kan måhända föranleda en viss förvirring, ty det oftast uppfattas tvärtom, när det gäller släkt och familj hos människor. Ett släkte består av en eller flera **arter** (species).

Familjen *Iridaceae* består av enhjärtbladiga växter. *Iridaceae* har ett nittiototal släkten med omkring 1800 arter. Här studeras några arter i släktet *Iris*.

SLÄKTET IRIS OCH UPPGIFTER OM DET

Det finns, liksom kan förväntas, ett flertal olika arter av släktet *Iris*. Dessa arter kännetecknas av olika typer av kronblad och foderblad.

- länk: Tuviris - *iris setosa* / laji.fi
- länk: Brokiris – *Iris versicolor* /laji.fi
- länk: Tuviris - *iris setosa* / laji.fi



Modifierat av Georg Buzin - Eget arbete, CC BY-SA 4.0, <https://commons.wikimedia.org/w/index.php?curid=109307159>

Olika Irisarter kan separeras kvantitativt med hjälp av längden och bredden på deras blad. Det finns en datamängd känd som Fishers Iris dataset, som regelbundet används för att testa olika metoder i klassificering/maskininlärning/AI.

Uppgift 1. I denna uppgift arbetar vi med att urskilja två irissorter från varandra. Nu är det tuviris *Iris setosa* och brokiris *Iris versicolor* som skall grupperas. Sätt igång genom att ta dig till Wikipedia: https://en.wikipedia.org/wiki/Iris_flower_data_set (se också originalartikeln¹). Du hittar där Fishers Iris dataset, en datamängd med olika längder och bredder på kronblad och foderblad hos några irisarter.

En del av datasettet (egenskapsvektorer med komponenternas benämningar) är här:

Iris setosa: (foderblad)

längd	5.1	4.9	4.7	4.6	5.0	5.4	4.6	5.0	4.4	4.9
bredd	3.5	3.0	3.2	3.1	3.6	3.9	3.4	3.4	2.9	3.1

¹Fisher, R. A. 138: *The Use of Multiple Measurements in Taxonomic Problems*. Annals of Eugenics, 7: 179-188 (1936).

Iris versicolor:

foderbladslängden	foderbladsbredden
7,0	3,2
6,4	3,2
6,9	3,1
5,5	2,3
6,5	2,8
5,7	2,8
6,3	3,3
4,9	2,4
6,6	2,9
5,2	2,7

Mata in dessa featurevektorer i GeoGebra. Använd olika färger för tuviris *Iris setosa* och brokiris *Iris versicolor*, så att det är så lätt som möjligt för dig att associera en punkt i planet till dess art.

Uppgift 2. Du har nu två punktmängder i Geogebra: en av dem representerar arten *Iris setosa* och den andra representerar arten *Iris versicolor*. Rita en linje mellan de två punktmolnen. Du kan använda glidaren. Försök att hitta med 'trial and error' en rät linje som har ett så stort avstånd från båda molnen som möjligt.

Uppgift 3. Här är de nästa tio featurevektorerna för *Iris setosa* och *Iris versicolor* från Fishers Iris Data:

Iris setosa:

foderbladslängden	foderbladsbredden
5,4	3,7
4,8	3,4
4,8	3,0
4,3	3,0
5,8	4,0
5,7	4,4
5,4	3,9
5,1	3,5
5,7	3,8
5,1	3,8

Iris versicolor:

fodersbladslängden	foderbladsbredden
5,0	2,0
5,9	3,0
6,0	2,2
6,1	2,9
5,6	2,9
6,7	3,1
5,6	3,0
5,8	2,7
6,2	2,2
5,6	2,5

Mata nu dessa data till GeoGebra. Använd samma färger som tidigare så att du ser till vilken art som varje punkt tillhör.

Uppgift 4. Hur många av de nya punkterna klassificerades rätt (träffsäkerhet), alltså hur många av dem hamnade på den rätta sidan av linjen?

Uppgift 5. Skulle du byta linjens plats enligt den nya datamängden som du just lade till, eller är du nöjd med hur linjen separerar *Iris versicolor* från *Iris setosa*?

Uppgift 6. Ta nya punkter från Fishers Iris Data (https://en.wikipedia.org/wiki/Iris_flower_data_set) och testa härefter. (Tabellerna ovan motsvarar de 20 första *Iris setosa* och *Iris versicolor*. (Sepal=Foderblad och petal=kronblad.)

Uppgift 7. Låt oss nu betrakta kronblad. Data:

Iris setosa:

kronbladslängden	kronbladsbredden
1,4	0,2
1,4	0,2
1,3	0,2
1,5	0,2
1,4	0,3
1,7	0,4
1,4	0,3
1,5	0,2
1,4	0,2
1,5	0,1

Iris versicolor:

kronbladslängden	kronbladsbredden
4,7	1,4
4,5	1,5
4,9	1,5
4,0	1,3
4,6	1,5
4,5	1,3
4,7	1,6
3,3	1,0
4,6	1,3
3,9	1,4

Mata dem in i Geogebra med olika färger, och separera mängderna med hjälp av en rät linje.

Uppgift 8. Kan du göra modellen för separering enklare, alltså skulle det räcka med att ha en endimensionell modell?

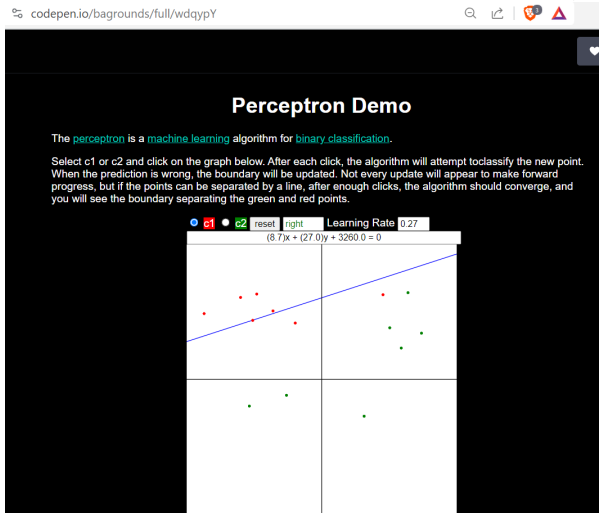
Uppgift 9. En irisplanta har foderbladslängden 7,7 cm, foderbladbredden 3,0 cm, kronbladslängden 6,1 cm och kronbladbredden 2,3 cm.

- (1) Verkar det mer sannolikt att denna planta är *Iris setosa* eller *Iris versicolor*?
- (2) Det finns också andra arter i släktet, t.ex. *Iris virginica*. Verkar det sannolikt att plantan är *Iris setosa* eller *Iris versicolor* eller kanske en annan art?

FÖRDJUPANDE UPPGIFTER

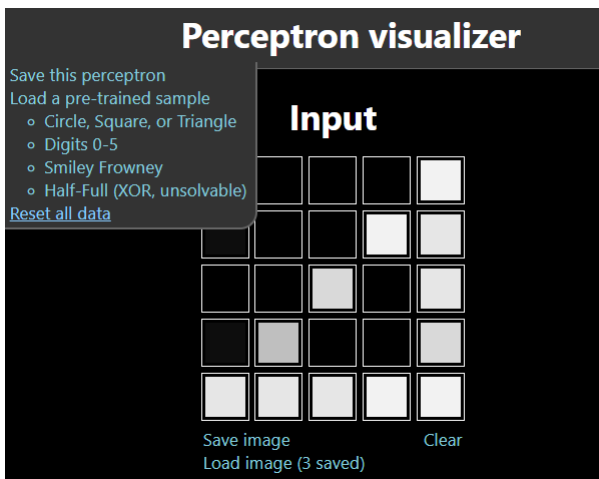
I detta avsnitt studeras den ovan framställda metodologin med hjälp av webbdemonstrationer samt pythonkoder, vilka implementerar den basala artificiella intelligensen. Härvid utvidgas fältet av applikationer från biologisk systematik. Det finns många andra likanda program som illustrerar AI. Till exempel, det finns GenAI Teachable Machine <https://tm.generation-ai-stn.fi/image/general> som kräves inte programmeringskunskaper.

Uppgift 10. Testa en online demo av Rosenblatts perceptron gemon att lägga till poäng. Här med punkter i 2D <https://codepen.io/bagrounds/full/wdqypY> av Bryan Grounds.



Olika länkar till demonstrationer av perceptroner online. <https://perceptrondemo.com>
Här har vi ett pixelfält med litet olika dataset för igenkänning med AI.

- geometriska figurer
- siffrorna 0-5
- smileys
- XOR (kan ej separeras)



Uppgift 11. Läs kapitel 1-4 eller kapitel 4 s.98 - s.101 ur boken: John D. Kelleher: *Syväoppiminen, kuinka tekoäly toimii*, Terra Cognita, Helsinki, 2020
eller motsvarande avsnitt i den engelska versionen John D. Kelleher: *Deep learning*. MIT Press, 2019.

- (1) Beskriv i ord inlärningsregeln för Rosenblatts perceptron.
- (2) Vad menas med en iteration av en inlärningsomgång i träning av en perceptron?
- (3) Diskutera inlärningshastighetens betydelse i träning av en perceptron med Rosenblatts perceptronalgoritm, se s. 102.
- (4) I vilken mening kan en perceptron ses som urtypen för dagens avancerade AI?

Uppgift 12. Pröva att rita en graf med ett python program.

Du kan ändra på punktmängderna och linjens koefficienter för olika grafvarianter.

- gå till tutorialspoint länk: till koden
- klicka på [Preview]
- länk: Online python på tutorialspoint

```
# CC BY 4.0 E.Krogius 2023
# pip install matplotlib

import matplotlib.pyplot as plt

def rita(x1, y1, x2, y2, col1, col2):
    # ritar punkter och en separerande hyperlinje
    plt.figure()

    # Punkterna
    plt.plot(x1, y1, marker='o', linestyle='', color=col1)
    plt.plot(x2, y2, marker='o', linestyle='', color=col2)

    # axlarna
    plt.xlabel('xaxeln')
    plt.ylabel('yaxeln')

    # Vi ritar in en separerande hyperlinje
    xm = 4 # x-coordinate of the point on the line
    ym = 3.4 # y-coordinate of the point on the line
    k = 0.2 # slope of the line

    plt.title(f'y = {k}x + {ym - k*xm:.2f}')

    # linjens punkter
    y_line_values = [ym + k * (x_val - xm) for x_val in x1]

    # rita linjen
    plt.plot(x1, y_line_values, linestyle='--', color='r')

x1 = [5.4, 3.6, 4.6, 5.0, 4.4, 4.5]
y1 = [3.9, 3.4, 5.1, 3.8, 5.0, 4.1]
x2 = [5.1, 3.6, 4.7, 5.0, 5.0, 4.5]
y2 = [2.1, 1.1, 3.1, 3.1, 2.1, 1.8]
rita(x1, y1, x2, y2, 'b', 'g')

plt.show()
```

