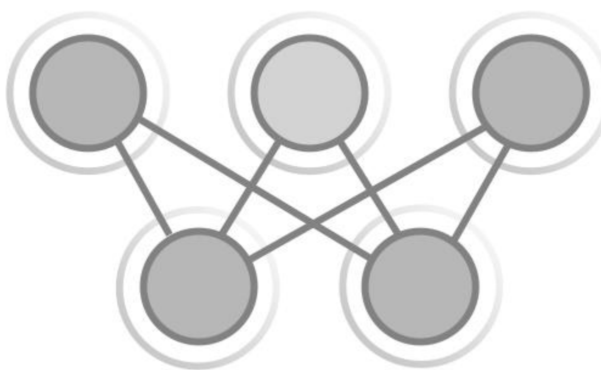
Prof.dr.sc. Bojana Dalbelo Bašić

Fakultet elektrotehnike i računarstva
Zavod za elekroniku, mikroelektroniku, računalne i inteligentne sustave

www.zemris.fer.hr/~bojana
bojana.dalbelo@fer.hr

Učenje na temelju primjera

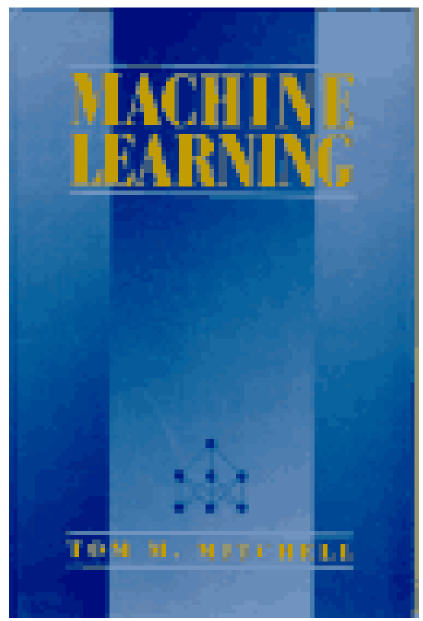


FER
Siječanj2010

Literatura za predavanje

- *Chapter 8*

  Instance-Based Learning



Instance-Based Learning = Učenje na temelju primjera

ili

- Memory based learning algorithms
- Lazy methods
- Nonparametric methods (E. Alpaydin, Chapter 8)

Učenje na temelju primjera

- Dosadašnje metode nastoje konstruirati eksplicitni opis ciljne funkcije.
- **Metode učenja na temelju primjera** pohranjuju primjere za učenje.
- Postupak generalizacije odgođen je do trenutka potrebe za klasifikacijom novog uzorka.
  - Metoda *k*-najbližih susjeda *(engl. k-nearest-neighbor)*
  - Metoda lokalne regresije s težinskim faktorima *(engl. weighted regression)*
  - Zaključivanje na temelju slučajeva *(engl. case based reasoning)*
  - Radijalne bazne funkcije *(engl. radial basis functions)*

Lazy vs. eager

- **Lazy** (lijene) metode
  - Odgađaju odluku o klasifikaciji sve do trenutka predočavanja novog primjera (upita).
  - Metoda *k* najbližih susjeda, metoda lokalne regresije s težinskim faktorima i zaključivanje na temelju slučajeva.
- **Eager** (marljive, nestrpljive) metode
  - Sve do sada iznesene metode (npr. ID3).
  - Od gornje navedenih radijalne bazne funkcije.

Prednosti lijenih metoda

- Dvije važne razlike, prednosti *lijenih* metoda, tj. metoda s odgodom, naspram ostalih metoda:
  - Konstruiraju različitu aproksimaciju ciljne funkcije za svaki različiti novi upit–primjer koji treba biti klasificiran.
  - Umjesto procjene ciljne funkcije, jednom za cijeli prostor, te metode procjenjuju ciljnu funkciju samo lokalno, u okolini novog primjera. Takva lokalna procjena ciljne funkcije je pogodna za vrlo kompleksne ciljne funkcije.
- Nedostatak metoda učenja na temelju primjera:
  - visoka cijena klasificiranja novog primjera.
  - razmatraju se svi atributi nekog primjera prilikom klasifikacije iako samo neki mogu imati utjecaj na ciljnu funkciju (k-najbližih susjeda).

Metoda k najbližih susjeda

- *Engl. k-nearest-neighbors*, skraćeno k-nn.
- Ideja je da se novi primjer klasificira tako da se pogledaju njemu najbliži primjeri iz skupa za učenje.
- Primjeri su najčešće točke u n-dimenzionalnom prostoru R^n, a za račun udaljenosti koristi se Euklidska metrika.
  - Moguće je da primjeri budu npr. nizovi znakova, a za udaljenost da se koristi Levensteinova udaljenost.
- Zadatak:
  - Klasifikacija – vrijednosti ciljne funkcije su iz konačnog skupa.
  - Regresija – ciljna funkcija poprima realne vrijednosti.

Metoda k najbližih susjeda

- Klasifikacija točaka iz prostora R^n korištenjem Euklidske metrike.
- Primjer **x** opisan je vektorom značajki

  $[a_1(x), a_2(x), \ldots, a_n(x)]$.
- Euklidska udaljenost dva vektora x_i i x_j je

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{n} \left[a_r(x_i) - a_r(x_j)\right]^2}$$

- Za ciljnu funkciju sa diskretnim vrijednostima:

  $f: R^n \to V$, gdje je $V = \{v_1, v_2, \ldots, v_s\}$

Algoritam k najbližih susjeda (k-nn)

- Algoritam za učenje
  - Za svaki primjer za učenje (x,f(x)) dodaj primjer na listu primjeri_za_učenje.

- Algoritam klasifikacije
  - Za dani primjer x_q s nepoznatom klasifikacijom
  - Neka $x_1, x_2, \ldots, x_k$ označavaju k primjera koji su najbliži x_q.
- Vrati

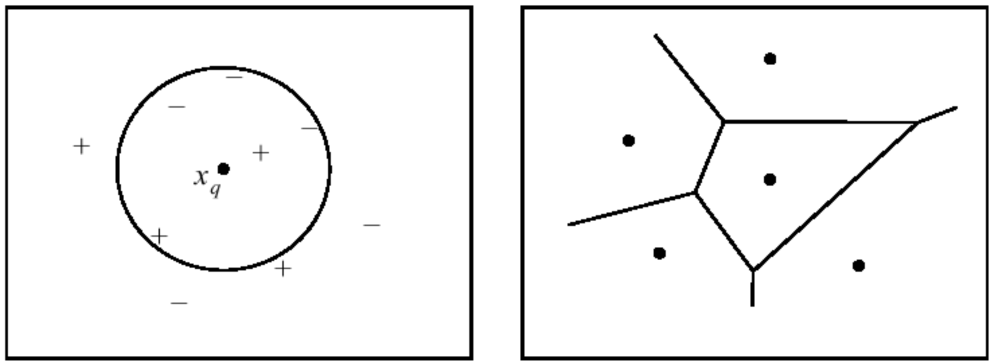
$$\hat{f}(x_q) \leftarrow \arg\max_{v \in V} \sum_{i=1}^{k} \delta(v, f(x_i))$$

gdje je $\delta(a,b)=1$ ako a=b, 0 inače.

- $\hat{f}(x_q)$ je najčešća vrijednost ciljne funkcije koja se pojavljuje među k primjera za učenje koji su najbliži upitu x_q

Algoritam k najbližih susjeda (k-nn)

- Razlika kod 1-nn i 5-nn algoritma:



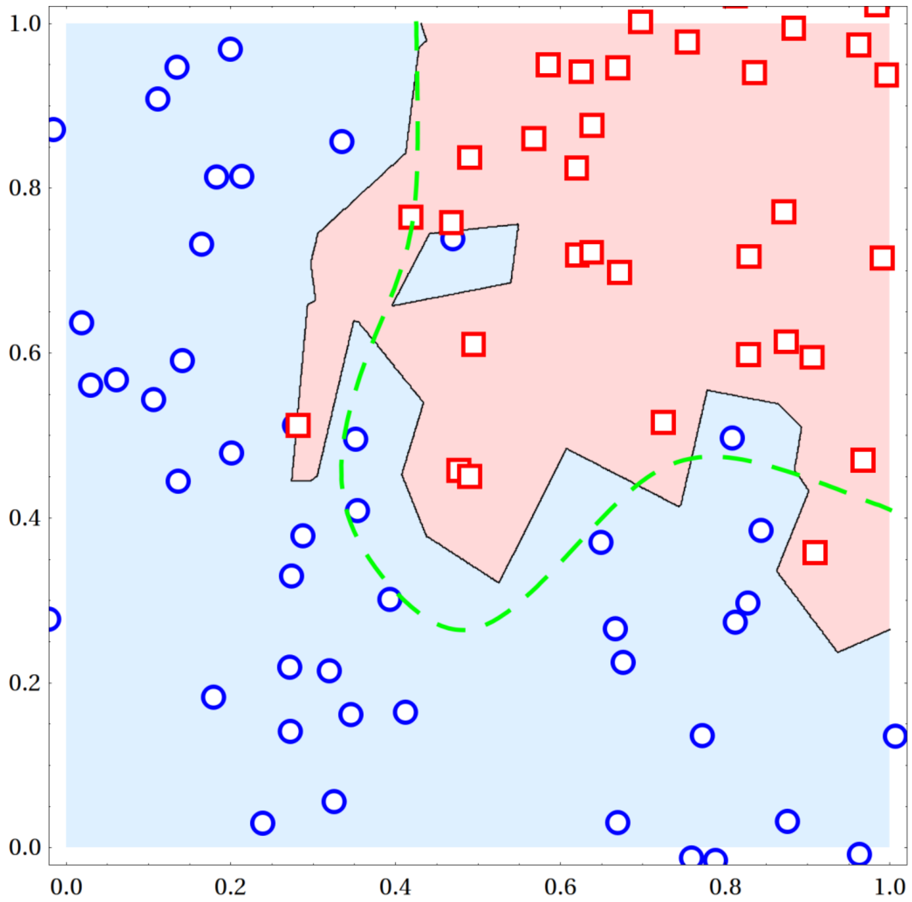
- k-nn algoritam nikad ne oblikuje eksplicitnu hipotezu za ciljnu funkciju f.
- Za 1-nn možemo je predočiti Voronoi dijagramom. Decizijska površina je konveksni poliedar koji okružuje svaki primjer za učenje.

Primjer rada k-nn algoritma

- Na primjeru će se vidjeti da se rastom k smanjuje varijanca, ali i povećava se pristranost.
- Uz poznatu pravu distribuciju svih primjera konstruiran je optimalan klasifikator čija je decizijska granica prikazana zelenom iscrtkanom crtom.

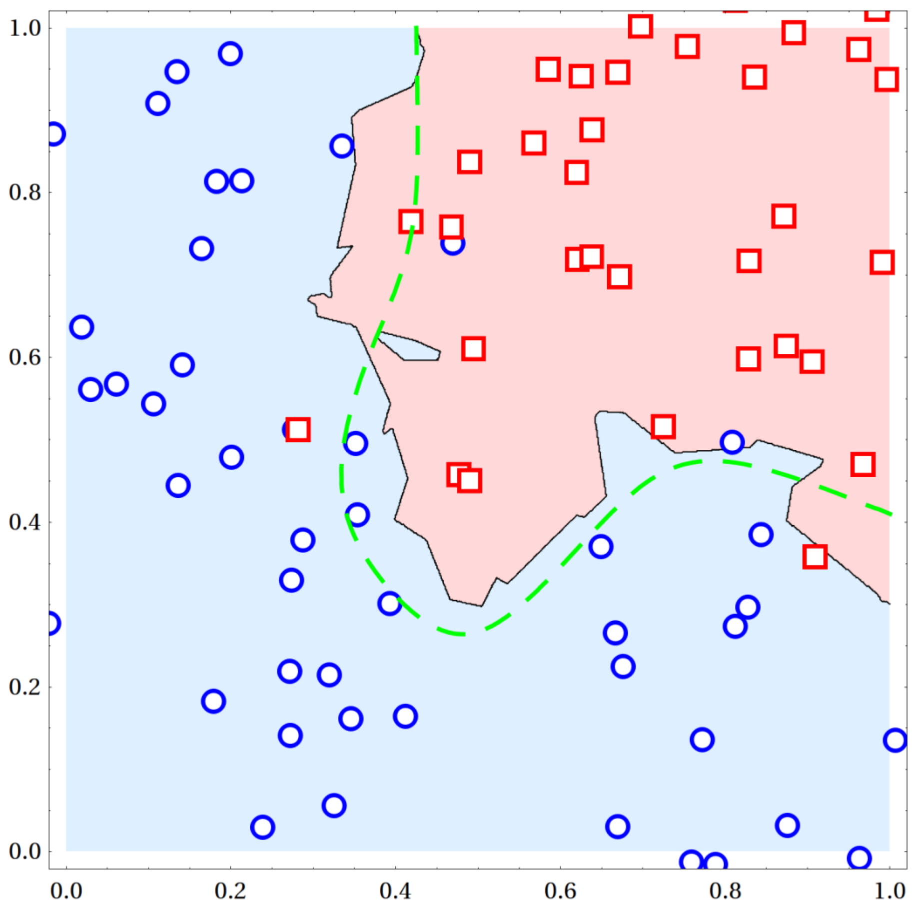
Primjer rada k-nn algoritma

- k = 1, 100 primjera za učenje



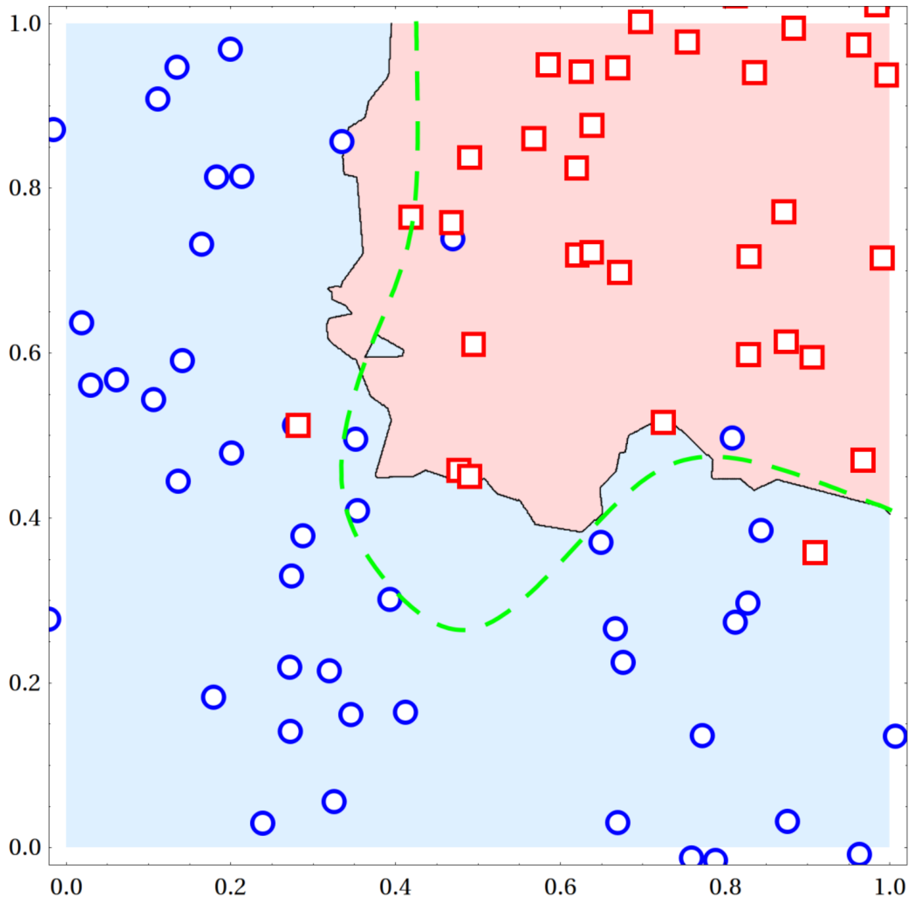
Primjer rada k-nn algoritma

- k = 3, 100 primjera za učenje



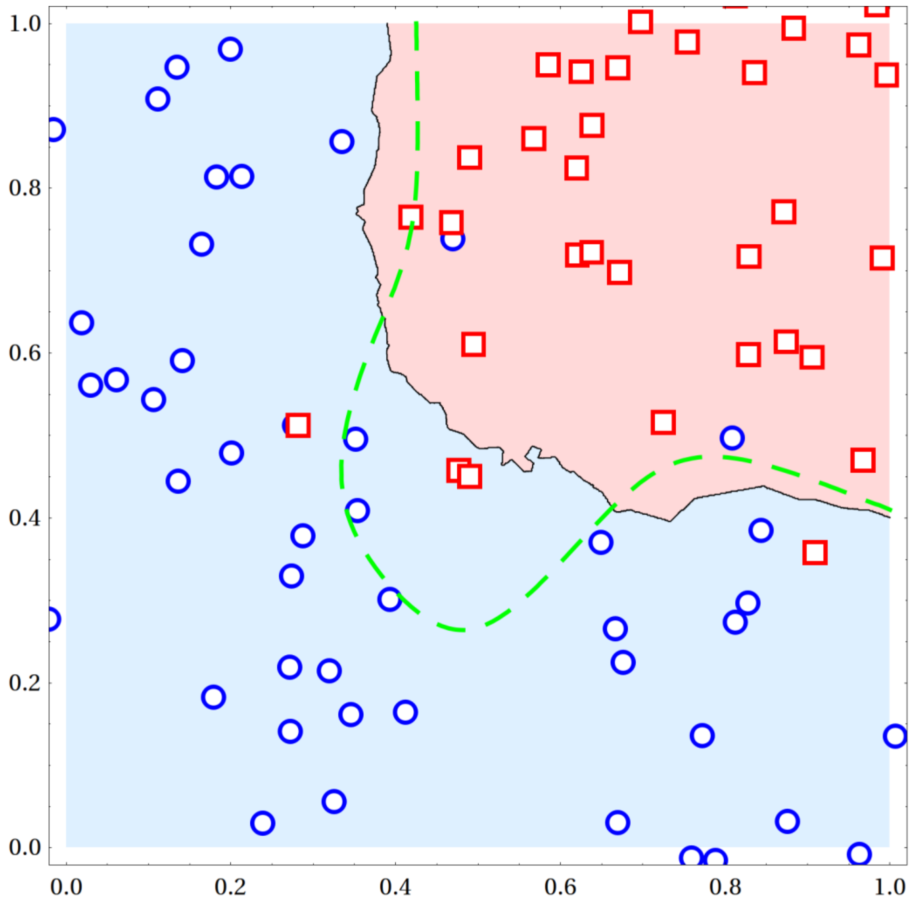
Primjer rada k-nn algoritma

- k = 5, 100 primjera za učenje



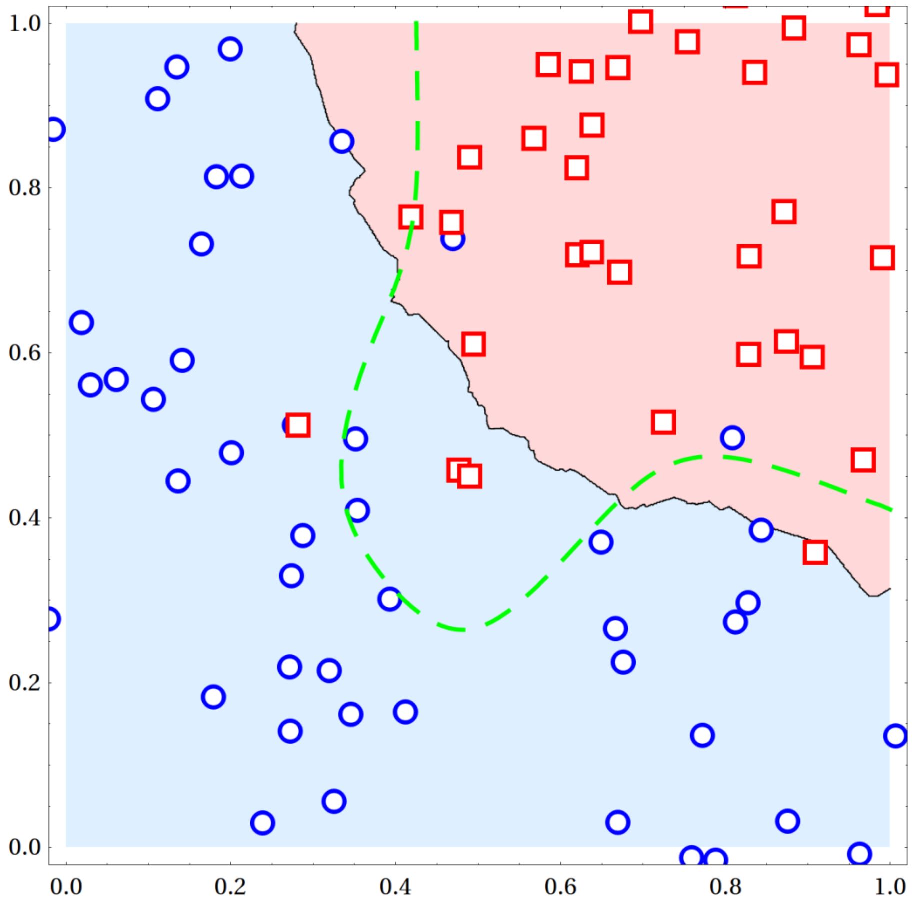
Primjer rada k-nn algoritma

- k = 15, 100 primjera za učenje



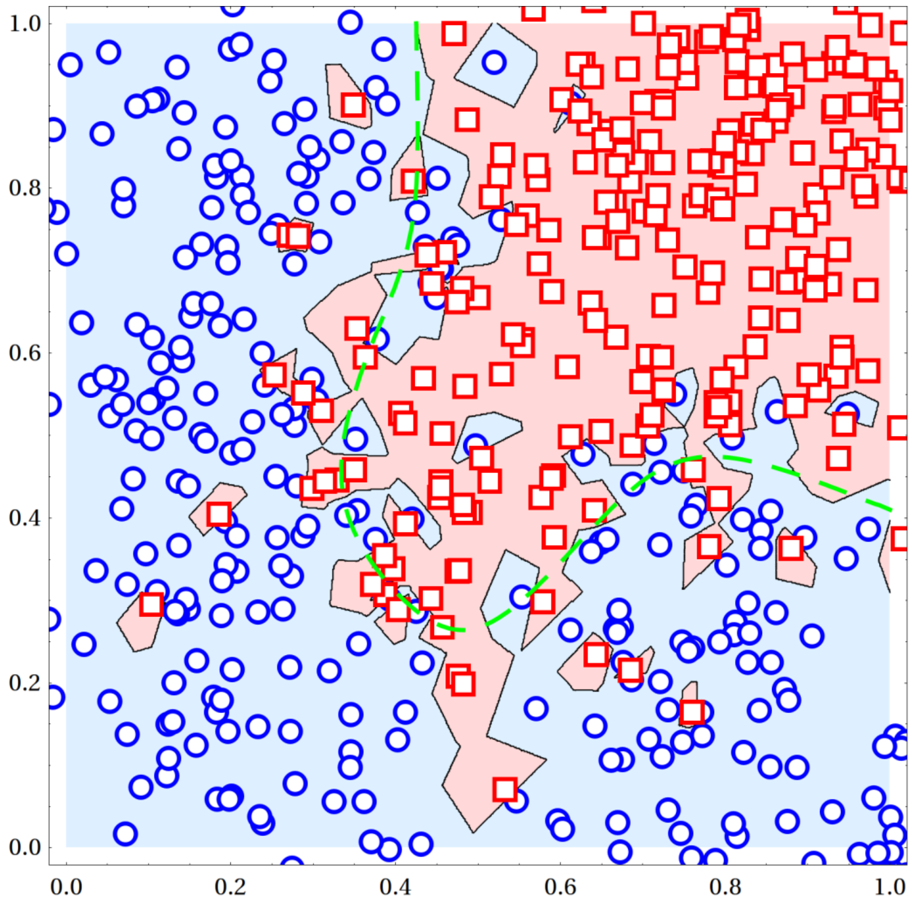
Primjer rada k-nn algoritma

- k = 31, 100 primjera za učenje



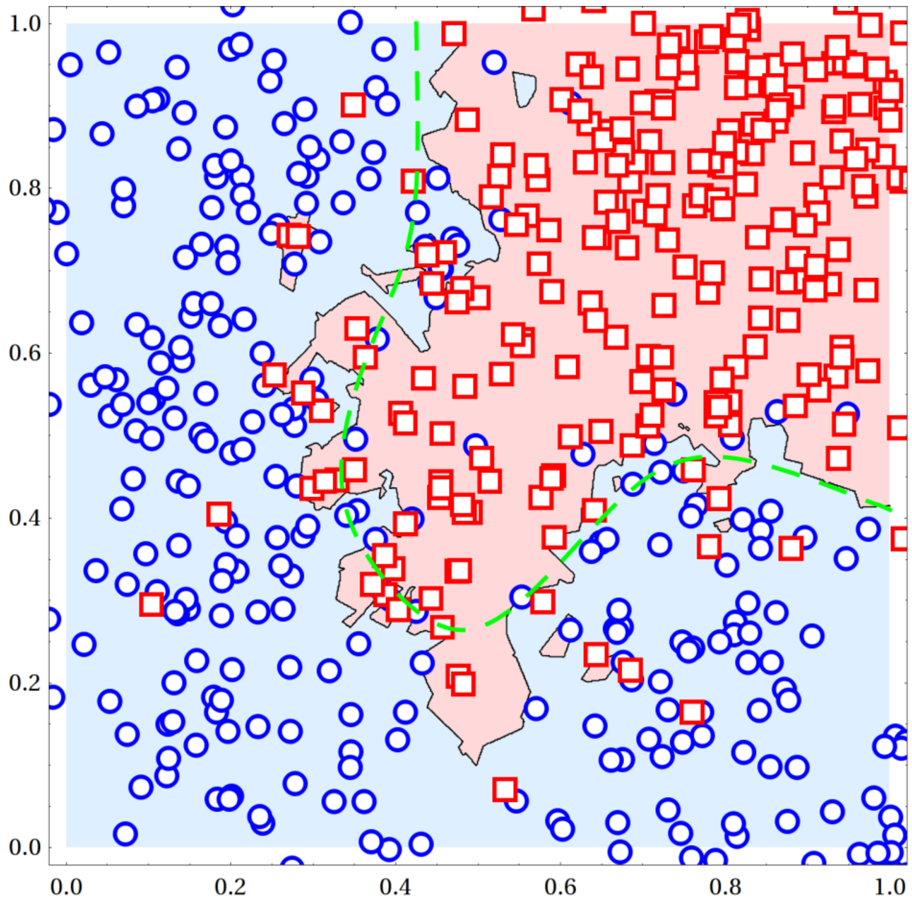
Primjer rada k-nn algoritma

- k = 1, 600 primjera za učenje



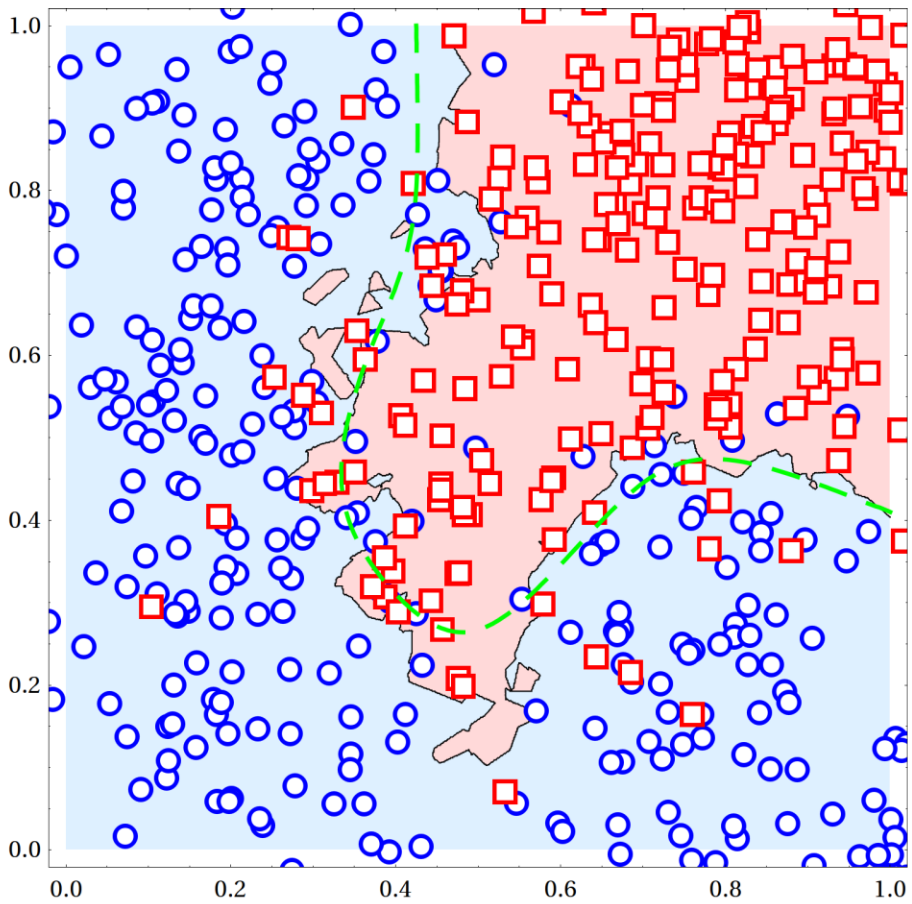
Primjer rada k-nn algoritma

- k = 3, 600 primjera za učenje



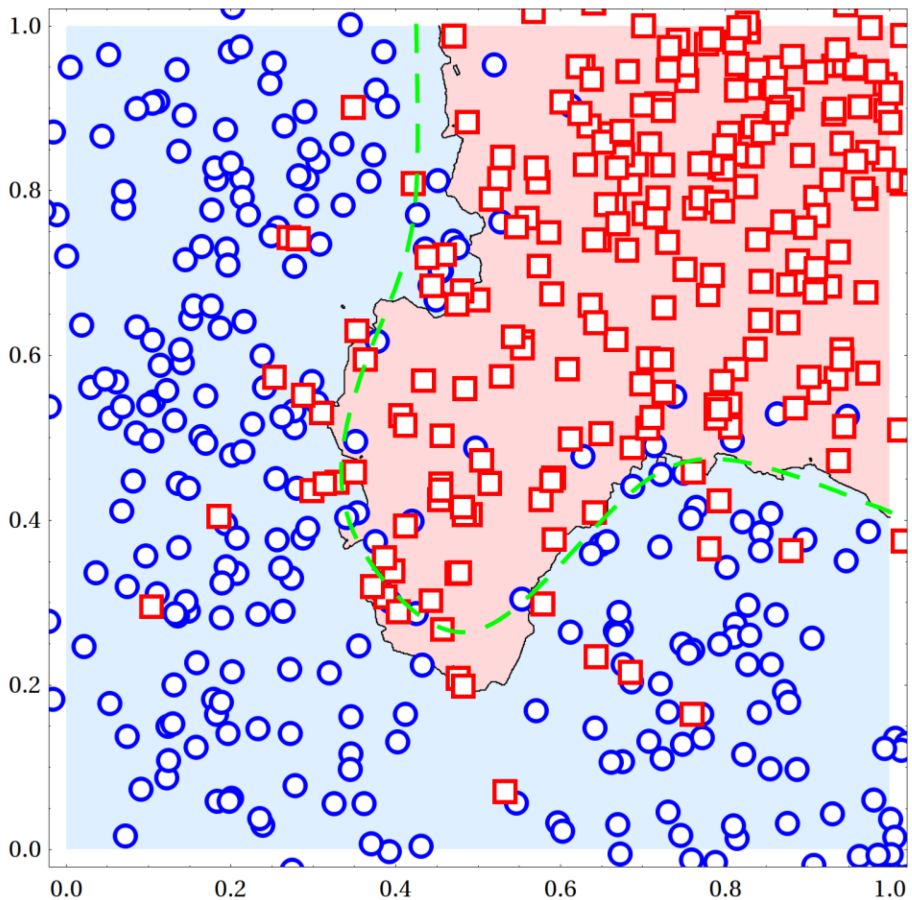
Primjer rada k-nn algoritma

- k = 5, 600 primjera za učenje



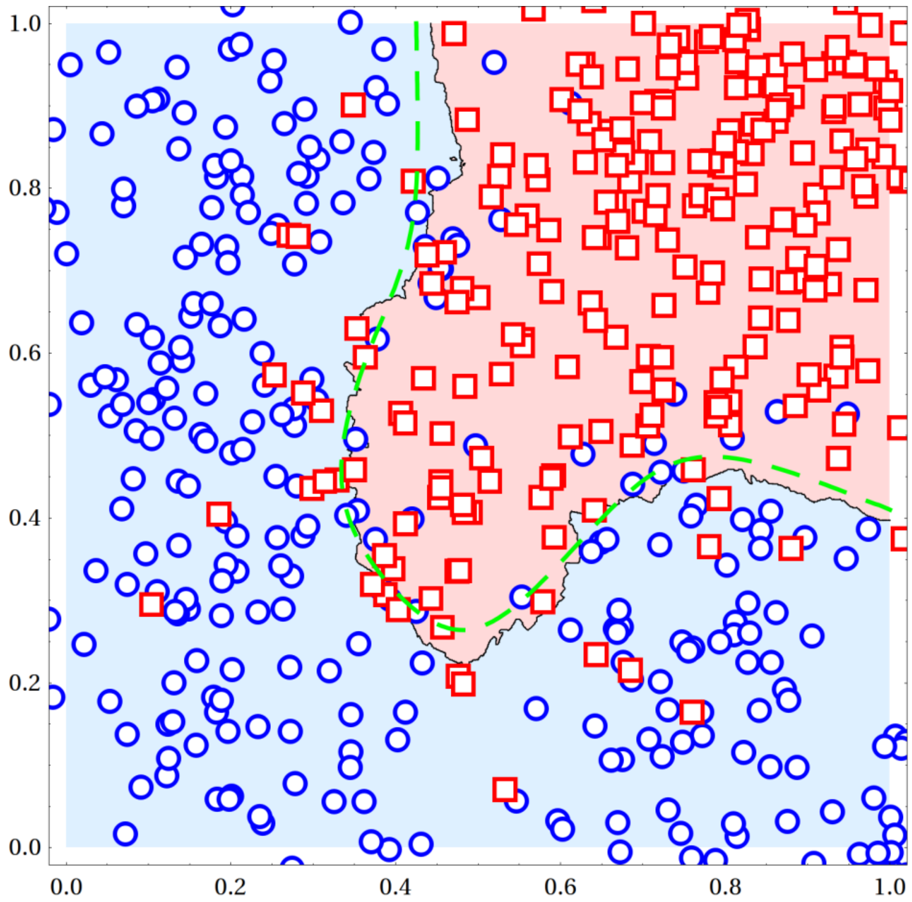
Primjer rada k-nn algoritma

- k = 15, 600 primjera za učenje



Primjer rada k-nn algoritma

- k = 31, 600 primjera za učenje



Regresija pomoću k-nn algoritma

- Umjesto najčešće pojavljivane vrijednosti ciljne funkcije odgovor na upit je srednja vrijednost ciljnih funkcija k najbližih susjeda.

$$\hat{f}(x_q) \leftarrow \frac{1}{k}\sum_{i=1}^{k} f(x_i)$$

Modifikacija k-nn algoritma uvođenjem težinskih faktora udaljenosti

- uvođenje težinskih faktora w_i za svaki od k susjeda, koji ovisi o njegovoj udaljenosti od upita x_q.

$$\hat{f}(x_q) \leftarrow \underset{v \in V}{\arg\max} \sum_{i=1}^{k} w_i \, \delta(v, f(x_i))$$

$$w_i = \frac{1}{d(x_i, x_q)^2}$$

- U slučaju $x_i = x_q$ tada pridružujemo funkciji vrijednost funkcije $f(x_i)$.

Modifikacija k-nn algoritma uvođenjem težinskih faktora udaljenosti

- U slučaju regresije (kontinuirane ciljne funkcije):

$$\hat{f}(x_q) \leftarrow \frac{\sum_{i=1}^{k} w_i \, f(x_i)}{\sum_{i=1}^{k} w_i}$$

- Ovakva globalna metoda naziva se **Shepardova** metoda.

Primjedbe na k-nn algoritam

- Prednosti:
  - efikasna induktivna metoda
  - robusna na šum u primjerima za učenje
  - (Cover i Hart) Ako broj primjera za učenje teži u beskonačno onda je greška 1-nn klasifikatora najviše dva puta veća od greške optimalnog Bayesovog klasifikatora.
- *Induktivna pristranost:*
  - pretpostavka da je klasifikacija upita x_q slična klasifikaciji primjera u blizini.
- Nije prednost:

Udaljenost se računa na temelju svih atributa (za razliku od ID3 ili učenja skupova pravila koji selektiraju podskupove atributa pri formiranju hipoteze).

Primjedbe na k-nn algoritam

- *«Curse of dimensionality»* – osjetljivost k-nn na sve atribute bez obzira na dimenziju prostora (broj atributa) i njihov značaj za ciljnu funkciju.
  - Rješenje: rastezanje ili stiskanje osi Euklidskog prostora (množenje vrijednosti atributa s faktorima) da bi se smanjio utjecaj nevažnih atributa.

- Praktična tema vezana za k-nn je efikasno indeksiranje memorije zbog brzog dohvata primjera kod novog upita.

Nazivlje

- **Metode s odgodom** - područje statističkog raspoznavanja uzoraka.
- **Regresija** – način aproksimacije ciljne funkcije s realnim vrijednostima.
- **Rezidual** (ostatak) – pogreška $\hat{f}(x) - f(x)$.
- **Jezgrena funkcija** (*engl. kernel function*) – funkcija udaljenosti koja se koristi za određivanje težinskih faktora primjera za učenje, tj. jezgrena funkcija **K** je takva da je

$$w_i = K\left(d(x_i, x_q)\right)$$

Lokalna regresija s težinskim faktorima

- *engl. locally weighted regression*
- Algoritam k-nn se može interpretirati kao aproksimiranje ciljne funkcije u f(x) u točci $x=x_q$.
- Regresija s težinskim faktorima generalizacija je te metode jer **konstruira eksplicitnu aproksimaciju ciljne funkcije na cijelom lokalnom području** oko x_q.
- Aproksimacija ciljne funkcije, može biti:
  - linearnom funkcijom
  - kvadratnom funkcijom
  - višeslojnom neuronskom mrežom

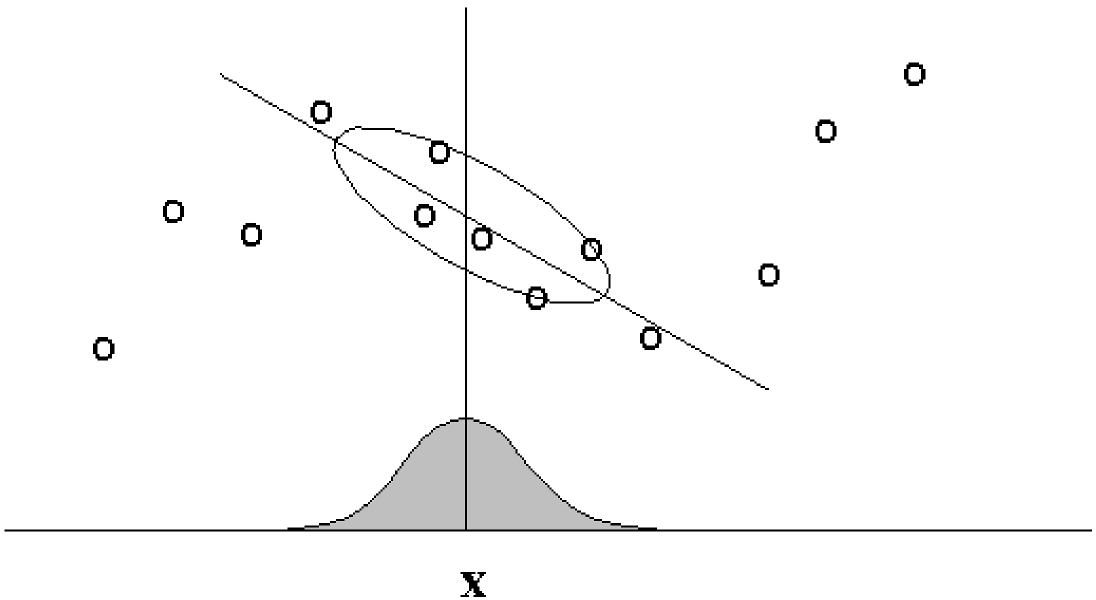
Lokalna regresija s težinskim faktorima

Lokalna regresija s težinskim faktorima

- Radi se aproksimacija samo u okolini točke upita x_q
- aproksimacija realne funkcije
- doprinos primjera za učenje ovisi o težinskom faktoru koji je funkcija udaljenosti

- Neka je dan je upit x_q
  - konstruira se aproksimacije $\hat{f}$ ciljne funkcije koja odgovara primjerima za učenje u okolini x_q
  - aproksimacija se koristi za izračun vrijednosti $\hat{f}(x_q)$

Lokalna regresija s težinskim faktorima



- f aproksimiramo linearnom funkcijom

$$\hat{f}(x) = w_0 + w_1 a_1(x) + \cdots + w_n a_n(x)$$

- $a_i(x)$ označava vrijednost i-tog atributa primjera x.

Lokalna regresija s težinskim faktorima

- Metoda globalne aproksimacije:

$$E = \frac{1}{2}\sum_{x\in D}\left(f(x) - \hat{f}(x)\right)^2$$

- Tri moguća kriterija prilagodbe ove metode za lokalnu aproksimaciju:
  1. Minimizacija kvadrata pogreške samo nad k najbližih susjeda.

$$E_1(x_q) = \frac{1}{2}\sum_{\substack{x\in \text{k-najbližih} \\ \text{susjeda od } x_q}}\left(f(x) - \hat{f}(x)\right)^2$$

Lokalna regresija s težinskim faktorima

2. Minimizacija kvadrata pogreške nad ciljnim skupom D uz umnožak s težinskim faktorima

$$E_2(x_q) = \frac{1}{2} \sum_{x \in D} \left(f(x) - \hat{f}(x)\right)^2 K\left(d(x_q, x)\right)$$

3. Kombinacija 1. i 2.

$$E_3(x_q) = \frac{1}{2} \sum_{\substack{x \in \text{k najbližih} \\ \text{susjeda od } x_q}} \left(f(x) - \hat{f}(x)\right)^2 K\left(d(x_q, x)\right)$$

Lokalna regresija s težinskim faktorima

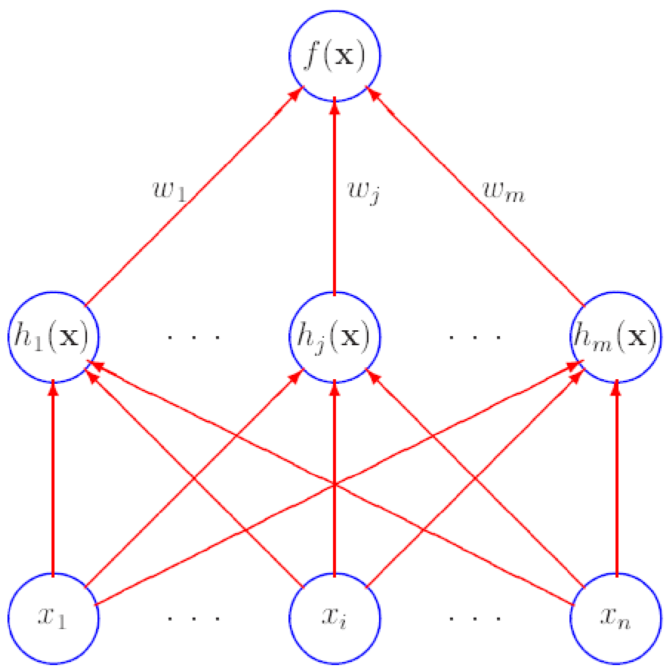
- Model pod 2 je računski najzahtjevniji.
- Ako usvojimo 3. model, pravilo učenja je:

$$\Delta w_j = \eta \sum_{\substack{x \in \text{k najbližih} \\ \text{susjeda od } x_q}} K\left(d\left(x_q, x\right)\right)\left(f(x) - \hat{f}(x)\right) a_j(x)$$

- Napomena:
  - Postoji niz varijanti metode linearne regresije s težinskim faktorima. Funkcija f je u našem slučaju linearna no koriste se još i kvadratna aproksimacijska funkcija, ali ne i složenije zbog cijene koja bi se platila za izračunavanje takve funkcije za svaki pojedini upit.

Radijalne bazne funkcije

- Metoda aproksimacije funkcije (povezana sa k-nn i lokalnom regresijom).
- Vrsta umjetnih neuronskih mreža, jedan skriveni sloj koji implementira tzv. jezgrenu funkciju h ili K (zavisi o primjeni)



- RBF Imaju izvrsna svojstva, mogu aproksimirati nelinearna preslikavanja

Radijalne bazne funkcije

- Hipoteza je funkcija oblika:

$$\hat{f}(x) = w_0 + \sum_{u=0}^{k} w_u K_u\big(d(x_u, x)\big) \qquad (1)$$

gdje su:

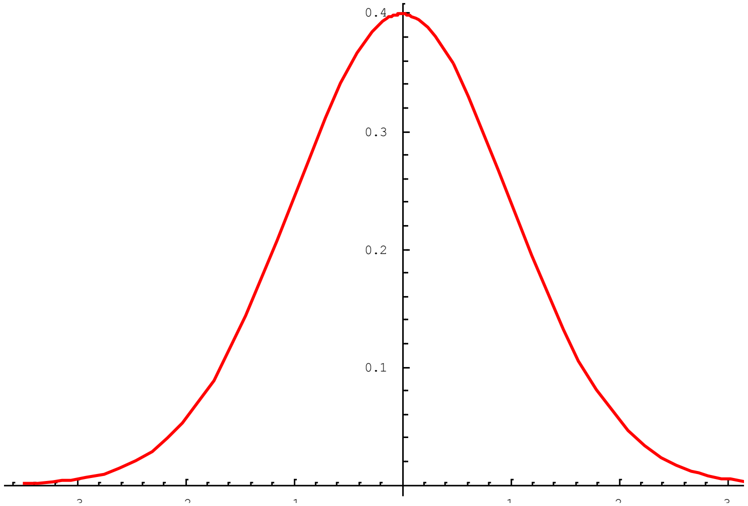
- x_u primjeri za učenje,
- $K(d(x_u,x))$ jezgrena funkcija čija se vrijednost smanjuje kada udaljenost raste od centroida x_u raste.
- k proizvoljan broj jezgrenih funkcija

- Iako je $\hat{f}(x)$ globalna aproksimacija f(x), doprinos svake $K_u(d(x_u,x))$ je lokalan - samo u okolini x_u.

Radijalne bazne funkcije

- Uobičajen izbor za $K_u(d(x_u,x))$ su Gaussove funkcije s centrom u x_u i varijancom σ^2.

$$K_u(d(x_u,x)) = e^{-\frac{1}{2\sigma_u^2}d^2(x_u,x)}$$



x_u

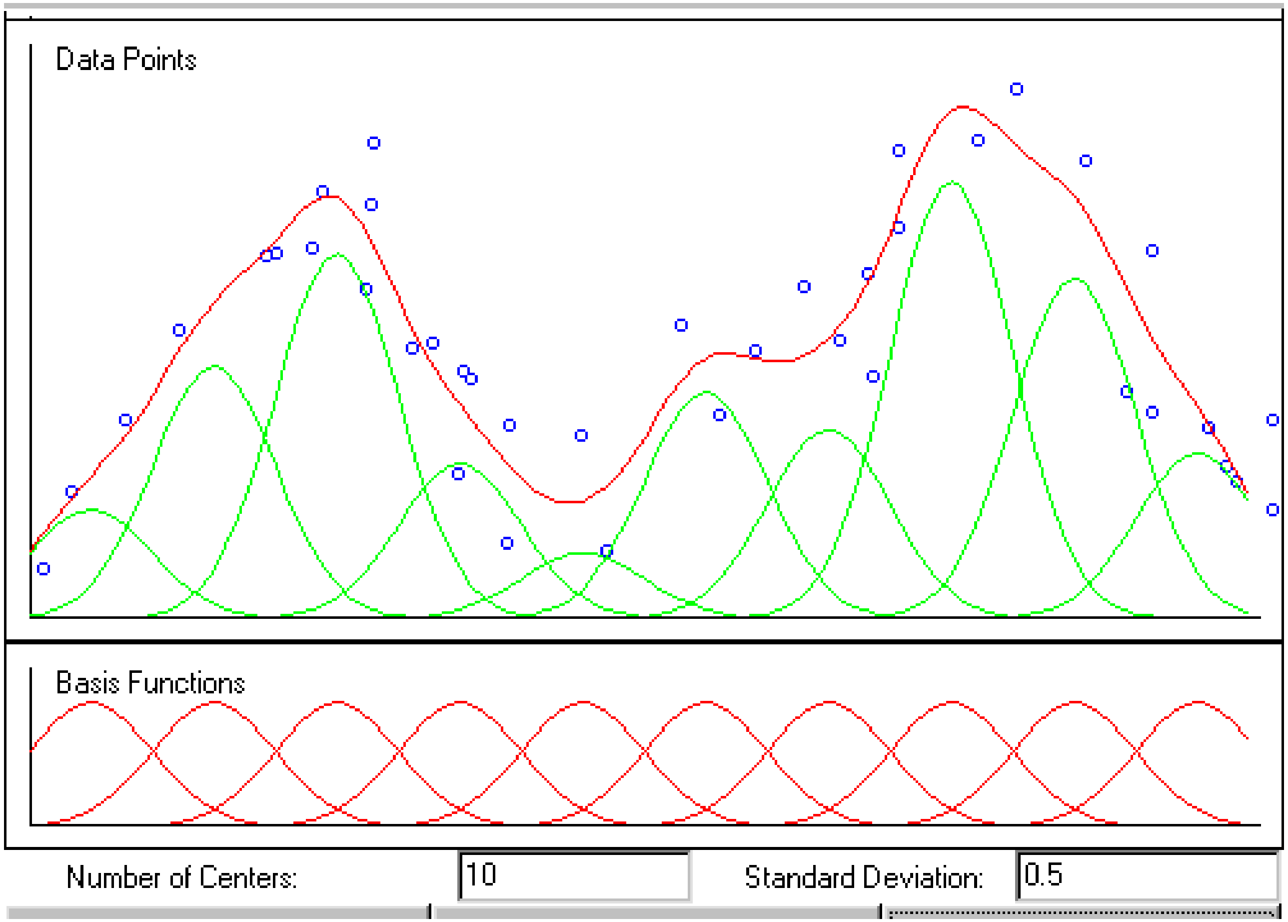
Radijalne bazne funkcije

- Prema (Hartman *et al.*, 1990) izraz

$$\hat{f}(x) = w_0 + \sum_{u=0}^{k} w_u K_u\big(d(x_u, x)\big)$$

može aproksimirati bilo koju funkciju proizvoljno točno za dovoljno veliki broj Gaussovih jezgri uz uvjet da se varijance mogu nezavisno odrediti.

Radijalne bazne funkcije



http://diwww.epfl.ch/mantra/tutorial/english/rbf/html/index.html

Radijalne bazne funkcije

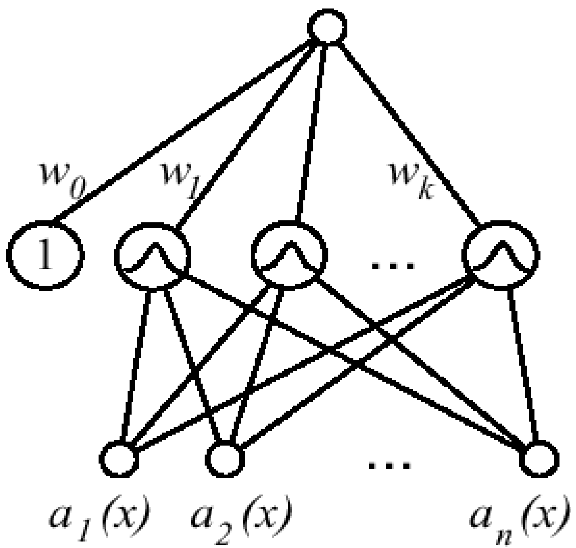
- Funkcija (1) se može interpretirati kao dvoslojna neuronska mreža prvi sloj računa $K_u(d(x_u,x))$, drugi sloj je linearna kombinacija Gaussove distribucije vjerojatnosti, vrijednosti prvog sloja – od kojih svaki K modelira jednu grupu (klaster) podataka za učenje.

- Kod klasifikacije, K () se interpretira kao *posteriorna vjerojatnost*

  K () -> p(klaster k | x),

  a težine w_i kao *posteriorna vjerojatnost*

  w_i -> p(klasa Ci | klaster k)



Radijalne bazne funkcije

- Postoji više algoritama za učenje RBF-a temeljenih na nadzirano učenje, metoda najmanjih kvadrata

$$\min \sum_i (\widehat{f}(x_i) - y_i)^2$$

- Modificirani Back propagation, Ortogonalni najmanji kvadrati, Hibridna strategija
- Hibridna strategija - RBF mreže se treniraju u dva koraka:
  1. određuje se broj skrivenih jedinica k i parametri jezgrene funkcije K(), tj. određuje se centroidi x_u i varijance σ. (nenadzirana faza)
     određuju se težinski faktori w_i tako da mreža odgovara podacima za učenje – na temelju minimizacije sume kvadrata pogreške.
     Za vrijeme te faze jezgrene funkcije se ne mijenjaju pa je učenje efikasno.

Radijalne bazne funkcije

- Nekoliko metoda za izbor broja k:
  1. za svaki primjer za učenje $(x_i,f(x_i))$ – jedna jezgrena funkcija s centrom u x_i i sa istim varijancama.
     na ovaj način RBFu potpunosti odgovara primjerima za učenje
  2. broj jezgrenih funkcija < broj primjera za učenje. efikasniji način.

  Centri RBFa mogu biti smješteni
  - uniformno po X
  - neuniformno,
  - slučajnim izborom, izvlačeći primjere iz skupa za učenje u skladu s njihovom distribucijom
  - prototipovima grupa primjera za učenje (uz uporabu algoritma grupiranja)

Radijalne bazne funkcije

- Zaključak:
  - RBF daju globalnu aproksimaciju ciljne funkcije kao **linearnu kombinaciju više lokalnih jezgrenih funkcija**.
  - mogu biti **trenirane efikasnije** od unaprijednih neuronskih mreža s *backpropagation* algoritmom (algoritam radi u dva koraka).