11. Neparametarske metode

Strojno učenje 1, UNIZG FER, ak. god. 2021./2022.

Jan Šnajder, natuknice s predavanja, v1.3

1 Parametarske vs. neparametarske metode

- **Parametarske metode** – hipoteza je definirana do na parametre $\boldsymbol{\theta}$
  - broj parametara modela n (složenost modela) ne ovisi o broju primjera N
  - pretpostavljaju da se podatci ravnaju po nekom modelu (distribuciji)
  - primjeri imaju **globalan** utjecaj na izgled hipoteze
- **Neparametarske metode** – hipoteza nije eksplicitno definirana
  - broj parametara ovisi o broju primjera
  - ne pretpostavljaju model (distribuciju) podataka
  - **lokalna** aproksimacija hipoteze u okolici pohranjenih primjera
- NB: Neparametarski modeli imaju parametre (ali nemaju parametre distribucije)!
- Predikcija se ne radi unaprijed nego na zahtjev ⇒ **lijene metode** (*lazy methods*)
- **Induktivna pristranost** neparametarskih metoda: slični primjeri imaju slične oznake
- Preporuke:
  - malo podataka i/ili poznat model/distribucija ⇒ parametarski postupci
  - mnogo podataka i nepoznat model/distribucija ⇒ neparametarski postupci

2 SVM

- SVM model:

$$h(\mathbf{x}) = \underbrace{\mathbf{w}^\intercal \mathbf{x} + w_0}_{\text{Primarno}} = \underbrace{\sum_{i=1}^{N} \alpha_i y^{(i)} \mathbf{x}^\intercal \mathbf{x}^{(i)} + w_0}_{\text{Dualno}}$$

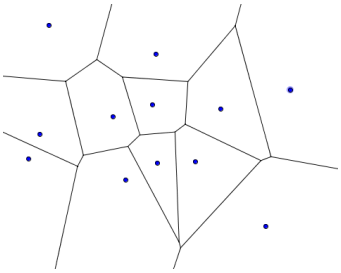
- Primarna formulacija ⇒ parametarski; dualna formulacija ⇒ neparametarski
- Broj parametara proporcionalan broju potpornih vektora, koji ovisi o N
- Prikladno kada $N \ll n$ (algoritam SMO ima složenost $\mathcal{O}(N^2)$)

3 Algoritam k-NN

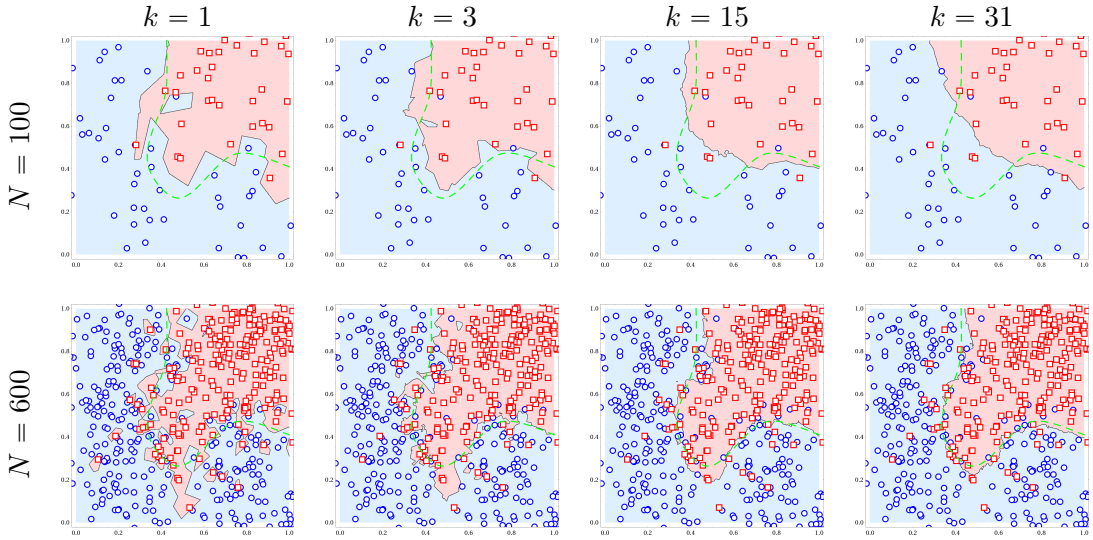
- Neparametarski klasifikacijski algoritam
- Predikcija na temelju većinske oznake k **najbližih susjeda** (*nearest neighbors*):

$$h(\mathbf{x}) = \underset{j\in\{0,\ldots,K-1\}}{\operatorname{argmax}} \sum_{(\mathbf{x}^{(i)},y^{(i)})\in \mathrm{NN}_k(\mathbf{x})} \mathbf{1}\{y^{(i)} = j\}$$

- k je **hiperparametar** algoritma $\Rightarrow$ manji k daje složeniji model
- $k = 1 \Rightarrow$ ulazni prostor particioniran u **Voronoijev dijagram**:



- Primjer: binarna klasifikacija u $n = 2$ u ovisnosti o k za $N = 100$ i $N = 600$:

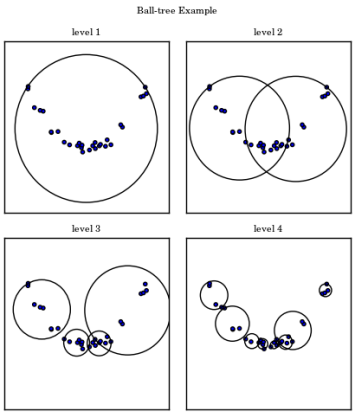


- **Težinski** k**-NN** – utjecaj primjera ovisi o udaljenosti/sličnosti $\Rightarrow$ **kernel**:

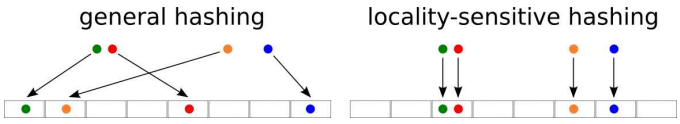
$$h(\mathbf{x}) = \underset{j\in\{0,\ldots,K-1\}}{\operatorname{argmax}} \sum_{i=1}^{N} \kappa(\mathbf{x}^{(i)},\mathbf{x})\mathbf{1}\{y^{(i)} = j\}$$

- Mjera udaljenosti ne mora biti euklidska (npr. Mahalonbisova udaljenost)
- Računalni problem: **nalaženje najbližeg susjeda** (*nearest neighbor search*)
- Alternative iscrpnom pretraživanju (bitno za velike skupove podataka):

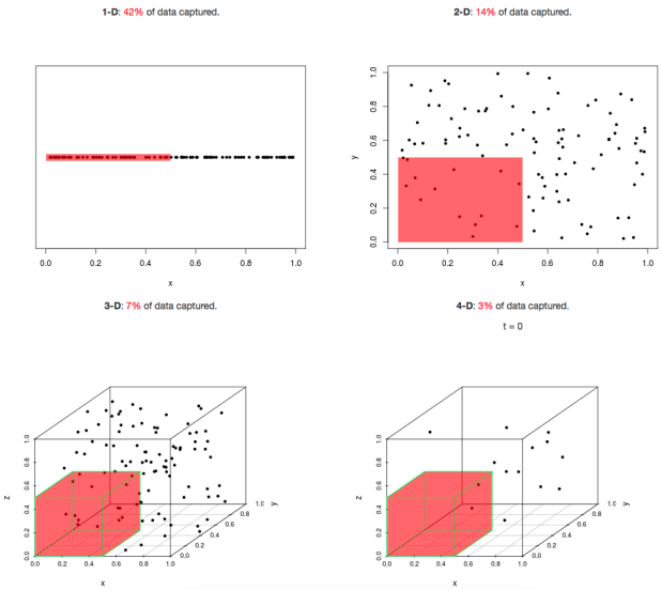
– egzaktne metode: indeksiranje prostora primjera (npr. **ball tree**)



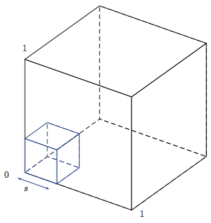
– aproksimativne metode: **locally sensitive hashing (LSH)**

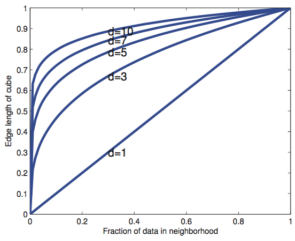


- **Prokletstvo dimenzionalnosti** (*curse of dimensionality*):
  - s porastom dimenzije n sve točke postaju međusobno vrlo udaljene
  - udaljenosti postaju nediskriminative
  - općenit problem svih algoritama u visokodimenzijskim prostorima
- Primjer: s porastom broja dimenzija udio podataka u jediničnoj hiperkocki opada:



- Primjer: s porastom broja dimenzija, udaljenost između susjeda raste:





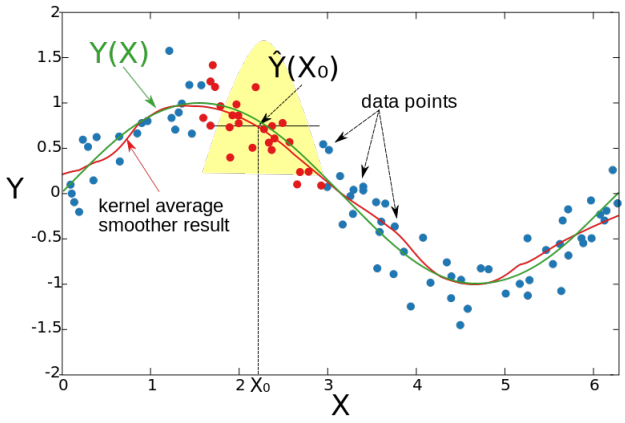
4 Neparametarska regresija

- Neparametarska regresija = **modeli zaglađivanja** (*smoothing models*)
- k-**nn smoother** - prosjek vrijednosti k najbližih susjeda:

$$h(\mathbf{x}) = \frac{1}{k} \sum_{(\mathbf{x}^{(i)}, y^{(i)}) \in \mathrm{NN}_k(\mathbf{x})} y^{(i)}$$

- **Jezgreno zaglađivanje** (*kernel smoothing*):

$$h(\mathbf{x}) = \frac{\sum_{i=1}^{N} \kappa(\mathbf{x}^{(i)}, \mathbf{x}) y^{(i)}}{\sum_{i=1}^{N} \kappa(\mathbf{x}^{(i)}, \mathbf{x})}$$



5 Stabla odluke

- Neparametarski model jer broj parametara ($\propto$ broj razina) raste s brojem primjera
- Ulazni prostor rekurzivno dijeli na lokalna područja (dva potprostora)