Sistemi za istraživanje podataka

Data Mining (*DM*)

Sadržaj

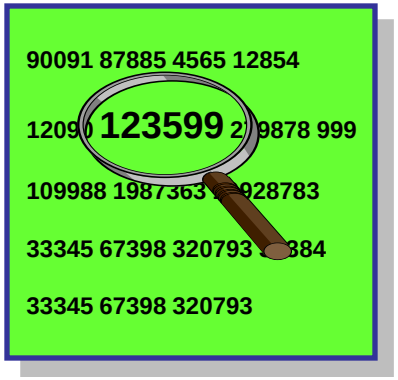
- Data Mining
- Prebrojavanje sličnih pojava
- Istraživanje zakonitosti
- Pravila, strukturirana u obliku stabla
- Klasterizacija
- Pronalaženje sličnih vremenskih serija
- Veštačke neuronske mreže

Data Mining

- **Data Mining - Istraživanje podataka**
  - istraživanje i analiza veoma velikih količina podataka u cilju pronalaženja ili potvrđivanja
    - značajnih uzoraka podataka ili činjenica
    - trendova u životnom ciklusu podataka
    - međuzavisnosti podataka
    - pravila (tj. zakonitosti)
  - **Alternativna terminologija**
    - **Otkrivanje znanja (Knowledge Discovery)**
    - **Data Surfing**
    - **Žetva podataka (Data Harvesting)**

Data Mining

- **Motivacija**
  - podrška procesa
    - predviđanja ponašanja poslovnog sistema i
    - odlučivanja
  - pronalaženje zakonitosti, pravila i trendova, koji su
    - skriveni, ili nisu očigledno prepoznatljivi
    - suviše kompleksni da bi bili otkriveni korišćenjem samo statističkih modela analize
      - upotreba tehnika kao što su
        - » induktivno i deduktivno rezonovanje
        - » veštačke neuronske mreže
        - » klasterizacija



Data Mining

- **Moguće oblasti primene**
  - Marketing, Prodaja, Razvoj proizvoda, Planiranje i analiza poslovanja
    - profilisanje korisnika proizvoda (kupaca)
    - profilisanje proizvoda na tržištu
    - segmentacija tržišta
    - otkrivanje zakonitosti prodaje
    - otkrivanje grešaka i uspeha u poslovanju
    - predviđanje i analiza rizika poslovanja
    - predviđanje i analiza uspešnosti poslovanja

Data Mining

- **Preduslovi za uspešnu primenu**
  - postojanje odgovarajuće **Data Warehouse** baze podataka
    - sa velikom količinom istorijskih podataka
  - podrška **odgovarajućih metoda istraživanja podataka**
    - statističkih
      - eksploratorne ("istraživačke") analize podataka
    - veštačke inteligencije
      - otkrivanje znanja i mašinsko učenje

Data Mining

- **Preduslovi za uspešnu primenu**
  - podrška **odgovarajućih alata za prezentaciju** rezultata
  - **edukovani i iskusni korisnici** u domenu primene

Data Mining

- **Rizici uspešne primene**
  - pojedinačne transakcije istraživanja podataka
    - mogu da traju satima
    - mogu da obuhvataju veoma veliku količinu podataka
  - potrebna moćna H/S infrastruktura
    - u cilju obezbeđenja zadovoljavajućih performansi
  - pravilno sagledavanje i razumevanje korisničkih zahteva
    - komunikacija s relevantnim, edukovanim i iskusnim krajnjim korisnicima

Data Mining

- **Koraci primene**
  - **kreiranje i održavanje DW baze podataka**
    - Selekcija, transformacija i transport podataka
      - Prečišćavanje i usaglašavanje podataka
    - Inicijalno punjenje i osvežavanje podataka u DW
  - **uzorkovanje podataka**
    - pravilan izbor reprezentativnih uzoraka za analizu
  - **istraživanje podataka (Data Mining)**
    - analize podataka i rezonovanje

Data Mining

- **Koraci primene**
  - **prezentacija rezultata istraživanja**
    - Vizuelizacija produkovanih rezultata
      - korišćenjem odgovarajućih softverskih alata
  - **ocenjivanje i analiza dobijenih rezultata**
    - ekspertsko ocenjivanje i analiza rezultata
      - od strane korisnika (menadžera, analitičara, projektanata)
  - **upotreba dobijenih rezultata u poslovanju**
    - primena na proces upravljanja, tj. odlučivanja

Data Mining

- **Moguće Data Mining tehnike**
  - **Prebrojavanje sličnih pojava**
    - **Counting Co-Occurrences**
  - **Istraživanje zakonitosti**
    - **Mining for Rules**
  - **Pravila, strukturirana u obliku stabla**
    - **Tree–Structured Rules**
  - **Klasterizacija**
    - **Clustering**

Data Mining

- **Moguće Data Mining tehnike**
  - **Pronalaženje sličnih vremenskih serija**
    - **Similarity Search over Sequences**
  - **Veštačke neuronske mreže**
    - **Artificial Neural Networks**
  - **Deduktivno rezonovanje**
    - **Deductive Reasoning**
  - **Genetski algoritmi**
    - **Genetic Algorithms**

Sadržaj

- Data Mining
- Prebrojavanje sličnih pojava
- Istraživanje zakonitosti
- Pravila, strukturirana u obliku stabla
- Klasterizacija
- Pronalaženje sličnih vremenskih serija
- Veštačke neuronske mreže

Prebrojavanje sličnih pojava

- **Counting Co-Occurrences**
  - **Motivacija**
    - **Analiza potrošačke korpe (Market basket analysis)**
      - **Potrošačka korpa - Market Basket**
        - » jedna transakcija kupca – kupovina više artikala
  - **Cilj**
    - Identifikovati artikle koji se kupuju zajedno
  - **Dva karakteristična tipa tehnika**
    - **Frequent Itemsets** – frekventni skupovi artikala
    - **Iceberg Queries** – upiti tipa ledenog brega

Primer:
uzorak podataka o prodaji artikala

Prebrojavanje sličnih pojava

Prodaja

TransID	KupacID	Datum	Artikal	Količina
111	201	10. 09. 01.	Penkalo	2
111	201	10. 09. 01.	Mastilo	1
111	201	10. 09. 01.	Mleko	3
111	201	10. 09. 01.	Sok	6
112	105	10. 09. 01.	Penkalo	1
112	105	10. 09. 01.	Mastilo	1
112	105	10. 09. 01.	Mleko	1
113	106	10. 09. 01.	Penkalo	1
113	106	10. 09. 01.	Mleko	1
114	201	11. 09. 01.	Penkalo	2
114	201	11. 09. 01.	Mastilo	2
114	201	11. 09. 01.	Sok	4

Prebrojavanje sličnih pojava

- **Frekventni skupovi artikala**
  - **Itemset (skup činilaca)**
    - skup artikala koji se pojavljuju zajedno u transakciji
  - **Support za itemset (podrška za skup činilaca)**
    - relativni broj (ili procenat) transakcija u kojima se svi artikli iz itemset-a pojavljuju zajedno
      - određuje se u odnosu na ukupan broj transakcija
  - **MinSup**
    - unapred definisani minimalni Support
  - **Frekventni Itemset**
    - Itemset za koji je **Support ≥ MinSup**

Prebrojavanje sličnih pojava

- **Frekventni skupovi artikala**
  - Primer
    - MinSup = 0.75
    - Frekventni itemset-ovi:
      - {*penkalo*}, {*mastilo*}, {*mleko*}
      - {*penkalo*, *mastilo*}, {*penkalo*, *mleko*}

Prebrojavanje sličnih pojava

- **Frekventni skupovi artikala**
  - **"A priori" pravilo**
    - svaki podskup frekventnog itemset-a mora biti frekventni itemset
  - **Algoritam pronalaženja frekventnih itemset-ova**
    - **Ulaz:** *N* - ukupan broj artikala
    - **Izlaz:** Skup svih frekventnih itemset-ova

Prebrojavanje sličnih pojava

- **Frekventni skupovi artikala**
  - **Algoritam pronalaženja frekventnih itemset-ova**
    - Odrediti skup I_1
      - skup svih **jednočlanih** frekventnih itemset-ova
    - Radi za $k = 2$ do N :
      - Formiranje kandidata za itemset-ove kardinalnosti k
        - » Napraviti sve kombinacije itemset-ova kardinalnosti k, polazeći od frekventnih itemset-ova iz skupa I_{k-1} (za koje važi Support $\geq$ MinSup)
      - Formiranje skupa frekventnih itemset-ova I_k
        - » Izabrati samo one kandidate za I_k, za koje važi Support $\geq$ MinSup

Prebrojavanje sličnih pojava

- **Frekventni skupovi artikala**
  - Primer
    - MinSup = 0.70
    - Frekventni itemset-ovi kardinalnosti 1 (l_1):
      - {*penkalo*}, {*mastilo*}, {*mleko*}
    - Kandidati za itemset-ove kardinalnosti 2:
      - {*penkalo*, *mastilo*}, {*penkalo*, *mleko*}, {*mleko*, *mastilo*}
    - Frekventni itemset-ovi kardinalnosti 2 (l_2):
      - {*penkalo*, *mastilo*}, {*penkalo*, *mleko*}
        - » {*mleko*, *mastilo*} se izbacuje, jer je Support = 0.50 < 0.70

Prebrojavanje sličnih pojava

- **Frekventni skupovi artikala**
  - Primer
    - MinSup = 0.70
    - Frekventni itemset-ovi kardinalnosti 1 (l_1):
      - {*penkalo*}, {*mastilo*}, {*mleko*}
    - Frekventni itemset-ovi kardinalnosti 2 (l_2):
      - {*penkalo*, *mastilo*}, {*penkalo*, *mleko*}
        - » {*mleko*, *mastilo*} se izbacuje, jer je Support = 0.50
    - Frekventni itemset-ovi kardinalnosti 3 (l_3):
      - ne postoje
        - » {*penkalo*, *mastilo*, *mleko*} nije kandidat, jer {*mleko*, *mastilo*} nije frekventni itemset

Prebrojavanje sličnih pojava

- **Upiti tipa ledenog brega**
  - Primer
    - Selektovati samo one parove (*KupacID*, *Artikal*), za koje važi da je njihova ukupna prodata količina ne manja od 5

```
SELECT KupacID, Artikal, SUM(Količina)
FROM Prodaja
GROUP BY KupacID, Artikal
HAVING SUM(Količina) >= 5;
```

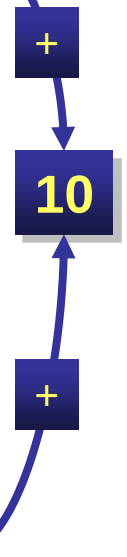
    - Problem: po pravilu, broj parova (*KupacID*, *Artikal*) je
      - veliki, u ukupnom obimu
      - vrlo mali, koji zadovoljava uslov SUM(*Količina*) >= 5
        - "vrh ledenog brega" - SUM(*Količina*) >= 5

Primer:
uzorak podataka o prodaji artikala

Prebrojavanje sličnih pojava

Prodaja

TransID	KupacID	Datum	Artikal	Količina
111	201	10. 09. 01.	Penkalo	2
111	201	10. 09. 01.	Mastilo	1
111	201	10. 09. 01.	Mleko	3
111	201	10. 09. 01.	Sok	6
112	105	10. 09. 01.	Penkalo	1
112	105	10. 09. 01.	Mastilo	1
112	105	10. 09. 01.	Mleko	1
113	106	10. 09. 01.	Penkalo	1
113	106	10. 09. 01.	Mleko	1
114	201	11. 09. 01.	Penkalo	2
114	201	11. 09. 01.	Mastilo	2
114	201	11. 09. 01.	Sok	4



Prebrojavanje sličnih pojava

- **Upiti tipa ledenog brega - algoritmi**
  - Ukoliko je broj parova (*KupacID*, *Artikal*) dovoljno mali da može stati u operativnu memoriju:
    - Pretražiti relaciju, uz kontinuirano ažuriranje sume artikala za svaki par (*KupacID*, *Artikal*)
  - Ukoliko je broj parova (*KupacID*, *Artikal*) vrlo veliki:
    - SUBP će, klasično, izvršiti:
      - uređivanje (soritiranje) grupa podataka (*KupacID*, *Artikal*)
      - sumiranje količina po svakoj grupi i
      - eliminaciju grupa koje ne zadovoljavaju traženi uslov

Prebrojavanje sličnih pojava

- **Upiti tipa ledenog brega - algoritmi**
  - Ukoliko je broj parova (*KupacID*, *Artikal*) vrlo veliki:
    - Tehnike DM
      - iskoristiti činjenicu da je broj traženih parova vrlo mali, u odnosu na ukupan broj parova
      - primeniti analogiju sa problemom frekventnih itemset-ova
    - "A priori" pravilo
      - posmatrati samo one parove (*KupacID*, *Artikal*), za koje *KupacID* i *Artikal*, pojedinačno, zadovoljavaju HAVING uslov
        - » ostali parovi sigurno ne zadovoljavaju uslov

Prebrojavanje sličnih pojava

- **Upiti tipa ledenog brega - algoritmi**
  - Ukoliko je broj parova (*KupacID*, *Artikal*) vrlo veliki:
    - Algoritam DM – tri prolaza kroz relaciju *Prodaja*
      - Pronaći kupce koji su kupovali artikle u ukupnoj količini ≥ 5
      - Pronaći artikle, koji su kupovani u ukupnoj količini ≥ 5
      - Na osnovu prethodnih rezultata, formirati sve parove kandidata (*KupacID*, *Artikal*)
      - Pronaći, među dobijenim kandidatima, samo one parove (*KupacID*, *Artikal*), za koje je SUM(*Količina*) >= 5

Prebrojavanje sličnih pojava

- **Upiti tipa ledenog brega - algoritmi**
  - Primer
    - U relaciji *Prodaja* - ukupno 12 parova (*KupacID*, *Artikal*)
      - jedan prolaz zahteva održavanje sume za 12 parova
    - Kupci koji su ukupno kupovali najmanje 5 artikala:
      - **(*KupacID*, *Količina*): {(*201*, *20*)}**
    - Artikli, koji su ukupno kupovani u količini od najmanje 5:
      - **(*Artikal*, *Količina*): {(*sok*, *10*), (*mleko*, *5*), (*penkalo*, *6*)}**
    - Kandidati parova (relevantni parovi) (*KupacID*, *Artikal*):
      - **{(*201*, *sok*), (*201*, *mleko*), (*201*, *penkalo*)}**
    - Parovi koji zadovoljavaju uslov SUM(*Količina*) >= 5
      - **{(*201*, *sok*)}**

Sadržaj

- Data Mining
- Prebrojavanje sličnih pojava
- Istraživanje zakonitosti
- Pravila, strukturirana u obliku stabla
- Klasterizacija
- Pronalaženje sličnih vremenskih serija
- Veštačke neuronske mreže

Istraživanje zakonitosti

- **Mining for Rules**
  - Postoji dosta algoritama, namenjenih za istraživanje zakonitosti podataka
  - Neki primeri
    - **Pravila povezanosti**
      - **Association rules**
    - **Klasifikaciona i regresiona pravila**
      - **Classification and regression rules**
    - **Sekvencijalni uzorci**
      - **Sequential patterns**

Istraživanje zakonitosti

- **Pravila povezanosti**
  - **Pravilo povezanosti podataka – izraz oblika:**

$$LHS \Rightarrow RHS$$

    - *LHS* i *RHS* – skupovi činilaca poslovanja (artikala)
  - Interpretacija pravila poslovanja:
    - Ukoliko se, u jednoj transakciji, kupe svi artikli iz *HLS*, vrlo je verovatno da će biti kupljeni i svi artikli iz *RHS*
  - Primer:
    - {*penkalo*} ⇒ {*mastilo*}

Istraživanje zakonitosti

- **Pravila povezanosti**
  - **Support za skup artikala - Sup(S):**
    - relativni broj (ili procenat) transakcija u kojima se zajedno pojavljuju svi artikli iz S
  - Support za pravilo $LHS \Rightarrow RHS$
    - Sup(S) = Sup($LHS \cup RHS$)
  - **Poverenje (Confidence) pravila $LHS \Rightarrow RHS$**
    - Conf($LHS \Rightarrow RHS$) = Sup($LHS \cup RHS$) / Sup(LHS)
    - predstavlja meru "jačine" (verodostojnosti) pravila
      - uslovna (Bajesova) verovatnoća

Primer:
uzorak podataka o prodaji artikala

Istraživanje zakonitosti

Prodaja

TransID	KupacID	Datum	Artikal	Količina
111	201	10. 09. 01.	Penkalo	2
111	201	10. 09. 01.	Mastilo	1
111	201	10. 09. 01.	Mleko	3
111	201	10. 09. 01.	Sok	6
112	105	10. 09. 01.	Penkalo	1
112	105	10. 09. 01.	Mastilo	1
112	105	10. 09. 01.	Mleko	1
113	106	10. 09. 01.	Penkalo	1
113	106	10. 09. 01.	Mleko	1
114	201	11. 09. 01.	Penkalo	2
114	201	11. 09. 01.	Mastilo	2
114	201	11. 09. 01.	Sok	4

Istraživanje zakonitosti

- **Pravila povezanosti**
  - Primer
    - Pravilo: *penkalo* $\Rightarrow$ *mastilo*
      - Sup({*penkalo*}) = 1.00
      - Sup({*penkalo*, *mastilo*}) = 0.75
      - **Conf({*penkalo*} $\Rightarrow$ {*mastilo*}) = 0.75/1.00 = 0.75**
    - Pravilo: *mleko* $\Rightarrow$*sok*
      - Sup({*mleko*}) = 0.75
      - Sup({*mleko*, *sok*}) = 0.25
      - **Conf({*mleko*} $\Rightarrow$ {*sok*}) = 0.25/0.75 = 0.33**

Istraživanje zakonitosti

- **Pravila povezanosti - algoritam**

  – **Ulaz:** MinSup - minimalni support
  MinConf - minimalno poverenje

  – **Izlaz:** Skup svih pravila $LHS \Rightarrow RHS$, takvih da je

$$\textbf{Conf}(LHS \Rightarrow RHS) \geq \textbf{MinConf}$$

Istraživanje zakonitosti

- **Pravila povezanosti - algoritam**
  - **Postupak**
    - primeniti algoritam određivanja svih frekventnih itemset-ova S, za koje je Sup(S) $\geq$ MinSup
    - za svaki frekventni itemset vrši se:
      - generisanje kandidata za pravila povezanosti, pravljenjem svih kombinacija podskupova frekventnog itemset-a
      - Za svakog kandidata za pravilo povezanosti $LHS \Rightarrow RHS$:
        - » izračunava se Conf($LHS \Rightarrow RHS$)
        - » prihvataju se samo ona pravila za koja je

$$\mathbf{Conf(}\boldsymbol{LHS \Rightarrow RHS}\mathbf{) \geq MinConf}$$

Istraživanje zakonitosti

- **Pravila povezanosti - algoritam**
  - Primer
    - Ulaz: MinSup = 0.75, MinConf = 0.80
    - Algoritam za određivanje frekventnih itemset-ova:
      - Sup({*penkalo*}) = 1.0
      - Sup({*mastilo*}) = 0.75
      - Sup({*mleko*}) = 0.75
      - Sup({*penkalo*, *mastilo*}) = 0.75
      - Sup({*penkalo*, *mleko*}) = 0.75

Istraživanje zakonitosti

- **Pravila povezanosti - algoritam**
  - Primer
    - Ulaz: MinSup = 0.75, MinConf = 0.80
    - Određivanje svih kandidata za pravila i određivanje poverenja:
      - Conf({*penkalo*} ⇒ {*mastilo*}) = 0.75 / 1.00 = 0.75 < 0.80
      - Conf({*mastilo*} ⇒ {*penkalo*}) = 0.75 / 0.75 = 1.00 ≥ 0.80
      - Conf({*penkalo*} ⇒ {*mleko*}) = 0.75 / 1.00 = 0.75 < 0.80
      - Conf({*mleko*} ⇒ {*penkalo*}) = 0.75 / 0.75 = 1.00 ≥ 0.80
    - Skup verodostojnih pravila:
      - **{*mastilo*} ⇒ {*penkalo*}**
      - **{*mleko*} ⇒ {*penkalo*}**

Istraživanje zakonitosti

- **Pravila povezanosti - primena**
  - široko upotrebljavana u **predviđanjima**
  - Primer upotrebe u predviđanjima
    - Con($LHS \Rightarrow RHS$)
      - uslovna verovatnoća kupovine artikala iz RHS, pod pretpostavkom kupovine artikala iz LHS
    - Pravilo $LHS \Rightarrow RHS$ s visokim poverenjem može voditi ka zaključku da je poželjno
      - **promovisati bolju prodaju artikala iz *LHS*,** davanjem odgovarajućih popusta
      - u cilju **očekivanja povećanja prodaje artikala iz *RHS***
        - » u prethodnom primeru: **{*mastilo*} $\Rightarrow$ {*penkalo*}**

Istraživanje zakonitosti

- **Pravila povezanosti - primena i "zamke"**
  - predviđanja **moraju biti zasnovana** i na dodatnim analizama u oblasti istraživanja
    - predviđanja su opravdana ako postoji uslovna povezanost artikala iz *LHS* sa artiklima iz *RHS*
      - kupovina penkala **izaziva** kupovinu mastila
    - moguće je, međutim, izvesti pravilo sa visokim poverenjem i support-om, ali tako da ne postoji uslovna zavisnost artikala iz *LHS* sa artiklima iz *RHS*
      - Conf({*mleko*} $\Rightarrow$ {*penkalo*}) = 1.00, ali
      - kupovina mleka **ne izaziva** kupovinu penkala
        - » promovisanjem kupovine mleka neće biti izazvano povećanje prodaje penkala

Istraživanje zakonitosti

- **Klasifikaciona i regresiona pravila**
  - Izrazi oblika

$$P_1(X_1) \wedge P_2(X_2) \wedge \ldots \wedge P_k(X_k) \Rightarrow Y = c$$

- X_i $(i = 1, 2, \ldots, k)$ - atributi predviđanja
- $P_i(X_i)$ $(i = 1, 2, \ldots, k)$ - predikati (svojstva)
- Y - zavisni atribut
- c - konstanta (vrednost)
- $Y = c$ - specijalni (ciljni) predikat

Istraživanje zakonitosti

- **Klasifikaciona i regresiona pravila**

$$P_1(X_1) \wedge P_2(X_2) \wedge \ldots \wedge P_k(X_k) \Rightarrow Y = c$$

  - upotreba
    - Na osnovu uočenih svojstava atributa predviđanja $P(X_i)$, predviđa se vrednost zavisnog atributa Y
      - svojstva (predikati) atributa se pronalaze na osnovu sadržaja DW baze podataka

Istraživanje zakonitosti

- **Klasifikaciona i regresiona pravila**

$$P_1(X_1) \wedge P_2(X_2) \wedge \ldots \wedge P_k(X_k) \Rightarrow Y = c$$

  - vrste atributa X_i i Y, s obizirom na pridruženi domen mogućih vrednosti
    - **numerički** (intervalni)
      - definisan je totalni poredak vrednosti domena
      - $P_i(X_i)$: $v_{i1} \leq X_i \leq v_{i2}$
    - **kategorijalni** ("nabrojani")
      - nije definisan totalni poredak vrednosti domena
      - $P_i(X_i)$: $X_i \in \{v_1, v_2, \ldots, v_n\}$

Istraživanje zakonitosti

- **Klasifikaciona i regresiona pravila**
  - **Klasifikaciono pravilo**
    - ako je zavisni atribut *Y* kategorijalni
  - **Regresiono pravilo**
    - ako je zavisni atribut *Y* numerički

Istraživanje zakonitosti

- **Klasifikaciona i regresiona pravila**
  - Primer
    - ***Osiguranje*(*Starost*, *TipVozila*, *Rizik*)**
      - *dom*(*Starost*) = $\{d \in \mathbf{N} \mid d \geq 18 \wedge d \leq 99\}$
      - *dom*(*TipVozila*) = {*sedan*, *sportski*}
      - *dom*(*Rizik*) = {*visok*, *nizak*}
    - Klasifikaciono prravilo, na osnovu tabele *Osiguranje*

$$(18 \leq Starost \leq 25) \wedge (TipVozila \in \{Sportski\}) \Rightarrow Rizik = visok$$

Istraživanje zakonitosti

- **Klasifikaciona i regresiona pravila**
  - **Support (podrška) predikata $P(X)$: Sup($P(X)$)**
    - relativni broj (ili procenat) torki koje zadovoljavaju $P(X)$
  - **Support (podrška) pravila $P_1(X_1) \Rightarrow P_2(X_2)$**
    - Sup($P_1(X_1) \Rightarrow P_2(X_2)$) = Sup($P_1(X_1) \wedge P_2(X_2)$)
  - **Poverenje pravila $P_1(X_1) \Rightarrow P_2(X_2)$**
    - Conf($P_1(X_1) \Rightarrow P_2(X_2)$) = Sup($P_1(X_1) \wedge P_2(X_2)$) / Sup($P_1(X_1)$)
    - uslovna verovatnoća poverenja (verodostojnosti) pravila

Istraživanje zakonitosti

- **Klasifikaciona i regresiona pravila**
  - **Donja granica poverenja**
    - MinConf
  - **MinConf** se zadaje u cilju procene poverenja pravila

Istraživanje zakonitosti

- **Klasifikaciona i regresiona pravila**
  - Primer
    - MinConf = 0.80
    - Klasifikaciono pravilo *P*
      - $(18 \leq Starost \leq 25) \wedge (TipVozila \in \{Sportski\}) \Rightarrow Rizik = visok$
      - Sup($18 \leq Starost \leq 25$) = 0.30
      - Sup($TipVozila \in \{Sportski\}$) = 0.15
      - Sup($18 \leq Starost \leq 25 \wedge TipVozila \in \{Sportski\}$) = 0.10
      - Sup($Rizik = visok$) = 0.12
      - Sup(*P*) = 0.09
      - Conf(*P*) = 0.09 / 0.10 = 0.90 ≥ MinConf

Istraživanje zakonitosti

- **Sekvencijalni uzorci**
  - često su velike količine podataka u BP smeštene u formi redosleda
    - zadatog po uočenom kriterijumu
      - koji je često vezan za vremensku dimenziju
  - postoji potreba utvrđivanja zakonitosti postojanja takvih redosleda
    - sekvencijalnih uzoraka podataka

Tehnike pronalaženja sličnih vremenskih serija

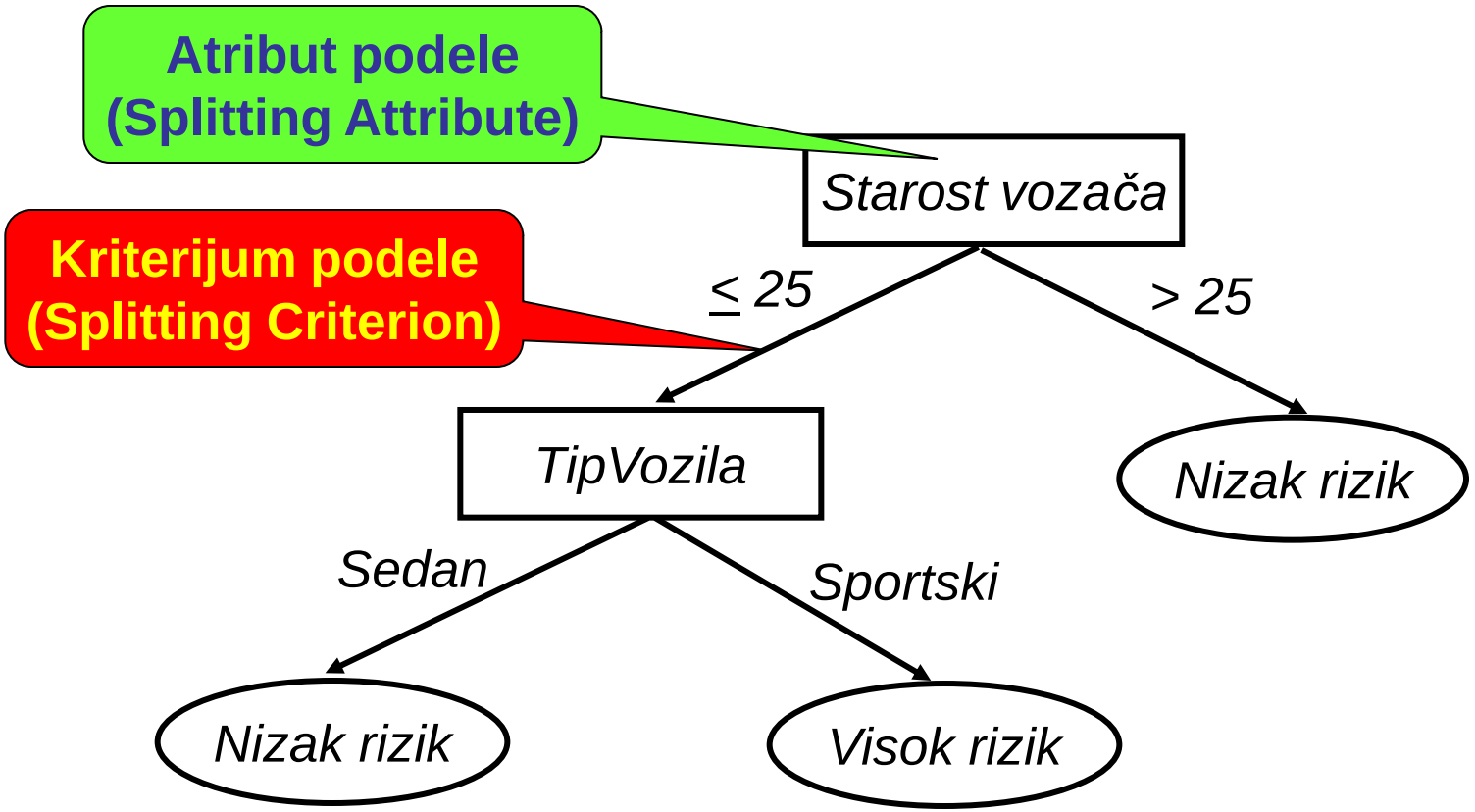
Sadržaj

- Data Mining
- Prebrojavanje sličnih pojava
- Istraživanje zakonitosti
- Pravila, strukturirana u obliku stabla
- Klasterizacija
- Pronalaženje sličnih vremenskih serija
- Veštačke neuronske mreže

Pravila, strukturirana u obliku stabla

- Specijalna regresiona i klasifikaciona pravila se mogu iskazati u obliku stabla
- **Stablo klasifikacije – stablo odlučivanja**
  - stablo koje iskazuje skup klasifikacionih pravila
- **Stablo regresije**
  - stablo koje iskazuje skup regresionih pravila
- Postoje različiti algoritmi za generisanje stabla klasifikacije i odlučivanja

Pravila, strukturirana u obliku stabla



Sadržaj

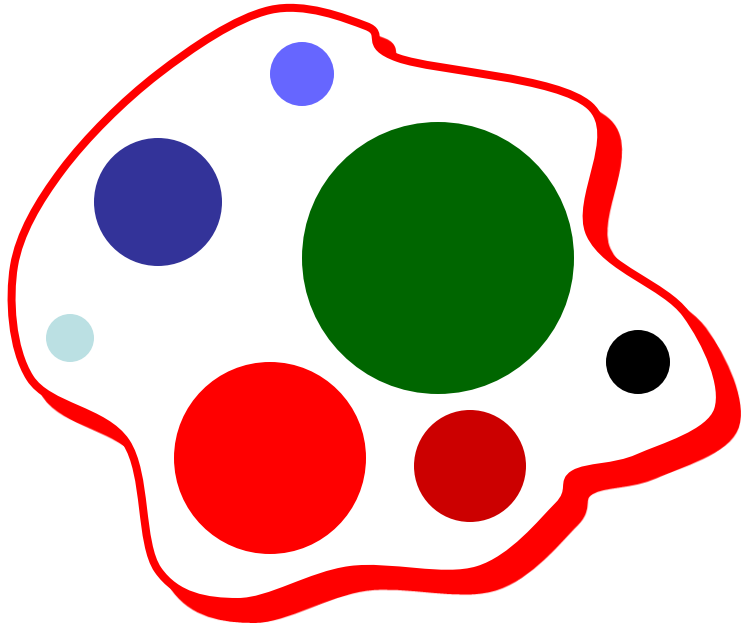
- Data Mining
- Prebrojavanje sličnih pojava
- Istraživanje zakonitosti
- Pravila, strukturirana u obliku stabla
- Klasterizacija
- Pronalaženje sličnih vremenskih serija
- Veštačke neuronske mreže

Klasterizacija

- **Clustering tehnika**
  - **Particioniranje skupa torki u grupe - klastere**
    - po nekom kriterijumu, vezanom za zadatu osobinu (atribut)
      - svi slogovi **u istom klasteru su "slični"** po zadatom kriterijumu
      - slogovi **iz različitih klastera su "različiti"** po zadatom kriterijumu
    - **Funkcija distance (Distanca)**
      - **mera (kriterijum) sličnosti slogova**
        - » u različitim aplikacijama, koriste se različite funkcije distance

Klasterizacija

- **Clustering tehnika**
  - Cilj
    - istraživanje zakonitosti ponašanja klastera u particiji
      - u cilju rezonovanja, odlučivanja i sprovođenja poslovnih akcija

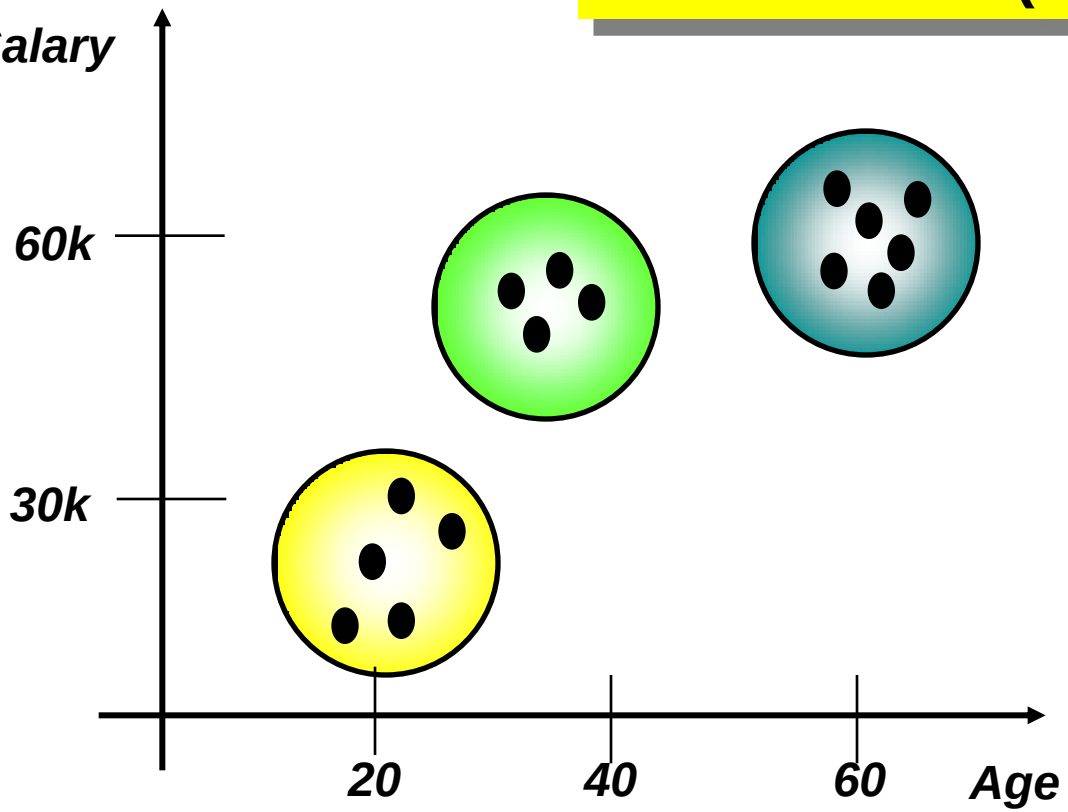


Klasterizacija

- **Clustering tehnika**
  - Primer

CustomerInfo(Age, Salary)

$$s = \frac{1}{n}\sum_{i=1}^{n} s_i$$



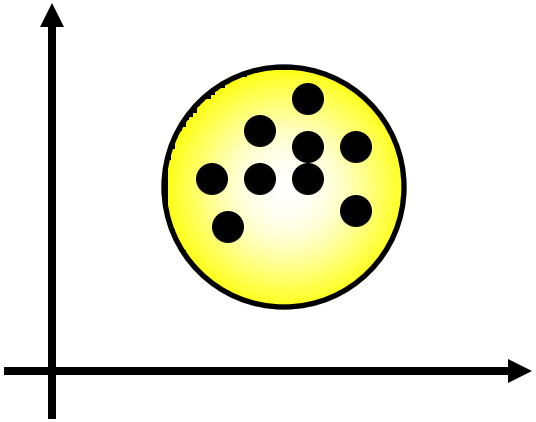
$$a = \frac{1}{n}\sum_{i=1}^{n} a_i$$

Klasterizacija

- **Clustering tehnika**
  - **Klaster - skup slogova**

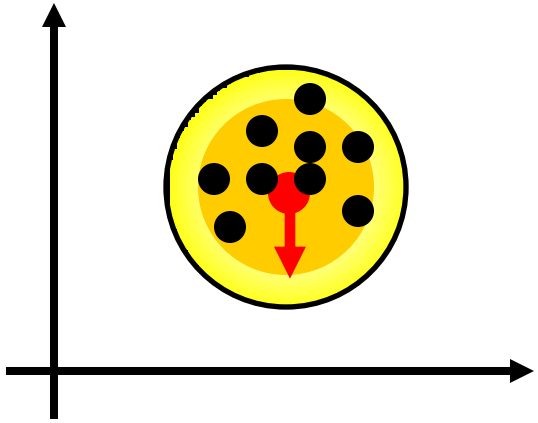
$$\{r_1, \ldots, r_n\}$$

- $(\forall i \in \{1,...,n\})(r_i = (a_i, s_i))$
  - $a_i \in dom(Age)$, $s_i \in dom(Salary)$



Klasterizacija

- **Clustering tehnika**
  - **Karakteristike klastera**
    - Centar *C* = (*a*, *s*)



      - srednja vrednost atributa po kojima je formiran klaster
    - **Radijus *R***

$$s = \frac{1}{n}\sum_{i=1}^{n} s_i \qquad a = \frac{1}{n}\sum_{i=1}^{n} a_i$$

      - srednje odstojanje torki od centra klastera

$$R = \sqrt{\frac{\sum_{i=1}^{n}((a - a_i)^2 + (s - s_i)^2)}{n}}$$

Klasterizacija

- **Algoritam formiranja klastera**
  - **Ulaz**
    - ε - maksimalni dozvoljeni radijus svakog klastera
    - skup numeričkih atributa po kojem se vrši klasterizacija
  - **Izlaz**
    - particija tabele po ε - skup klastera

Klasterizacija

- **Algoritam formiranja klastera**
  - **Postupak**
    - prva izabrana torka - formira se prvi klaster u skupu
    - **za svaku preostalu torku tabele $r_i = (a_i, s_i)$**
      - **za svaki, do tada formirani klaster $C = (a, s)$ iz skupa**
        - » izračunava se distanca r_i od C, po formuli

$$d_i = \sqrt{(a - a_i)^2 + (s - s_i)^2}$$

      - r_i se smešta u
        - » klaster $C = (a, s)$, za koji je d minimalno, ako za novoizračunati radijus klastera važi $R \leq \varepsilon$, inače
        - » novi klaster, koji se dodaje u skup formiranih klastera

Sadržaj

- Data Mining
- Prebrojavanje sličnih pojava
- Istraživanje zakonitosti
- Pravila, strukturirana u obliku stabla
- Klasterizacija
- Pronalaženje sličnih vremenskih serija
- Veštačke neuronske mreže

Pronalaženje sličnih vremenskih serija

- **Vremenski redosled (vremenska serija)**
  - Niz numeričkih podataka X, dužine k

$$X = \langle x_1, \ldots, x_k \rangle$$

    - poređanih saglasno kriterijumu, opredeljenom u vremenskoj dimenziji - hronološki
- **Podredosled redosleda**
  - $X = \langle x_1, \ldots, x_k \rangle$
  - $Z = \langle z_1, \ldots, z_j \rangle$ je podredosled od X ($Z \leq X$) ako je
    - $z_1 = x_i$, $z_2 = x_{i+1}$,..., $z_j = x_{i+j-1}$,
      - za bilo koji $i \in \{1,..., k - j + 1\}$

Pronalaženje sličnih vremenskih serija

- **Distanca redosleda iste dužine**
  - $X = \langle x_1, \ldots, x_k \rangle$ i
  - $Y = \langle y_1, \ldots, y_k \rangle$
    - intenzitet razlike vektora

$$\|X - Y\| = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2}$$

Pronalaženje sličnih vremenskih serija

- **Predstavlja model upita**
  - u kojem korisnik zadaje parametre upita:
    - traženi redosled podataka X i
    - prag sličnosti ε
  - cilj realizacije upita
    - pronaći sve slične redoslede X_i
      - redoslede za koje važi:

$$||X - X_i|| \leq \varepsilon$$

    - Similarity search

Pronalaženje sličnih vremenskih serija

- **Algoritam traženja sličnih redosleda**
  - **pretraživanje kompletnih redosleda**
    - pretraživanje svih torki u tabelama BP
    - selektovanje svakog redosleda, iste dužine kao zadati *X*
    - izračunavanje distance u odnosu na redosled upita
    - preuzimanje samo onih redosleda, čija je distanca unutar zadatog ε
  - **pretraživanje podredosleda**
    - traže se svi redosledi u BP, čiji neki podredosled je sličan sa zadatim redosledom *X*, do na zadatu distancu ε

Sadržaj

- Data Mining
- Prebrojavanje sličnih pojava
- Istraživanje zakonitosti
- Pravila, strukturirana u obliku stabla
- Klasterizacija
- Pronalaženje sličnih vremenskih serija
- Veštačke neuronske mreže

Veštačke neuronske mreže

- **Artificial Neural Network (ANN)**
  - Formalni sistem koji **simulira rad humanih neuronskih sistema**
    - Osposobljen da **uči na osnovu iskustva iz prošlosti**
      - da izvodi pravila rezonovanja, na osnovu istorijskih (eventualno statističkih) podataka
    - Osposobljen da **generalizuje, na osnovu prosleđenih uzoraka podataka**
      - da rezonuje (vrši predikciju, ili klasifikovanje) na osnovu naučenih pravila

Veštačke neuronske mreže

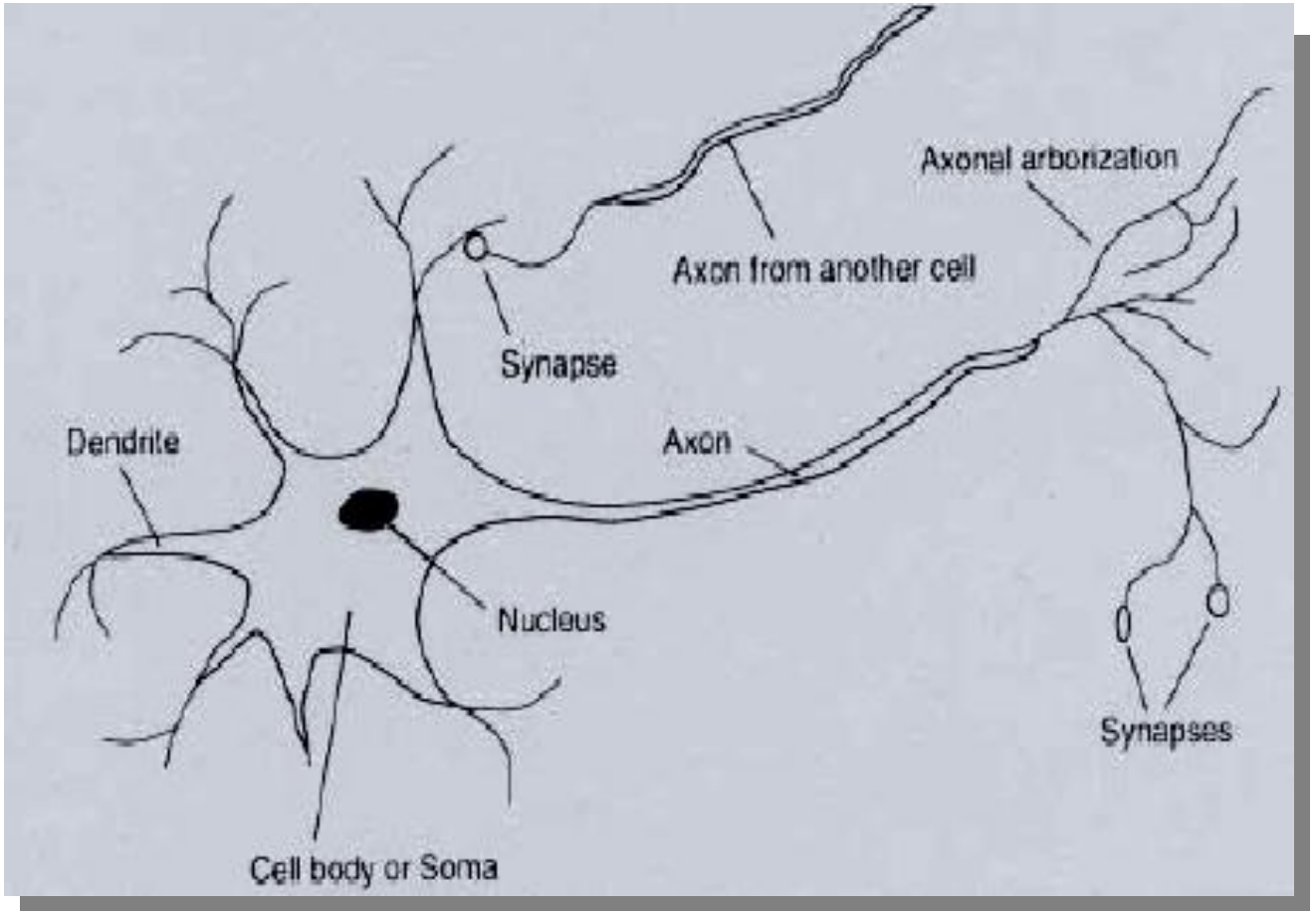
- **Artificial Neural Network (ANN)**
  - **Ideja**
    - sposobnost funkcionalne transformacije ulaznih veličina u izlaznu
    - **putem mreže neurona**
      - **visokog stepena kompleksnosti, tj. velikog broja neurona**
      - **neuron = procesor**
        - » sa relativno jednostavnom funkcionalnošću

Veštačke neuronske mreže

- **Artificial Neural Network (ANN)**
  - **Istorija**
    - definicija ANN 1943. godine
      - Warren McCulloch, neuropsiholog
      - Walter Pitts, logičar
    - pojačano interesovanje za ANN: $\geq$1980. godine
      - u kontekstu razvoja računarskih nauka

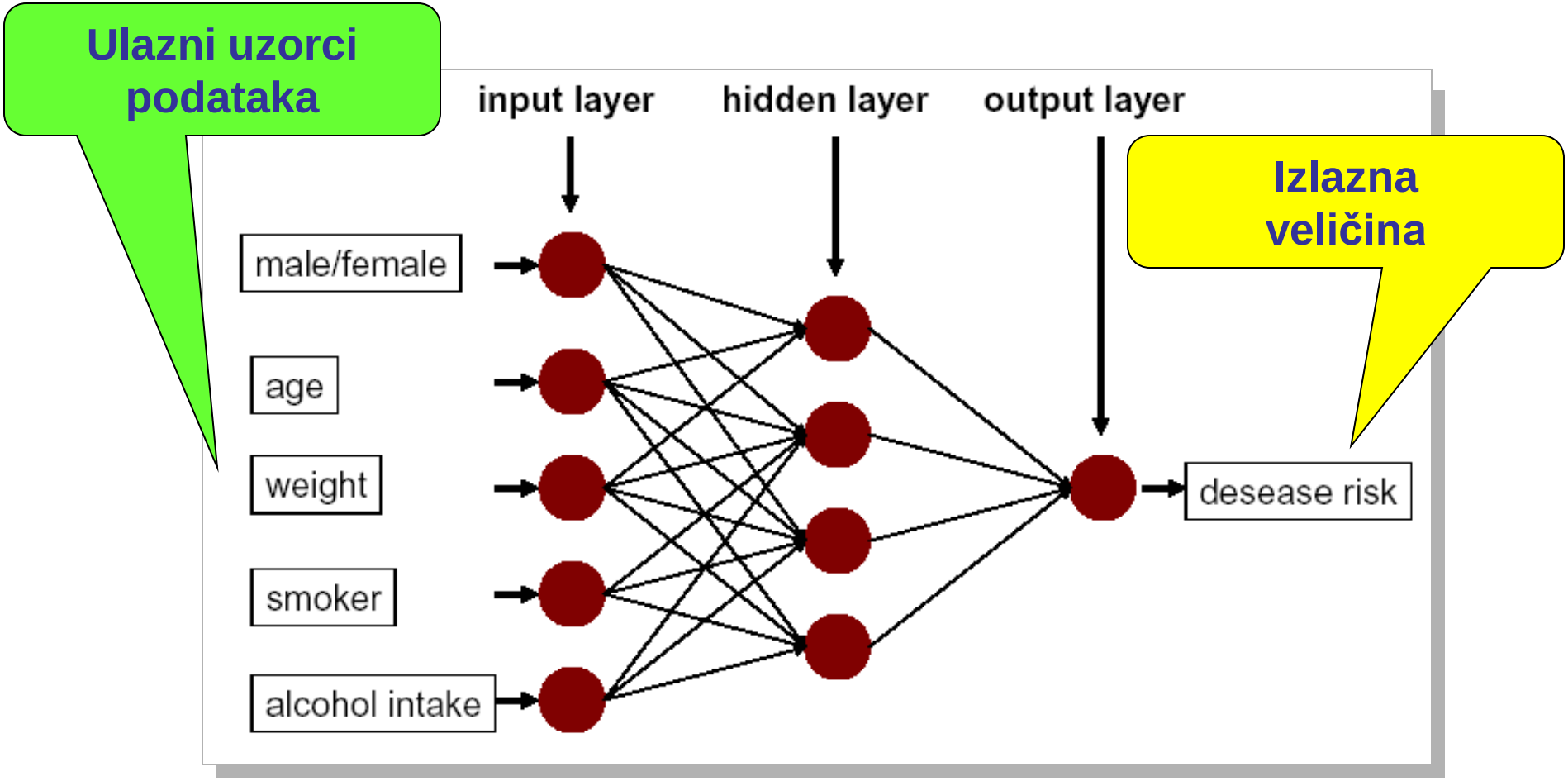
Veštačke neuronske mreže

- **Biološki neuron**



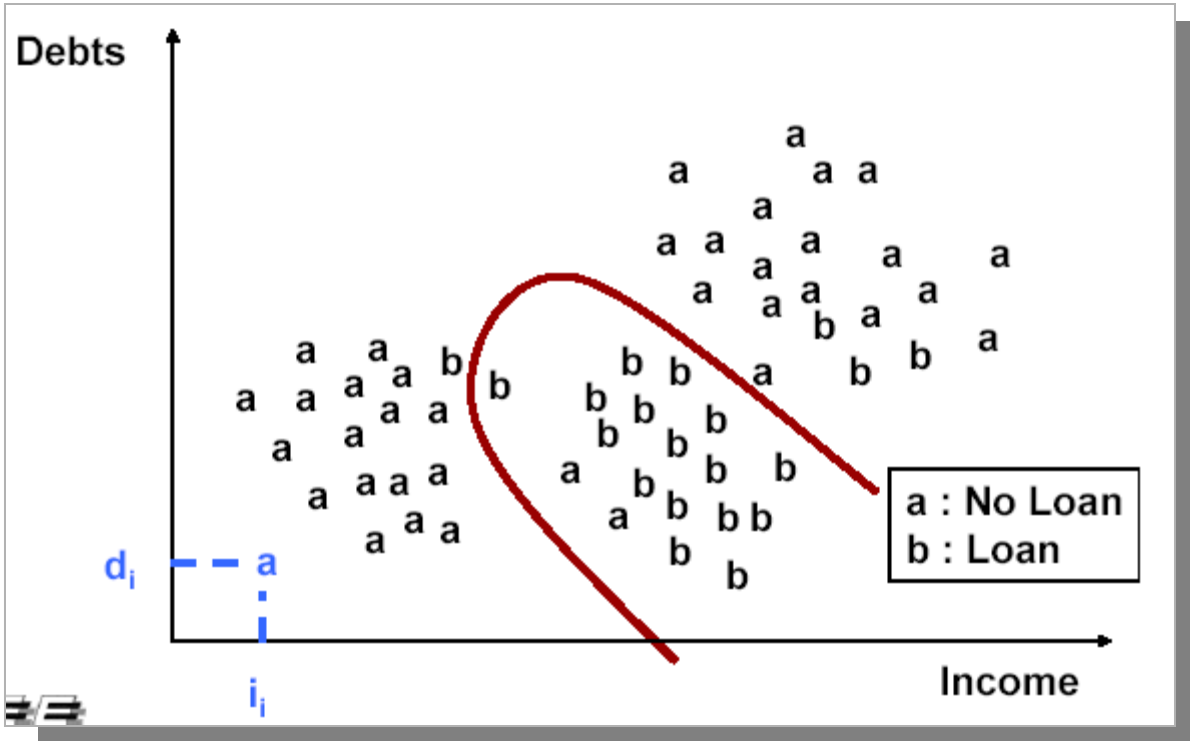
Veštačke neuronske mreže

- **ANN – ilustracija opšte strukture**



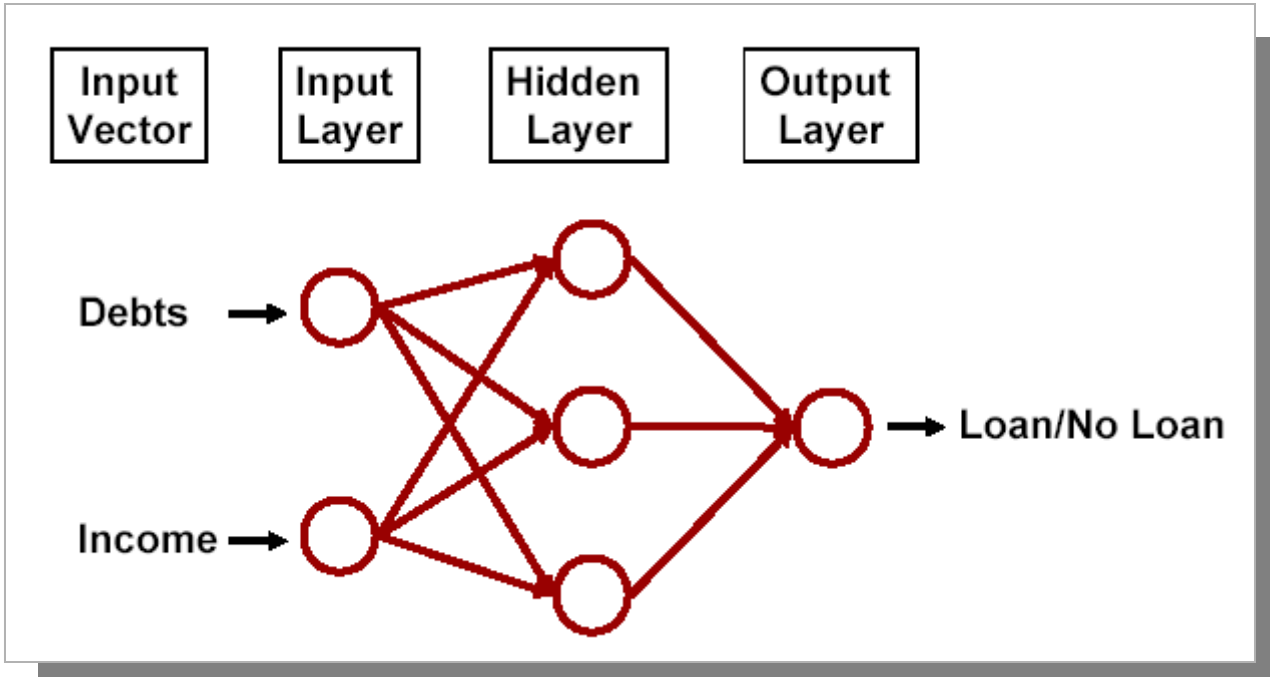
Veštačke neuronske mreže

- **ANN – ilustracija primene**
  - podela prostora uzorka podataka, primenom nelinearne logičke funkcije



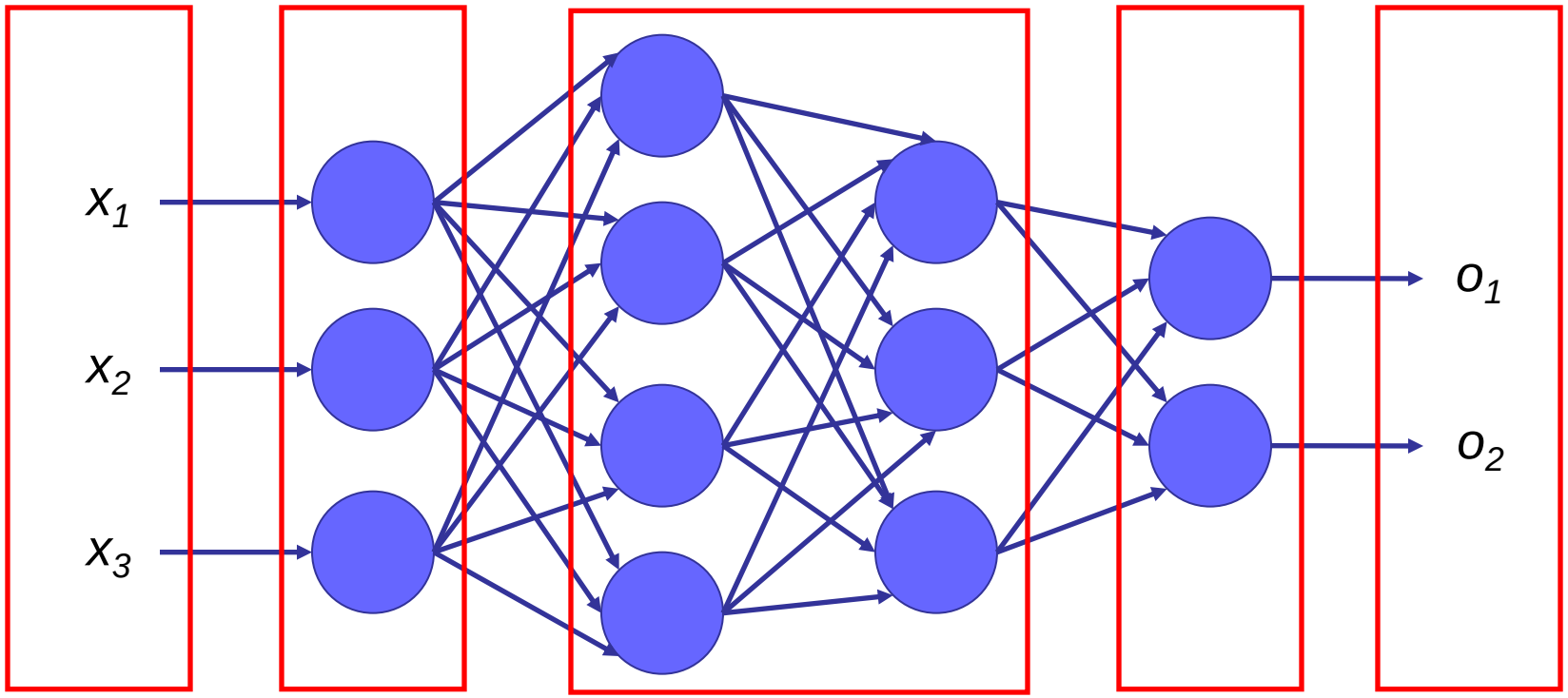
Veštačke neuronske mreže

- **ANN – ilustracija primene**
  - podela prostora uzorka podataka, primenom nelinearne logičke funkcije



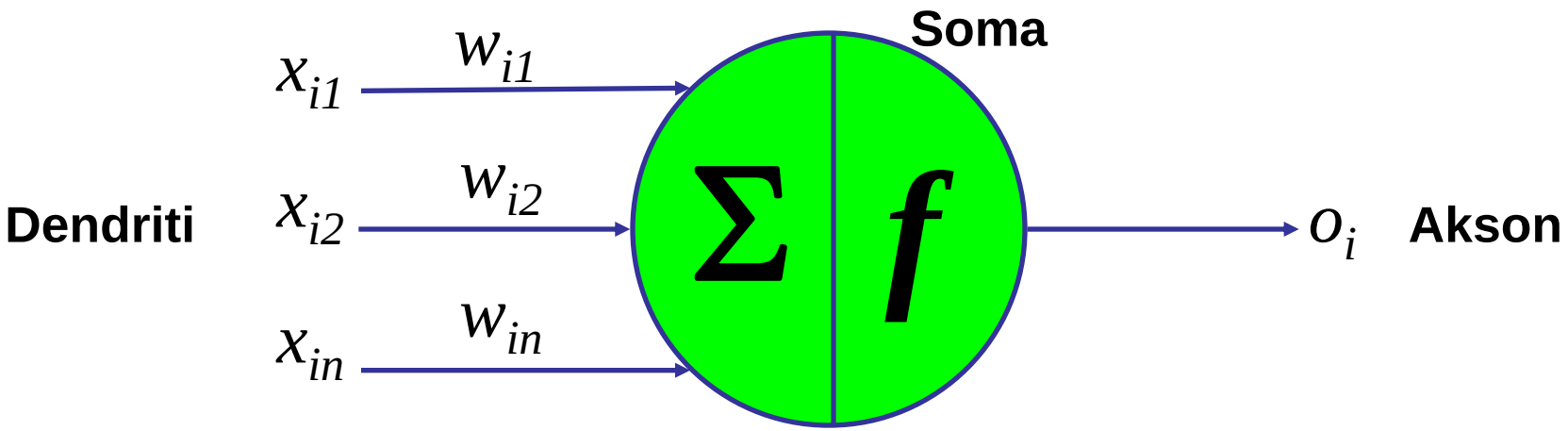
Veštačke neuronske mreže

- **ANN - opšta arhitektura (topologija)**



Veštačke neuronske mreže

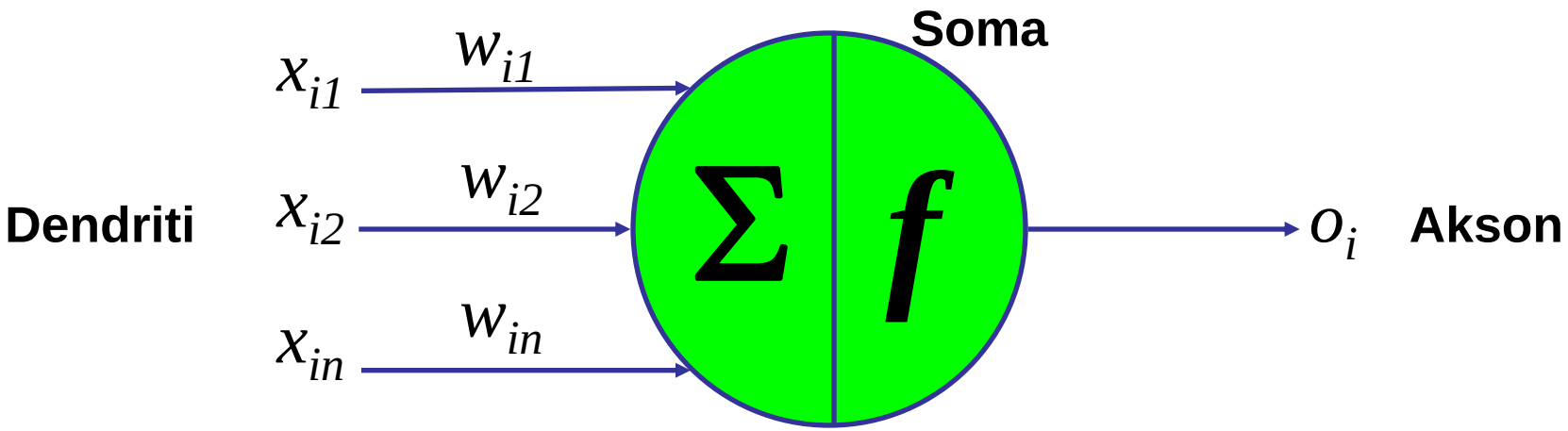
- **Veštački neuron**



- $(x_1,..., x_n)$ - ulazni vektor
- $(w_1,..., w_n)$ - vektor težinskih koeficijenata
- o_i - izlazna veličina neurona
- Σ - sumirajuća funkcija neurona
- f - aktivaciona funkcija (funkcija prelaza)

Veštačke neuronske mreže

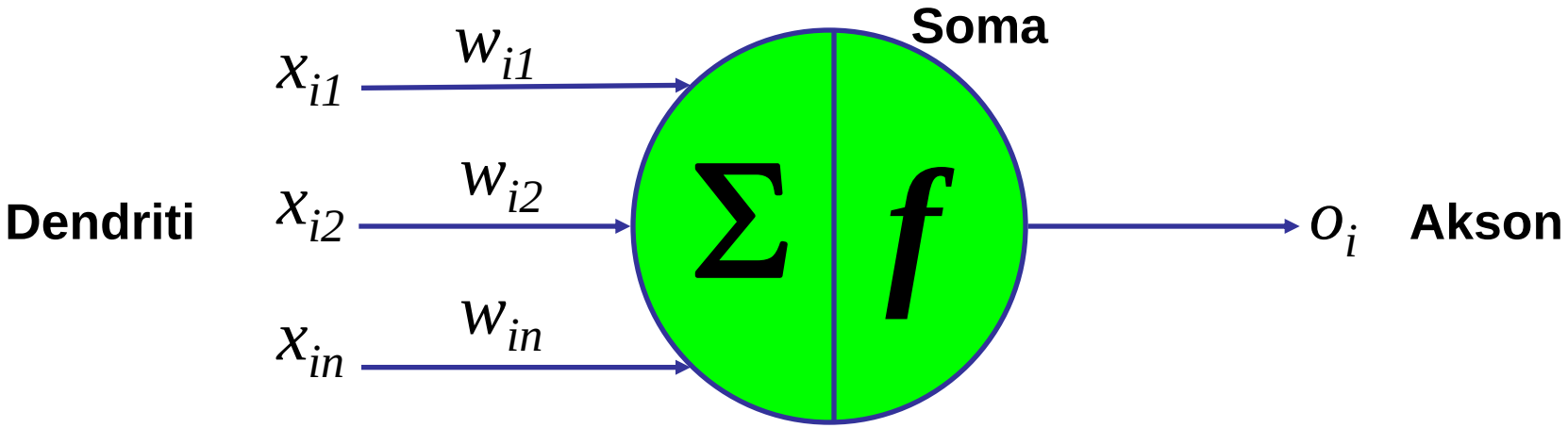
- **Veštački neuron**



- $(w_1, \ldots, w_n)$ - vektor težinskih koeficijenata
  - uobičajeno je $w_i \in [0, 1]$, ali nije obavezno
  - može biti $\sum_{j=1}^{n} w_{ij} = 1$, ali nije obavezno

Veštačke neuronske mreže

- **Veštački neuron**

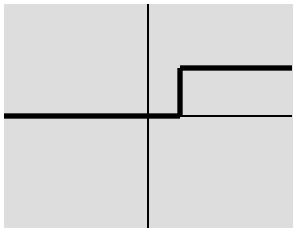


Aktivaciona funkcija neurona

$$o_i = f\left(\sum_{j=1}^{n} x_{ij} w_{ij}\right)$$

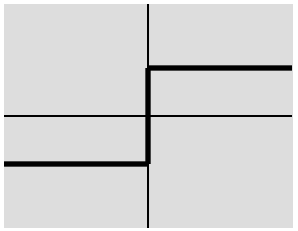
Veštačke neuronske mreže

- **Veštački neuron**
  - vrste aktivacionih funkcija
    - step
      - t - prag



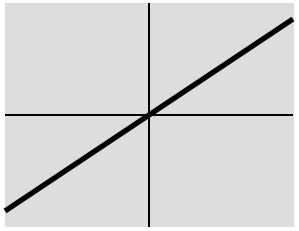
$$f(x) = \begin{cases} 1, & x \geq t \\ 0, & x < t \end{cases}$$

    - signum (sign)



$$f(x) = \begin{cases} +1, & x \geq 0 \\ -1, & x < 0 \end{cases}$$

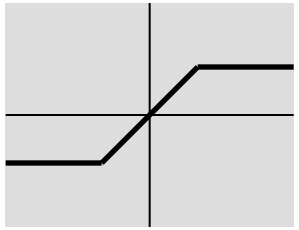
    - linearna



$$f(x) = ax + b$$

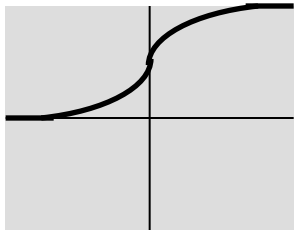
Veštačke neuronske mreže

- **Veštački neuron**
  - vrste aktivacionih funkcija
    - linearno-saturacijska



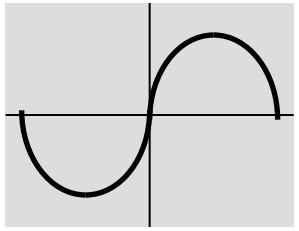
$$f(x) = \begin{cases} 1, & x \geq +1 \\ x, & -1 \leq x \leq 1 \\ -1, & x < -1 \end{cases}$$

    - sigmoid



$$f(x) = \frac{1}{1+e^{-x}}$$

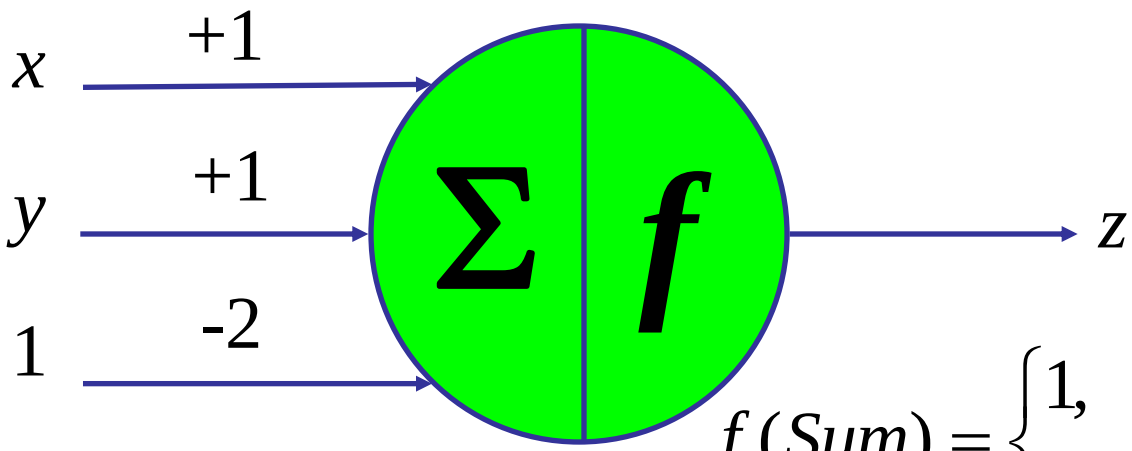
    - sinusna



$$f(x) = \mathrm{a}\ sin(\mathrm{b}x + \mathrm{c})$$

Veštačke neuronske mreže

- **ANN - primer**
  - realizacija funkcije **"logički AND"**: z = x ∧ y

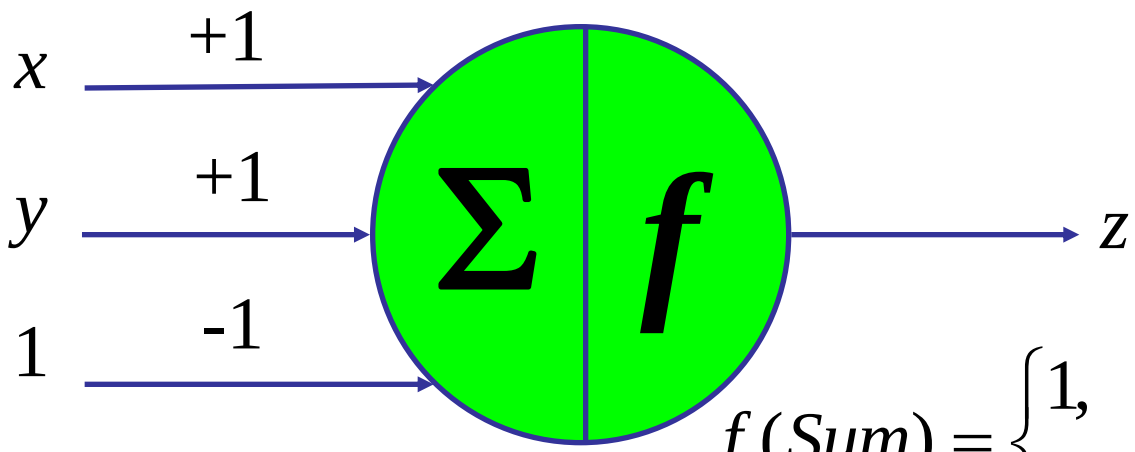


$$f(Sum) = \begin{cases} 1, & Sum \geq 0 \\ 0, & Sum < 0 \end{cases}$$

$$Sum(x, y) = x \cdot 1 + y \cdot 1 - 1 \cdot 2$$

Veštačke neuronske mreže

- **ANN - primer**
  - realizacija funkcije **"logički OR"**: z = x ∨ y



$$f(Sum) = \begin{cases} 1, & Sum \geq 0 \\ 0, & Sum < 0 \end{cases}$$

$$Sum(x, y) = x + y - 1$$

Veštačke neuronske mreže

- **Učenje neuronske mreže**
  - **Ideja učenja**
    - čovek percipira impulse (informacije) iz okruženja
    - poredi ih sa prethodno spoznatim vrednostima
    - koriguje prethodno spoznate vrednosti
      - ne samo da bi delovao,
      - već i da bi **povećao svoje znanje i sposobnosti kako da deluje u budućnosti**

Veštačke neuronske mreže

- **Učenje neuronske mreže**
  - **Operacije (aktivnosti) učenja**
    - kreiranje ili brisanje neurona
    - kreiranje ili brisanje veza između neurona (sinapsi)
    - modifikacija težinskih faktora grana mreže
    - modifikacija praga osetljivosti neurona
    - modifikacija ulaznih i aktivacionih funkcija neurona

Veštačke neuronske mreže

- **Učenje neuronske mreže**
  - **Uobičajeni pristup**
    - modifikacija težinskih faktora grana mreže, na osnovu opservacije izlaznog rezultata
    - pristupi učenja
      - **nadgledano učenje**
      - **nenadgledano učenje**
      - **vođeno učenje**
        - » izbor pristupa zavisi od okolnosti u kojima se mreža modelira

Veštačke neuronske mreže

- **Učenje neuronske mreže**
  - **Pristupi učenja**
    - **nadgledano učenje**
      - **supervised learning - "s učiteljem"**
      - slučajno se iz zadatog uzorka (domena) biraju ulazne veličine
      - slučajno se biraju težinski koeficijenti
      - data je vrednost izlazne veličine koja se smatra korektnom
        - » poredi se dobijeni izlaz sa zadatom vrednošću koja se smatra korektnom i izračunava se razlika
        - » na osnovu izračunate razlike, modifikuju se vrednosti težinskih koeficijenata, "unazad" kroz mrežu

Veštačke neuronske mreže

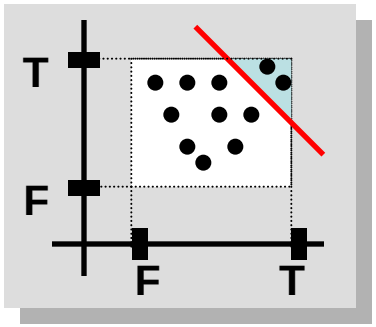
- **Učenje neuronske mreže**
  - **Pristupi učenja**
    - **nenadgledano učenje**
      - **unsupervised learning - "bez učitelja"**
      - slučajno se iz zadatog uzorka biraju ulazne veličine
      - slučajno se biraju težinski koeficijenti
      - ne postoje nikakve informacije o tome šta treba da predstavlja korektan izlaz
        - » porede se pojedinačni izlazi neurona s dobijenim izlazom
        - » izmenom težinskih koeficijenata, dodatno se favorizuje neuron s najvećim doprinosom izlaznoj veličini

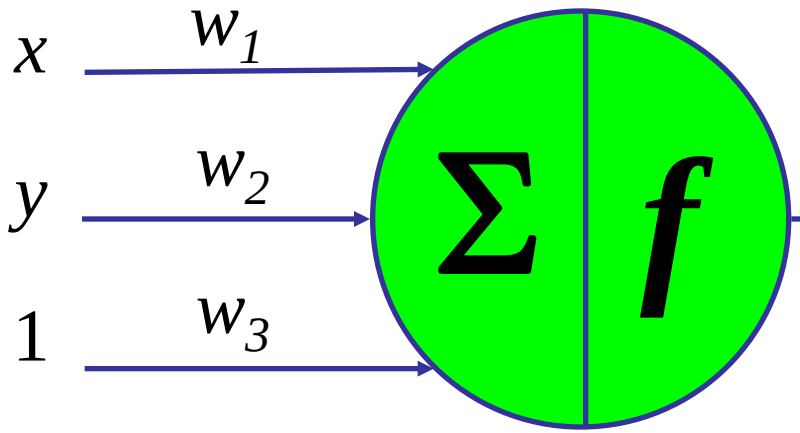
Veštačke neuronske mreže

- **Učenje neuronske mreže**
  - **Pristupi učenja**
    - **vođeno učenje**
      - **reinforcement learning - uz delimičnu pomoć učitelja**
      - ne postoji podatak koji treba da predstavlja korektan izlaz
      - postoje informacije o udaljenosti dobijenog izlaza od očekivane (korektne) vrednosti
        - » učitelj za dobijeni izlaz daje ocenu tipa "toplo-hladno", kojom se ocenjuje udaljenost izlaza od od korektne vrednosti
        - » na osnovu tih ocena, modifikuju se težinski koeficijenti grana

Veštačke neuronske mreže

- **ANN - primer učenja**
  - zadatak
    - da ANN nauči da primenjuje funkciju z = x ∧ y
      - na osnovu statistički branih podataka iz domena
      - s visokom verovatnoćom "pogodaka"





$$f(Sum) = \begin{cases} 1, & Sum \geq 0 \\ 0, & Sum < 0 \end{cases}$$

$$Sum(x, y) = x \cdot w_1 + y \cdot w_2 + 1 \cdot w_3$$

Veštačke neuronske mreže

- **ANN – osnovni koraci primene**
  - **Priprema podataka**
    - akvizicija i prethodna obrada podataka
  - **Konstrukcija mreže**
    - identifikacija promenljivih, slojeva i neurona u slojevima
  - **Učenje (obučavanje mreže) i testiranje**
    - nad testnim uzorkom podataka
  - **Sažimanje mreže**
    - optimizacija rada mreže – povratna veza na konstrukciju
  - **Interpretacija ("izvršenje") mreže**
    - primena mreže u radnim uslovima

Veštačke neuronske mreže

- **ANN - Prednosti primene**
  - Koncept, pogodan za modelovanje **složenih i nelinearnih sistema**
    - sposobnost upravljanja **velikim brojem promenljivih i parametara**
    - sposobnost procesiranja **velike količine podataka**
    - moguća upotreba promenljivih nad domenima, zasnovanim na **raznorodnim tipovima podataka i ograničenjima**
      - postoje postupci za prevođenje nenumeričkih promenljivih u numeričke promenljive

Veštačke neuronske mreže

- **ANN - Prednosti primene**
  - Koncept, pogodan za modelovanje **složenih i nelinearnih sistema**
    - sposobnost **razlikovanja korisnih podataka od nekorisnih** (tzv. "smetnji - šuma")
    - visoka **otpornost na pojavu grešaka** u učenju, tj. u testnom uzorku podataka
  - Sposobnost **mnogo bržeg odziva sistema**
    - u odnosu na primenu drugih metoda, u rešavanju istih problema

Veštačke neuronske mreže

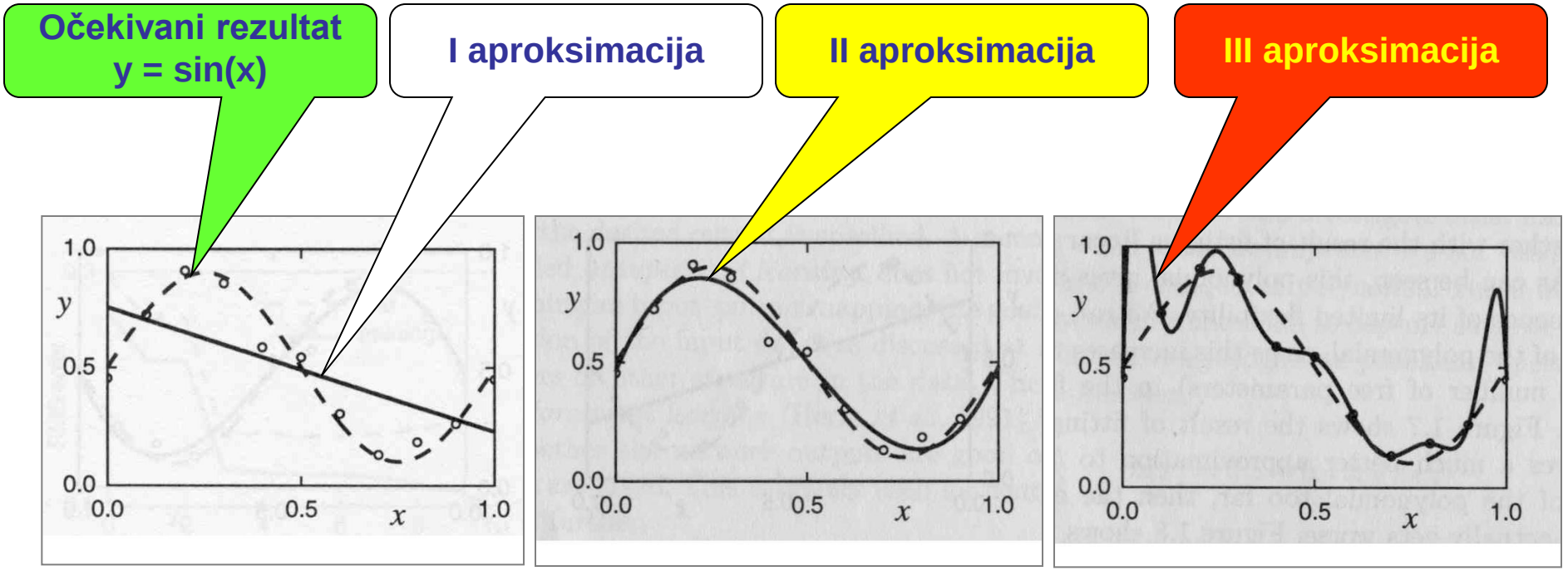
- **ANN - Problemi**
  - Nepogodnost za grafičko prezentovanje kompletne funkcionalnosti
  - Sistem, modelovan putem ANN, često, **nije lako percipirati i razumeti,** samo na osnovu analize mreže
  - Osposobljavanje mreže (obučavanje) je **dugotrajan** (vremenski zahtevan) proces
  - Priprema podataka može biti **složen i dugotrajan** proces
    - sam proces pripreme podataka se, često, modelira putem drugih neuronskih mreža

Veštačke neuronske mreže

- **ANN - Problemi**
  - **"Overlearning"**
    - preveliko upuštanje u analizu detalja
      - » učenje premnogo pojedinačnih primera "napamet"
    - gubi se sposobnost generalizacije pri rezonovanju

Veštačke neuronske mreže

- **ANN**
  - **"Overlearning"** - primer aproksimacije funkcije



Veštačke neuronske mreže

- **ANN – neke oblasti primene**
  - **Problemi klasifikacije**
    - medicinske dijagnoze
    - prepoznavanje pisanog teksta (npr. potpisa)
      - primer: ANN za prepoznavanje ručno napisanih cifara (poštanski čekovi i zip kodovi u USA) - procenat tačnosti 99,33%
    - prepoznavanje zvučnih zapisa ili slika
    - ocena rizika kreditiranja, osiguranja
  - **Problemi predviđanja**
    - predviđanje prodaje, marketinška predviđanja
    - vremenska prognoza

Veštačke neuronske mreže

- **ANN – neke oblasti primene**
  - **Problemi modelovanja ponašanja**
    - modelovanje procesa (dinamika sistema)
      - upravljanje robotskim sistemima
  - **Problemi aproksimacije funkcija**

Sadržaj

- Data Mining
- Prebrojavanje sličnih pojava
- Istraživanje zakonitosti
- Pravila, strukturirana u obliku stabla
- Klasterizacija
- Pronalaženje sličnih vremenskih serija
- Veštačke neuronske mreže

Pitanja i komentari

Sistemi za istraživanje podataka

Data Mining (*DM*)