Osnovi računarske inteligencije 2016/17

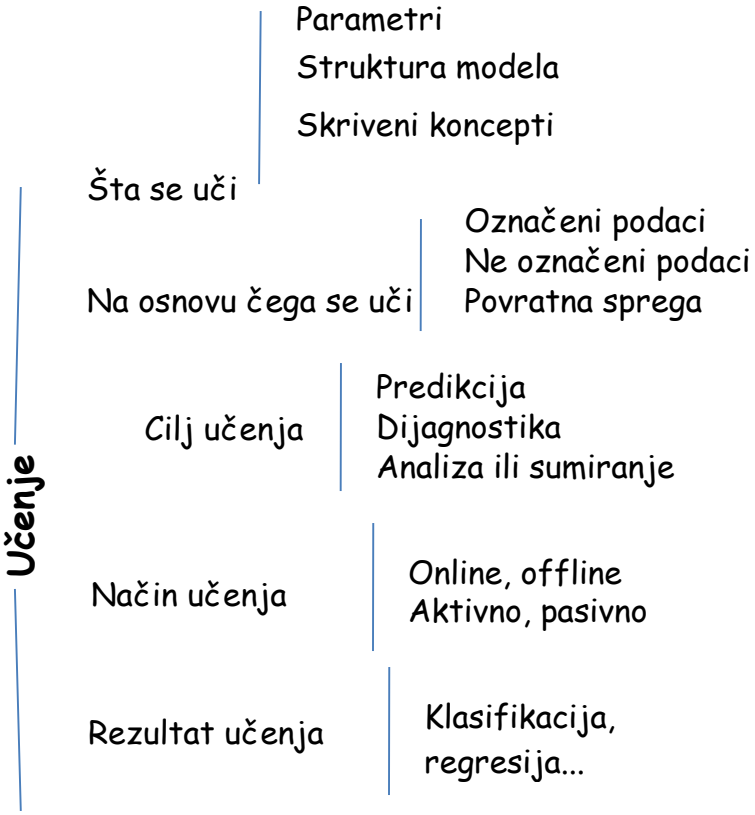
Mašinsko učenje

Arthur Samuel (1959):

> Disciplina koja omogućuje računarima da UČE bez da su eksplicitno programirani.

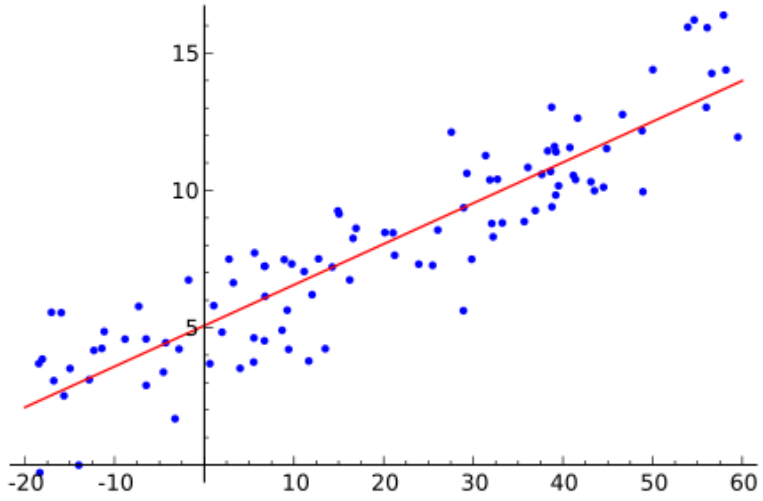
Tom Mitchel (1998):

> Dobro postavljen problem UČENJA.
>
> Za računarski program kažemo da UČI na osnovu ISKUSTVA (E) u odnosu na neki POSAO (T) i meru KVALITETA (P) ako svoje performanse merene sa P radeći posao T unapređuje korišćenjem iskustva E.

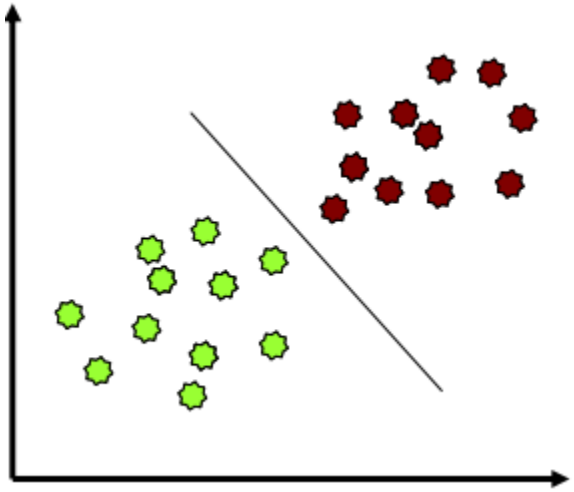


Nadgledano učenje

Regresija



Klasifikacija



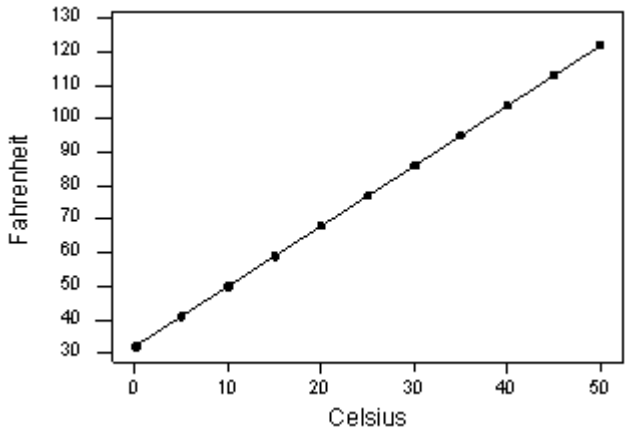
Linearna regresija

Statistički metod koji omogućava proučavanje veze između 2 kontinualne promjenljive:

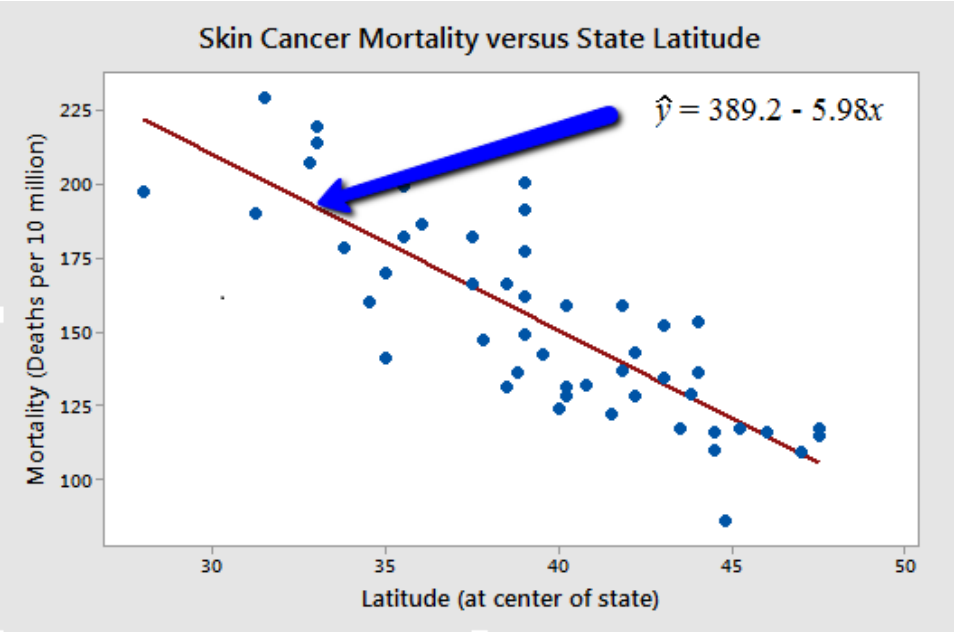
- **x** – nezavisna promenljiva (prediktor)
- **y** – zavisna promenljiva (odziv)

Deterministička (funkcionalna) zavisnost – tačno određivanje **y** na osnovu **x**:

$$\text{Fahr} = \frac{9}{5} * \text{Cels} + 32$$



Statistička zavisnost – veza između 2 promenljive nije savršeno tačna:



Kriterijum najmanjih kvadrata

Minimizovati sumu kvadrata grešaka predikcije:

- jednačina koja najbolje „fituje“ ***f(x) = ax + b***
- odrediti ***a*** i ***b*** tako da je vrednost funkcije greške minimalna:

$$\Phi(a,b) = \sum_{i=1}^{n} (b + ax_i - y_i)^2$$

Da bi odredili minimum funkcije greške, izjednačićemo njene parcijalne izvode po ***a*** i ***b*** sa nulom:

$$\frac{\partial\, \Phi(a,b)}{\partial\, a} = 0$$

$$\frac{\partial\, \Phi(a,b)}{\partial\, b} = 0$$

Nakon izvođenja dobijamo:

$$a = \frac{n\left(\sum_{i=1}^{n} x_i\, y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2}$$

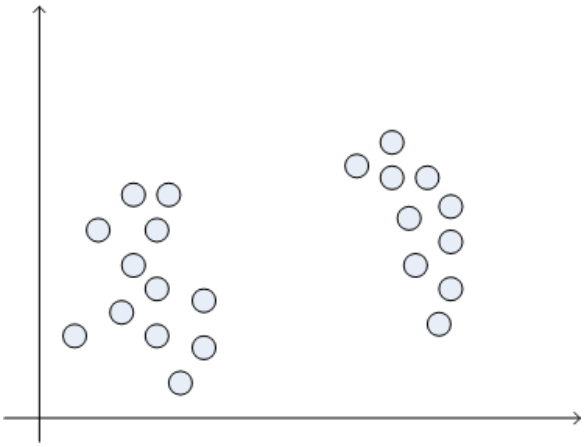
$$b = \frac{1}{n}\left(\left(\sum_{i=1}^{n} y_i\right) - a\left(\sum_{i=1}^{n} x_i\right)\right)$$

Nenadgledano učenje

Na osnovu podataka koji nisu označeni odrediti pravila ili parametre tako da se podaci mogu grupisati.

Nisu označeni jer se unapred ne zna kojoj grupi pripadaju.

Grupisanje/Klasterizacija



K-Means algoritam

Polazna pretpostavka:

broj grupa (K) je fiksan, konačan i unapred dat.

Cilj:

kreirati „kompaktne“ grupe.

Malo formalnije:

1. Inicijalizovati K grupa, odnosno K centara grupa μ_k

Za svaku iteraciju n (dok sistem konvergira):

2. Svaki objekat pridruži najbližem centru korišćenjem funkcije rastojanja
3. Izračunaj nove centre grupa na osnovu formule:

$$\mu_k = \frac{1}{n}\sum x_i$$

Gde je(su):

μ_k – centar grupa k

n – broj objekata koji pripadaju grupi k

x_i – objekti iz grupe k

Još malo formalnije:

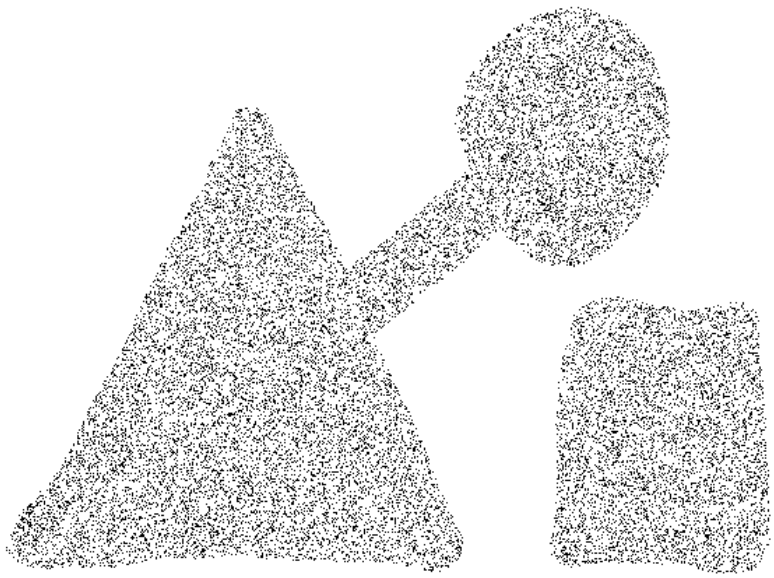
Tražimo minimum funkcije:

$$cost(\mu_1, \mu_2 \ldots, \mu_k) = \sum_{\mu_j} \sum_{x_i iz\ j} (x_i - \mu_j)^2$$

DBSCAN (Density Based Spatial Clustering)

Algoritam: DBScan(D, eps, G)

- za sve tačke iz D čija je epsilon gustina manja od G, kažemo da su ŠUM i izuzimamo ih iz daljeg razmatranja
- za preostale objekte (T) koristimo pretragu kako kod obeležavanja regiona i tako formiramo grupu



Zadaci

1. Linearna regresija:
   - u klasi **Program.cs** učitati skup podataka iz **data/skincancer.csv**
   - u klasi **LinearRegression.cs** implementirati metodu **fit** koja određuje parametre **k** i **n**
   - u klasi **LinearRegression.cs** implementirati metodu **predict** koja na osnovu prosleđene **x** vrednosti vraća predviđenu vrednost **y.**
   - u klasi **Program.cs** pozvati metodu **fit** za učitane podatke i ispisati dobijene parametre **k** i **n**.
   - u klasi **Program.cs** pozvati metodu **predict** i izvršiti predikciju za državu Hawaii (Lat = 20).

2. K-Means:
   - u klasi **KMeans.cs** inicijalizovati centre grupa na slučajan način
   - u klasi **KMeans.cs** dopuniti funkciju pomeriCentar() tako da računa novi centar u odnosu na elemente grupe
   - dopuniti funkciju za računanje rastojanja koristeći Euklidsko rastojanje. (za vežbu probati i sa drugim rastojanjima npr Menhetn)
   - grupisati države iz **skincancer** skupa podataka na osnovu geografske dužine i širine

Zadaci mogu biti implementirani u bilo kom programskom jeziku. Kostur projekta koji je dat u prilogu ovih vežbi ne morate koristiti!

DODATNI ZADATAK KLASTERIZACIJA: Implementirati DBScan algoritam i klasterizovati skincancer skup podataka kao u primeru za KMeans

DODATNI ZADATAK PREDIKCIJA: Implemenitrati višestruku regresiju. Izvršiti predviđanje za dobijanje raka na osnovu geografske širine i dužine.