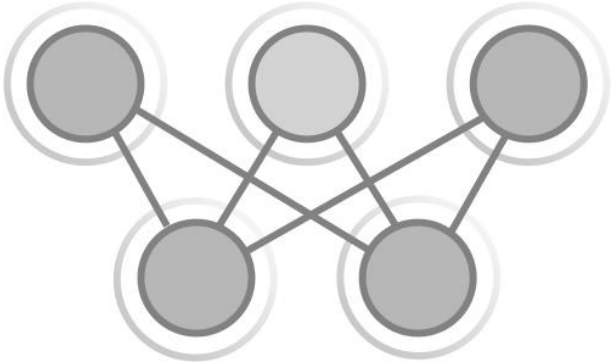
Prof.dr.sc. Bojana Dalbelo Bašić

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

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

Nadzirano učenje

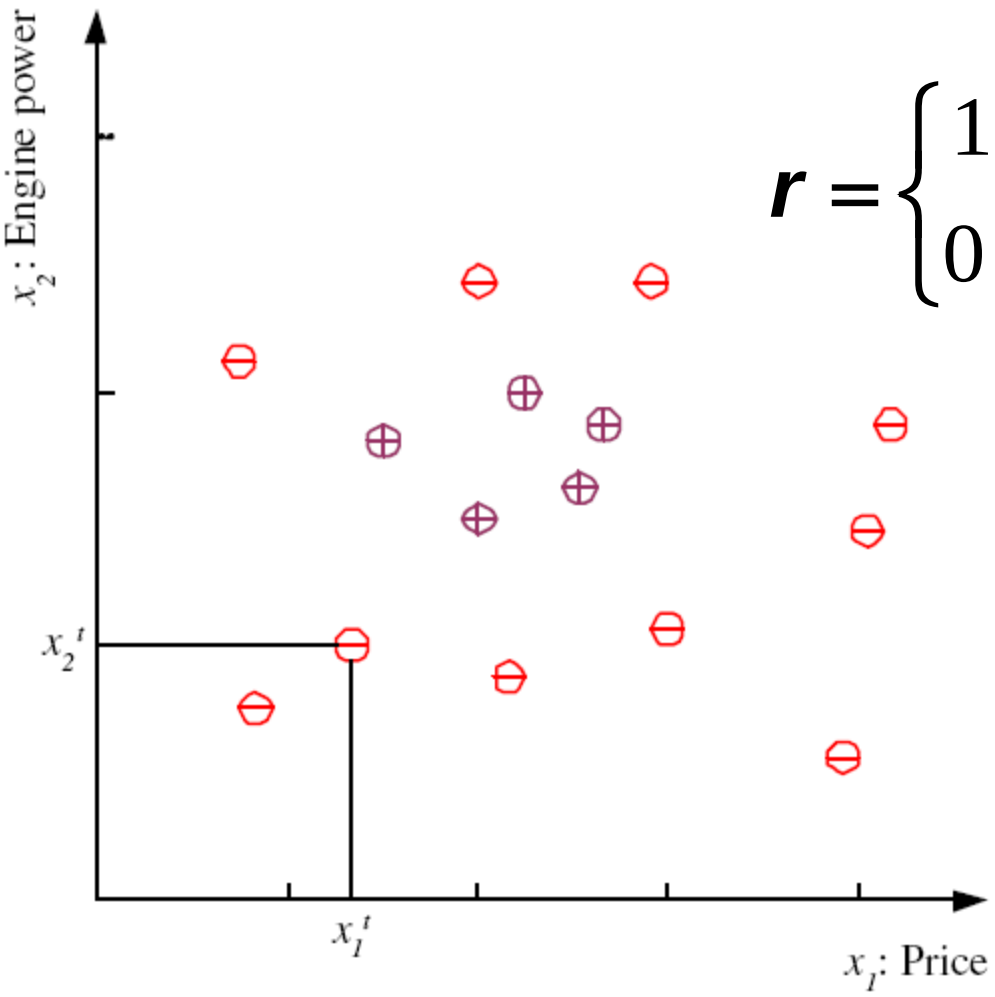


Chapter 2
Supervised Learning

Učenje koncepata (klasa,razreda) na temelju primjera

- Klasa C "obiteljski auto"
  - Predviđanje: Je li auto *x* obiteljski auto?
  - Crpljenje znanja: Što ljudi očekuju od obiteljskog automobila?
  - **Izlaz:**
    - Pozitivni (+) i negativni (-) primjeri
  - **Predstavljanje ulaza:**
    - x_1: cijena, x_2: snaga motora

Skup za učenje $\mathcal{X}$



$$\mathcal{X} = \{\mathbf{x}^t, r^t\}_{t=1}^N$$

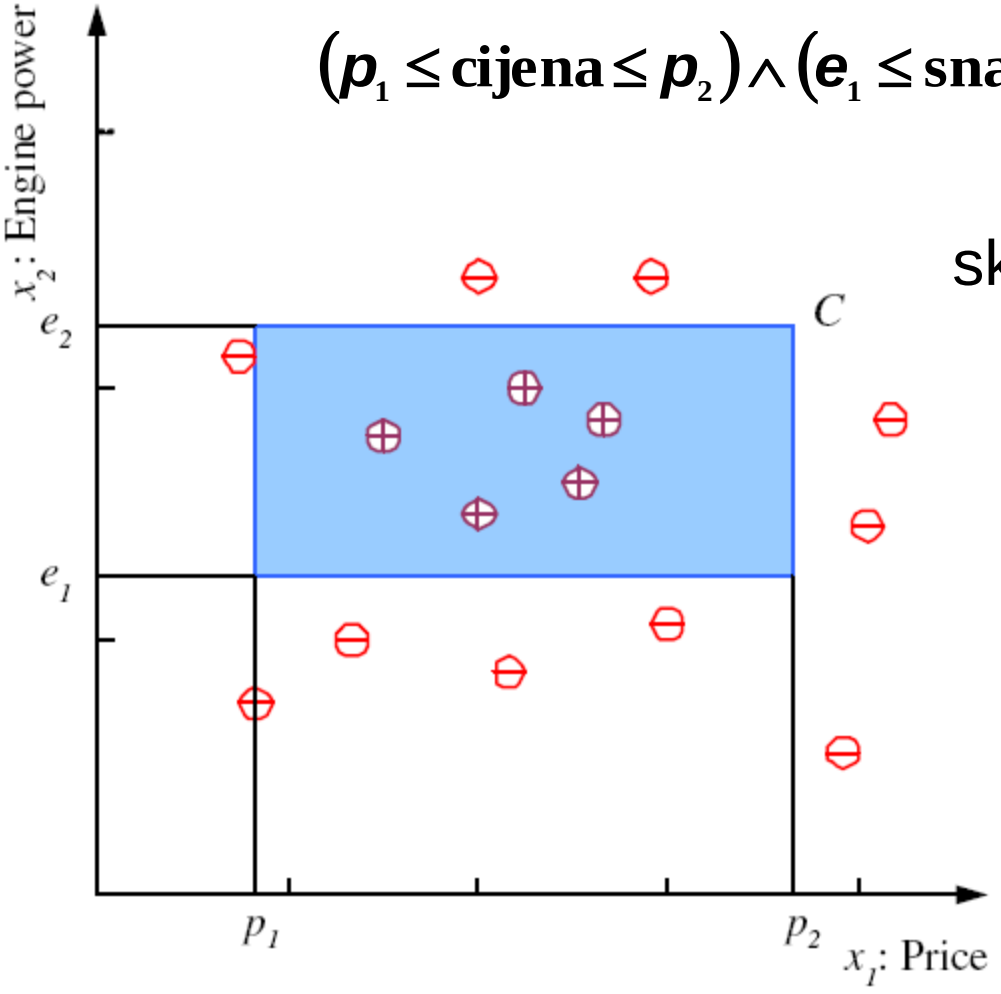
$$r = \begin{cases} 1 \text{ ako je x pozitivan} \\ 0 \text{ ako je x negativan} \end{cases}$$

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

Prostor hipoteza $\mathcal{H}$

Pretpostavka (na temelju prethodnog ispitivanja) o C:

$$(p_1 \leq \text{cijena} \leq p_2) \wedge (e_1 \leq \text{snaga motora} \leq e_2)$$

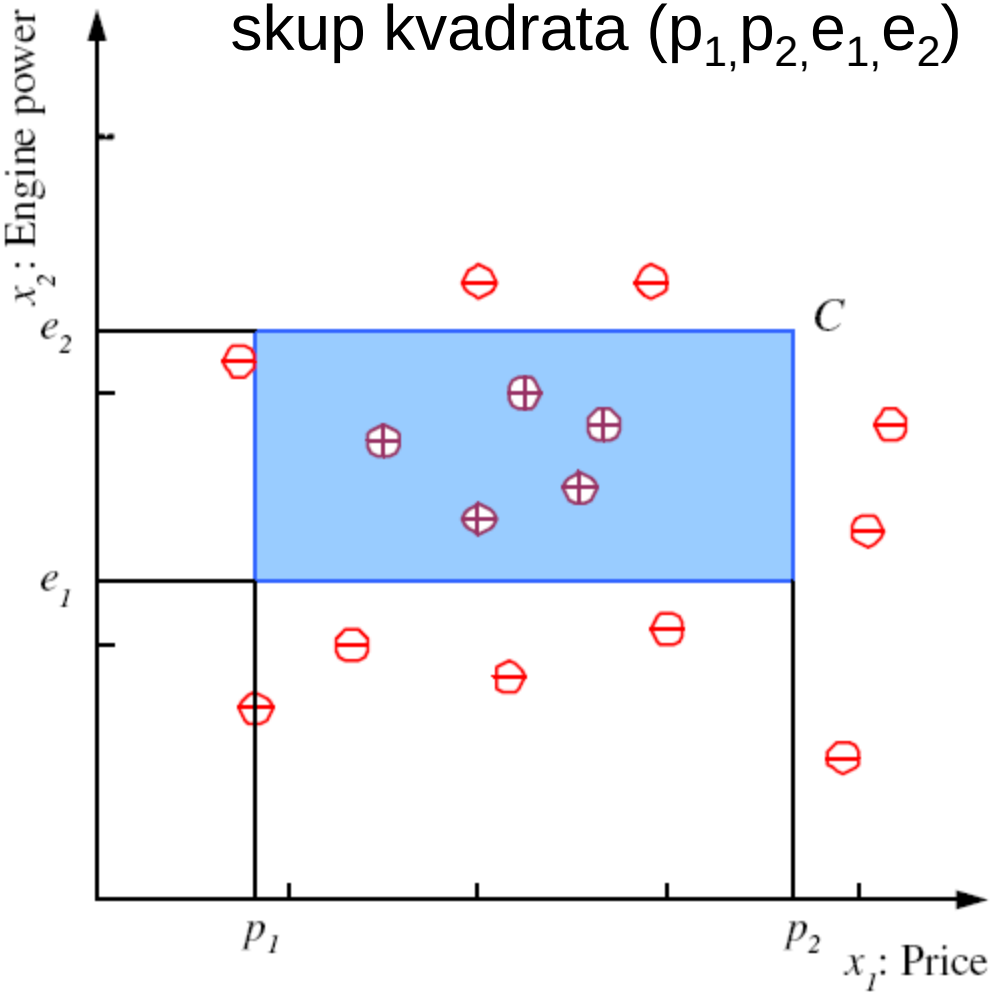


Prostor hipoteza $\mathcal{H}$ – skup kvadrata (p_1, p_2, e_1, e_2)

Algoritam učenje treba naći h iz $\mathcal{H}$ koji najbolje aproksimira C *(nalazi parametre)*

Prostor hipoteza $\mathcal{H}$

Prostor hipoteza $\mathcal{H}$ – skup kvadrata (p_1, p_2, e_1, e_2)



U stvarnosti ne možemo evaluirati kako dobro se h(x) podudara s c(x)

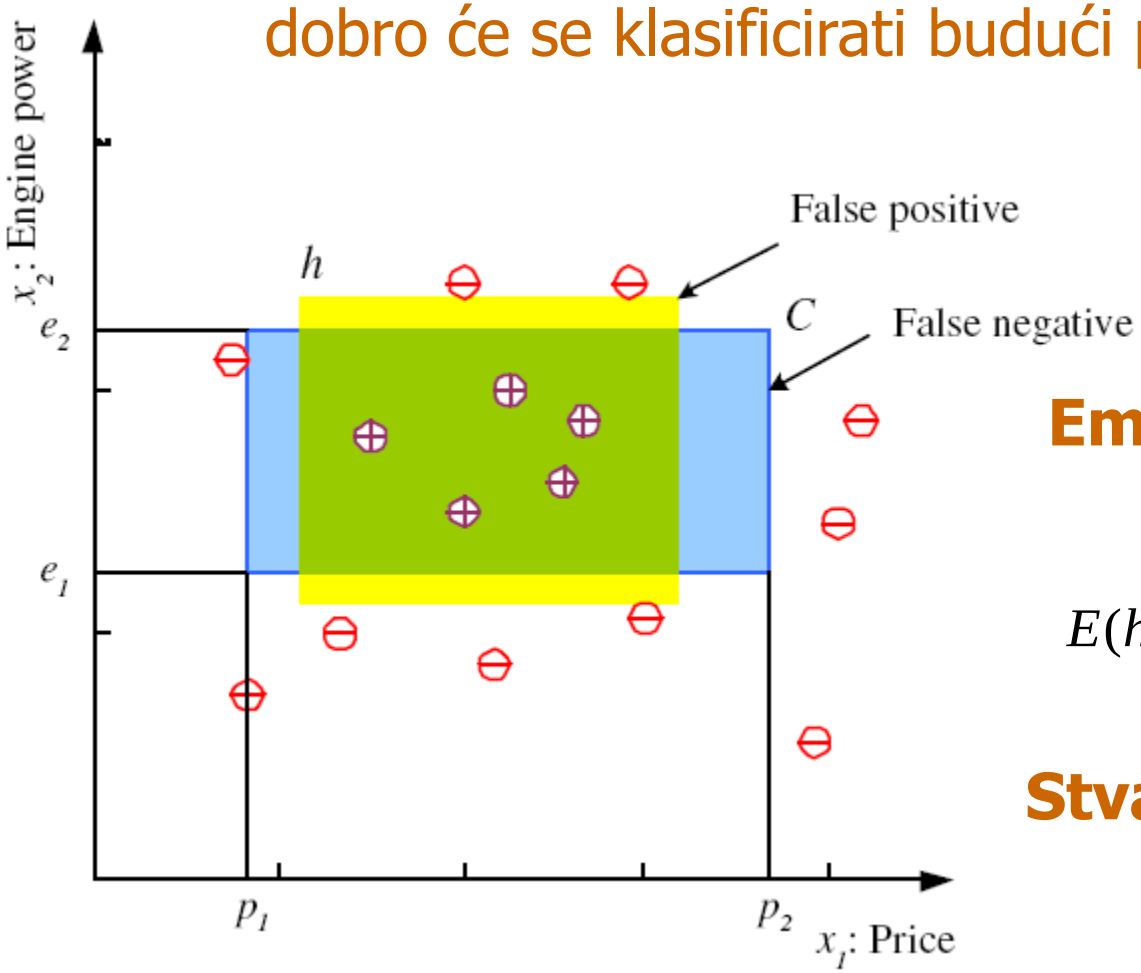
X je mali podskup

Empirijska greška *h* na *X*

$$E(h \mid \mathcal{X}) = \sum_{t=1}^{N} \delta\left(h\left(\mathbf{x}^t\right), r^t\right)$$

Prostor hipoteza $\mathcal{H}$

Problem **generalizacije** – kako dobro će se klasificirati budući primjeri



Empirijska greška

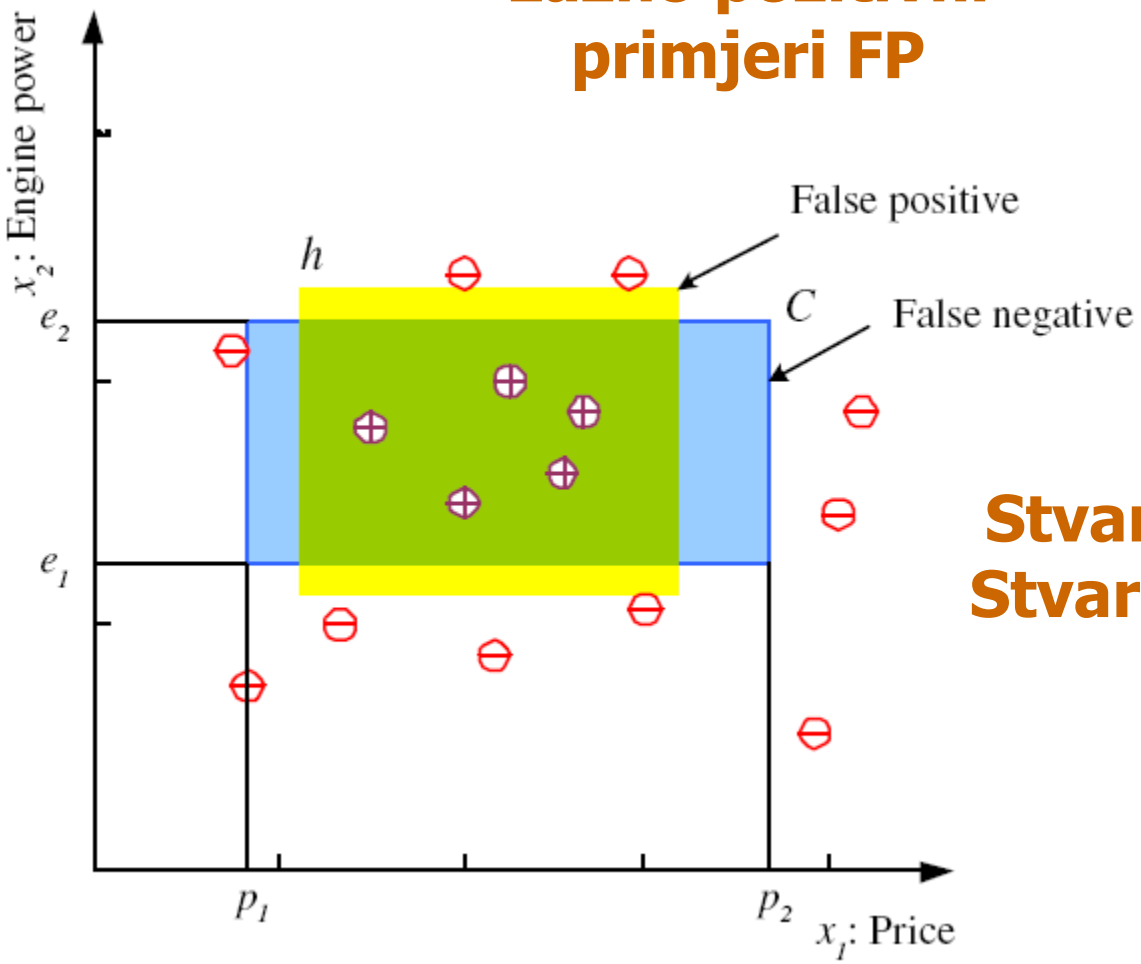
h na *X*

$$E(h \mid \mathcal{X}) = \sum_{t=1}^{N} \delta\left(h\left(\mathbf{x}^t\right), r^t\right)$$

Stvarna pogreška ?

Prostor hipoteza $\mathcal{H}$

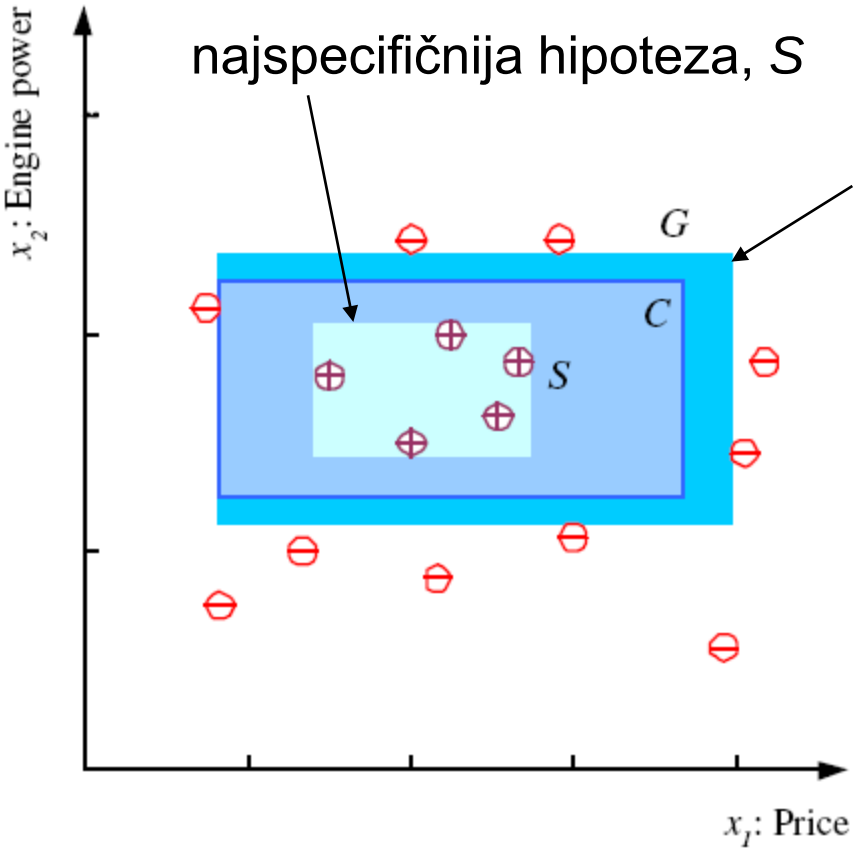
Lažno pozitivni primjeri FP



Lažno negativni primjeri FN

Stvarno pozitivni TP,
Stvarno negativni TN

S, G i prostor inačica



$h \in \mathcal{H}$, između S i G je **konzistentna**

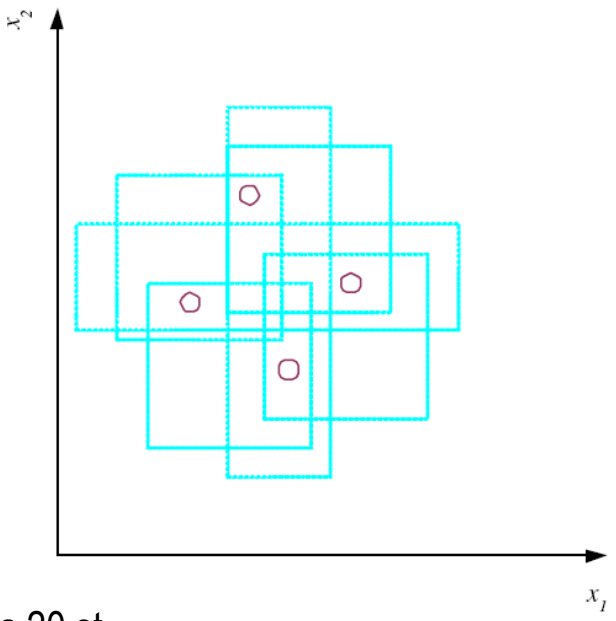
i čini
prostor inačica

(Mitchell, 1997)

Vapnik-Chervonenkisova dimenzija*

- Pretpostavili smo da je $c \in \mathcal{H}$, to znači da postoji $h \in \mathcal{H}$, $E(h|X) = 0$.

- Može se desiti da to nije slučaj
- Želimo osigurati da je $\mathcal{H}$ dovoljno fleksibilan (ili da ima kapacitet) da nauči C.
  - *N* primjera može se označiti na 2^N načina kao pozitivni ili negativni.



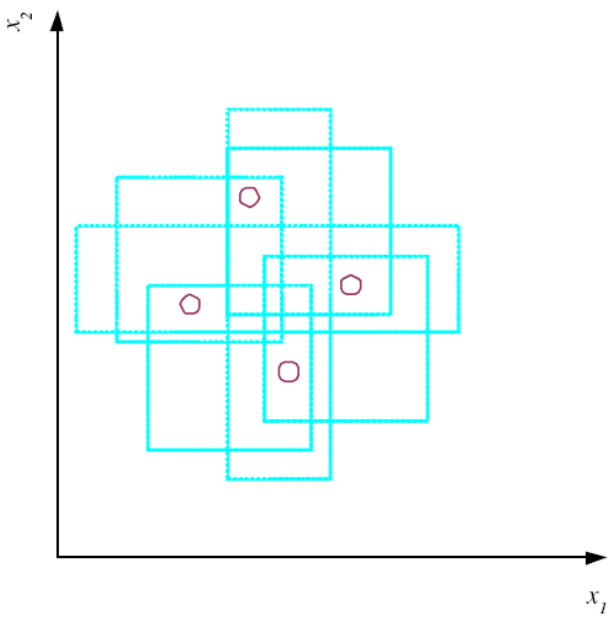
*Razvoj statističke teorije učenja započeo krajem 70tih godina 20.st.

Knjige o SLT : **Vladimir Vapnik (1995)** *The Nature of Statistical Learning Theory*, Springer, 1995,
(1995) *Statistical Learning Theory*

Vapnik-Chervonenkisova dimenzija

- Ako postoji skup od N primjera takav da za svako označavanje tih primjera postoji hipoteza $h \in \mathcal{H}$ koja je konzistentna s primjerima, onda kažemo da $\mathcal{H}$ **razdjeljuje (*shatter*)** *N* primjera.

- VC dimenzija skupa $\mathcal{H}$ je najveći N za koji $\mathcal{H}$ razdjeljuje N primjera.

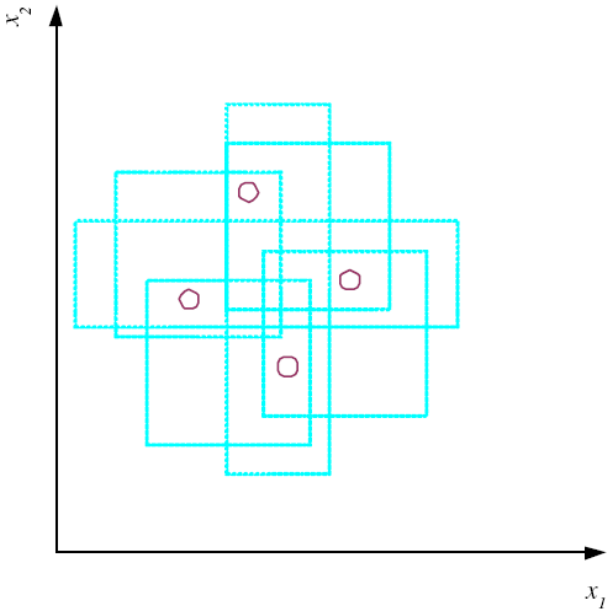


Ako je $\mathcal{H}$ skup pravokutnika poravnat po osima tada je $\mathrm{VC}(\mathcal{H}) = 4$ *jer može razdijeliti neke 4 točke za sva moguća označavanja.*

Vapnik-Chervonenkisova dimenzija

- Kod računanja VC dovoljno je da nađemo jedan skup od N točaka koje $\mathcal{H}$ **razdjeljuje** (za sva moguća označavanja!).

  Primjer: VC($\mathcal{H}$) = 4, ali ipak 4 točke na pravcu se ne mogu razdijeliti s $\mathcal{H}$.

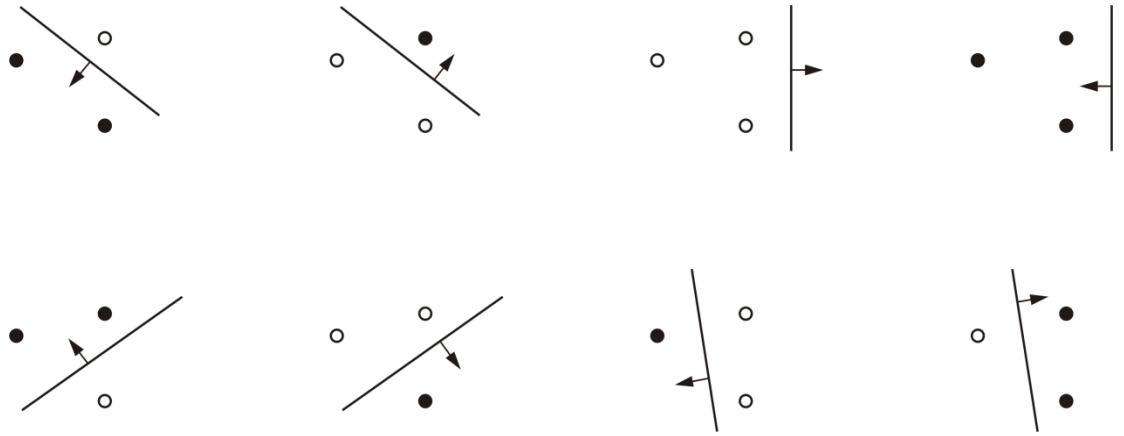


- VC dimenzija skupa je pesimistična
- VC dimenzija ne zavisi o distribuciji.

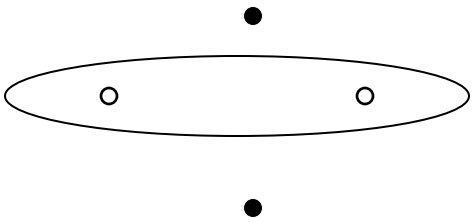
Pravokutnik poravnat po osima može razdijeliti 4 točke.

VC dimenzija – Primjer

Primjer: Skup funkcija→ pravci (hiperravnie) u R^2



VC dimenzija = 3 (2^3=8 označavanja).



Nije moguće razdijeliti 4 točke u R^2 sa pravcem.

VC dimenzija – Pitanja

Pitanja: Ako je $\mathcal{H}$

1. Skup svih kružnica u R^2
2. Skup svih trokuta R^2
3. Skup svih elipsa u R^2

Koliko je $VC(\mathcal{H})$?

Statistička teorija učenja

Statistička teorija učenja (ili VC teorija) pokazuje kako se može izbjeći prenaučenost

Važno je napraviti restrikciju klase funkcija iz koje biramo našu procjenu

Šum u podacima i složenost modela

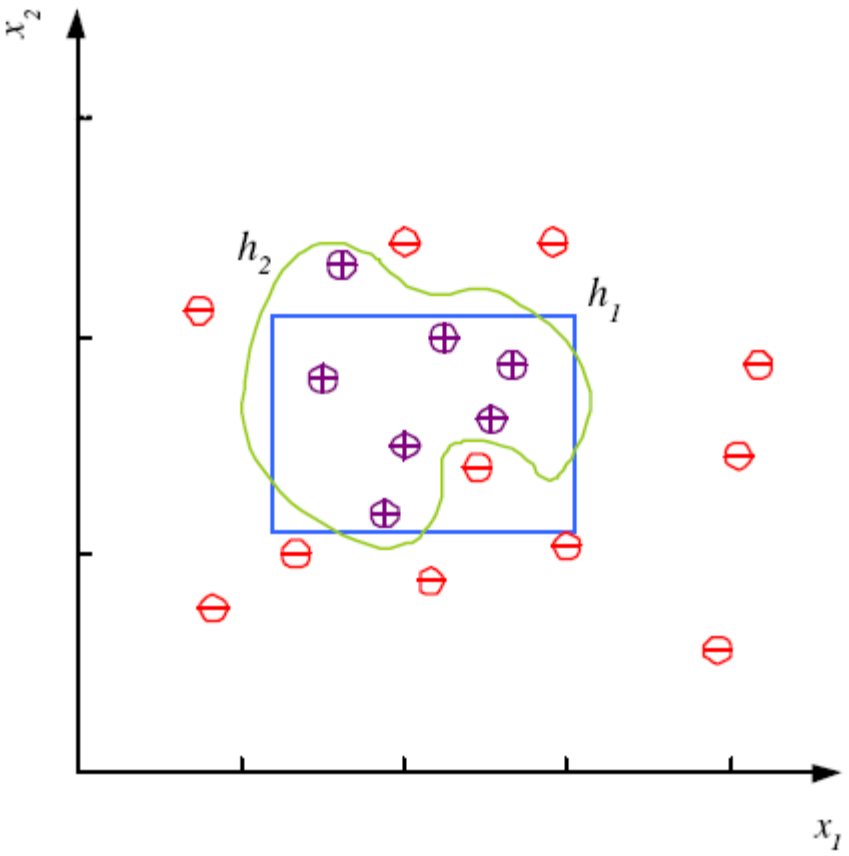
Šum - anomalije u podacima

Uzroci šuma …

Ako je šum prisutan - > nema jednostavne granice.

Opcije:

1. hipoteze većeg kapaciteta bez greške, ili
2. jednostavne hipoteze i dozvoliti (manju) grešku



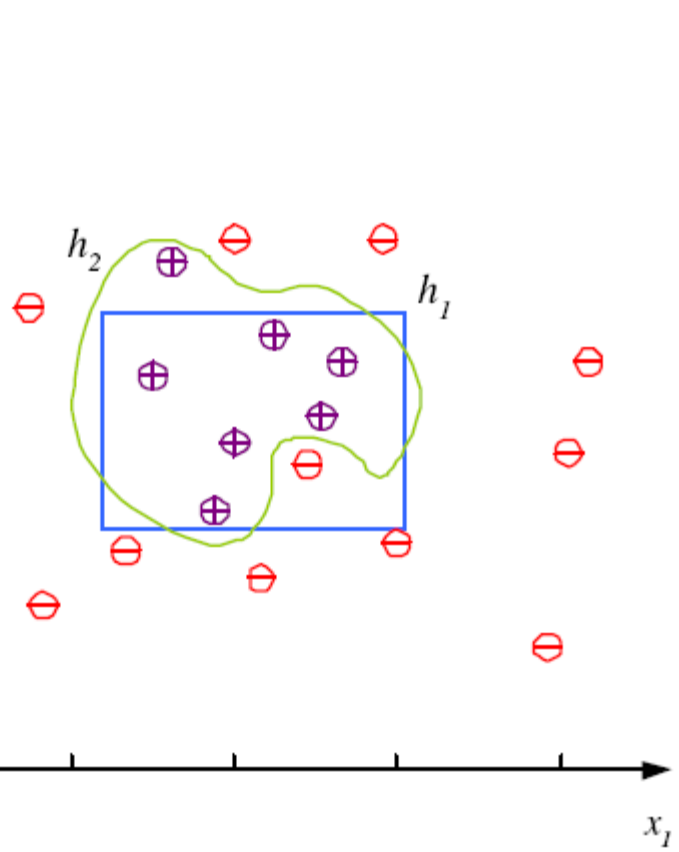
Šum u podacima i složenost modela

Koriste se jednostavniji modeli jer:

1. jednostavniji je za korištenje za buduće predviđanje (manja vremenska složenost)
2. lakše se uči (manje parametara modela manja, prostorna složenost)

Jednostavniji modeli – očekujemo manje promjene modela s manjom promjenom podataka -> **mala varijanca**

Jednostavniji model je krući i jači u pretpostavkama i može ne pronaći pravu hipotezu -> **velika pristranost**

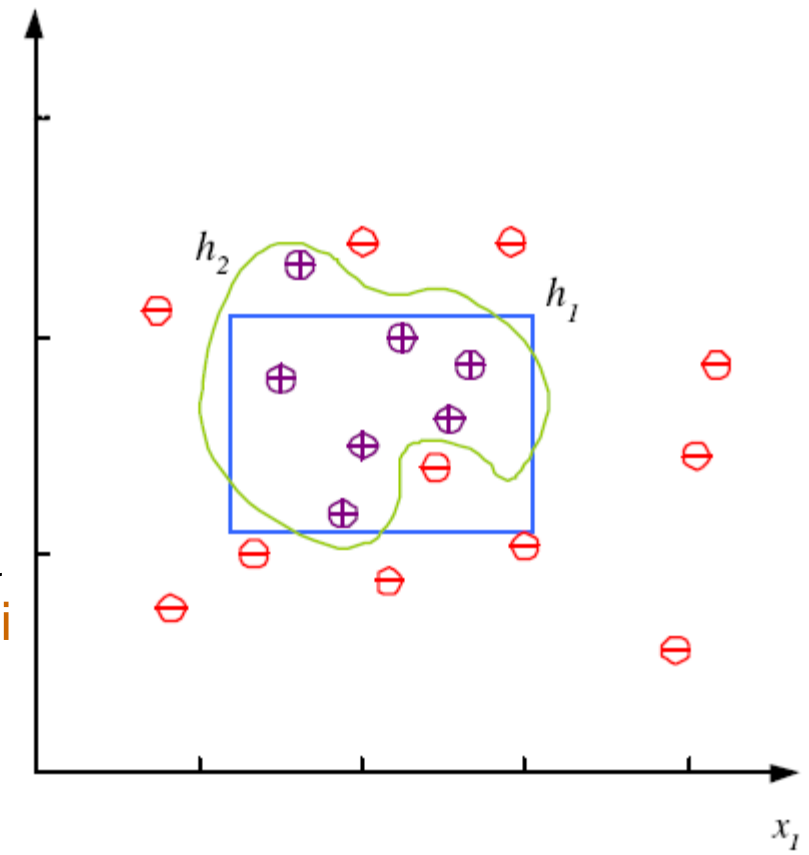


Šum u podacima i složenost modela

Koristi se jednostavniji model jer:

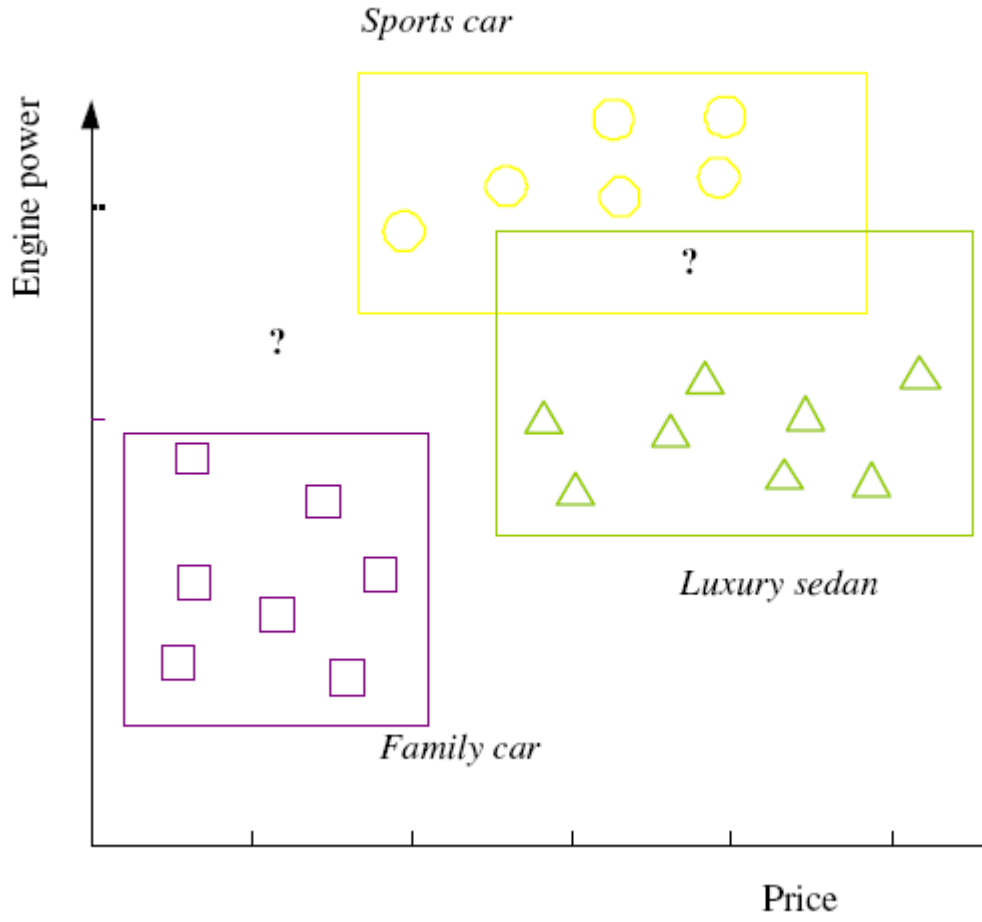
3. lakše se objašnjava (interpretabilniji je) – ekstakcija informacija
4. bolje generalizira
   Ako ima šuma jednostavniji model je manje podložan promjenama u podacima (manja varijanca) i bolje će generalizirati – čak i ako radi pogrešku na skupu za učenje

   Occamov princip)



Klasifikacija u više razreda, C_i i=1,...,K

$$\mathcal{X} = \{\boldsymbol{x}^t, r^t\}_{t=1}^N$$ r je k.dim vektor



$$r_i^t = \begin{cases} 1 \text{ if } \boldsymbol{x}^t \in C_i \\ 0 \text{ if } \boldsymbol{x}^t \in C_j, j \neq i \end{cases}$$

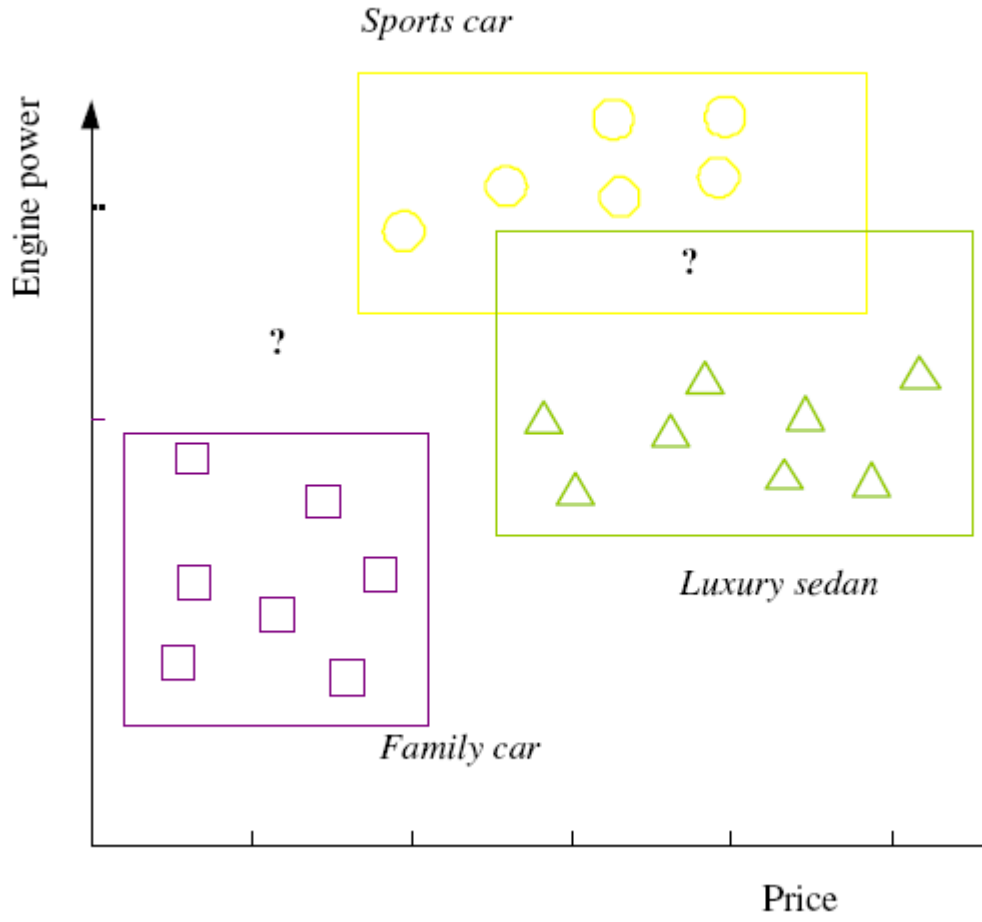
Klasifikacija u k klasa ->
k binarnih klasifikacija

Hipoteze za treniranje
$h_i(\boldsymbol{x})$, $i = 1,...,K$:

$$h_i\left(\boldsymbol{x}^t\right) = \begin{cases} 1 \text{ if } \boldsymbol{x}^t \in C_i \\ 0 \text{ if } \boldsymbol{x}^t \in C_j, j \neq i \end{cases}$$

Klasifikacija u više razreda, C_i i=1,...,K

$$\mathcal{X} = \{\boldsymbol{x}^t, r^t\}_{t=1}^N$$ r je k.dim vektor



$$r_i^t = \begin{cases} 1 \text{ if } \boldsymbol{x}^t \in C_i \\ 0 \text{ if } \boldsymbol{x}^t \in C_j, j \neq i \end{cases}$$

Klasifikacija u k klasa ->
k binarnih klasifikacija

Hipoteze za treniranje
$h_i(\boldsymbol{x}),\ i = 1,...,K$:

$$h_i\left(\boldsymbol{x}^t\right) = \begin{cases} 1 \text{ if } \boldsymbol{x}^t \in C_i \\ 0 \text{ if } \boldsymbol{x}^t \in C_j, j \neq i \end{cases}$$

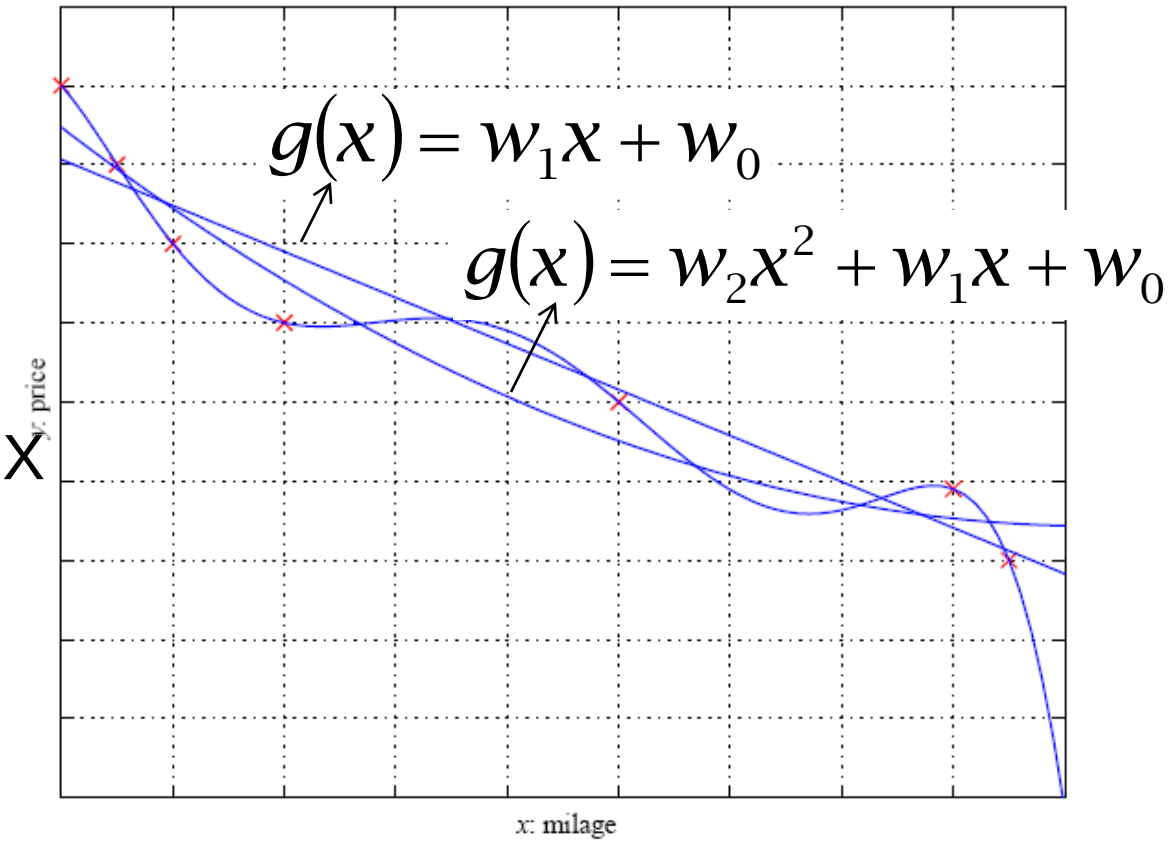
Regresija

$$\mathcal{X} = \left\{ x^t, r^t \right\}_{t=1}^{N}$$

$$r^t \in \Re$$

Ako nema šuma
-> interpolacija

Ako je x izvan ranga X
-> ekstrapolacija



Regresija:

$$r^t = f\left(x^t\right) + \varepsilon$$

nepoznata funkcija

slučajna pogreška

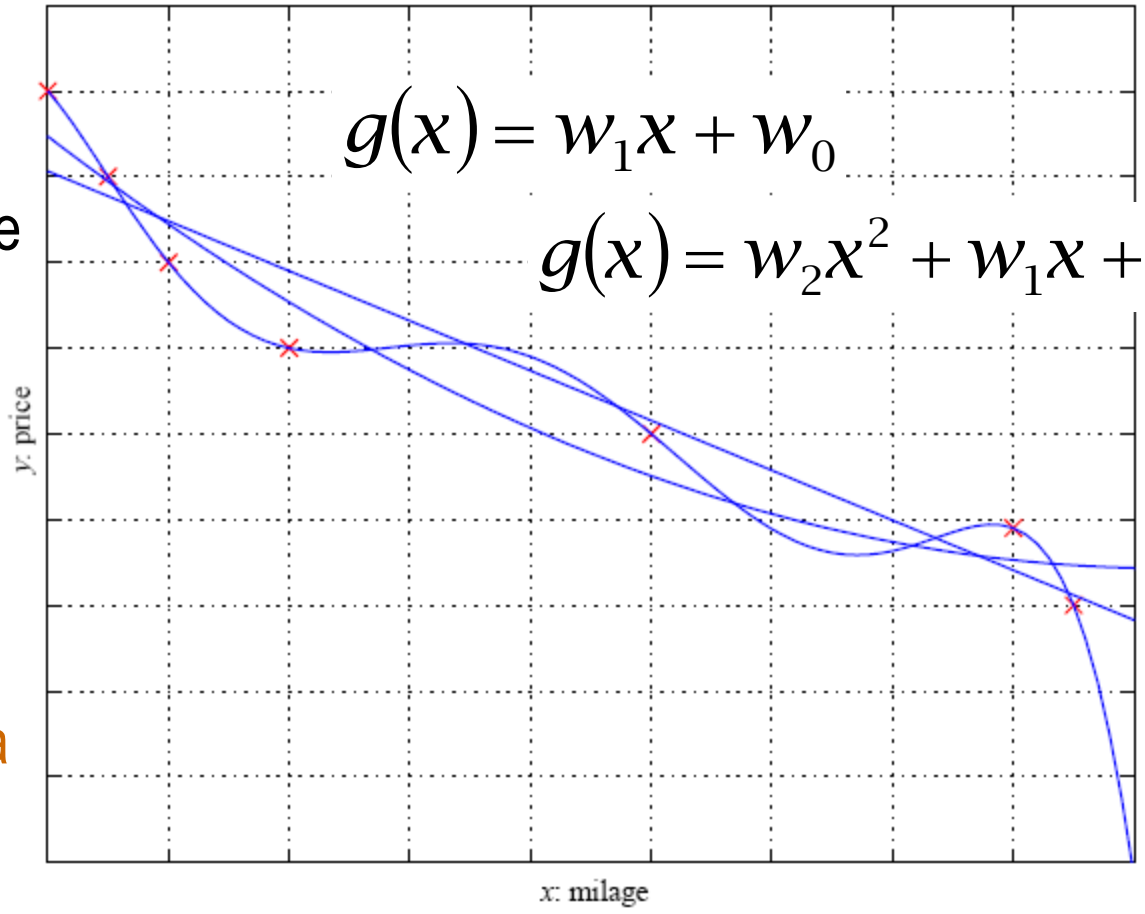
latentne
(skrivene) varijable

Regresija

Analitičo rješenje za w0 i w1 kroz parcijalne derivacije

Stupanj polinoma raste - empirijska greška pada,

ali polinomi višeg reda ne slijede generalni trend u podacima



Mala empirijska pogreška ne garantira malu očekivanu pogrešku!

Odabir modela i generalizacija

- Učenje je **loše postavljen problem** (engl. ill-posed problem); podatci nisu dovoljni za jedinstveno rješenje
  - Postoji potreba za **induktivnom pristranošću**, pretpostavkama o *H*
- Generalizacija:
  - Koliko dobro model radi na novim podatcima
- Za dobru generalizaciju kompleksnost hipoteza treba se poklapati s kompleksnošću podataka
- Podnaučenost (underfitting):
  - *H* je manje kompleksan od *C* ili *f*
- Prenaučenost (overfitting):
  - *H* je kompleksniji od *C* ili *f*

Trostruki balans, pogleška generalizacije

- Tri su faktora u balansu (Dietterich, 2003):
  1. Kompleksnost, tj. kapacitet prostora hipoteza $\mathcal{H}$
  2. Veličina skupa za učenje, *N*,
  3. Pogreška generalizacije na novim podatcima, *E*

- Kako *N* ↑, *E* ↓
- As *VC*($\mathcal{H}$) ↑, prvo*E*↑ onda *E* ↓

- Pogreška generalizacije* kompleksnih sustava smanjuje se povećanjem podataka, ali samo do neke točke

**generalisation error, expected (test) error, expected risk*

Unakrsna validacija

- Kako bismo procijenili pogrešku generalizacije, trebamo podatke neviđene za vrijeme treniranja
- Npr. podijelit ćemo podatke:
  - Skup za učenje(50%)
  - Skup za validaciju (25%)
  - Skup za testiranje (objavu) (25%) – očekivana pogreška
- Unakrsna validacija (*crossvalidation*) – tehnika ponovnog uzorkovanja – za selekciju modela
- Podaci se mogu uzrokovati više puta ukoliko je skup malen

Dimenzije nadziranog učenja - sažetak

$$\mathcal{X} = \left\{ x^t, r^t \right\}_{t=1}^{N}$$

Uzorak – poredak nevažan, svi elementi su izvučeni iz iste združene distribucije - nezavisni i identično distribuirani

(*independent and indentically distributed - **iid***)

1. Model: $g(\boldsymbol{x} \mid \theta)$
2. Funkcija gubitka(loss function)*: $E(\theta \mid \mathrm{X}) = \sum_t L\left(r^t, g\left(\mathbf{x}^t \mid \theta\right)\right)$

   (*loss function,cost fu*nction)
3. Optimizacijska procedura: $\theta^* = \arg\min_{\theta} E(\theta \mid \mathcal{X})$

*općenitije od empirijske pogreške

Dimenzije nadziranog učenja

Uvjeti:

1. Klasa hipoteza treba imati dovoljan kapacitet da može izraziti funkciju koja je generirala podatke
2. Treba biti dovoljno podataka za učenje da bi se oblikovala dovoljno dobra hipoteza
3. Dobar optimizacijski algoritam koji pronalazi hipotezu