Obučavanje ansmbla klasifikatora

Milan M.Milosavljević

Uvod

- Teorema o besplatnom ručku (No Free Lunch Theorem): Ne postoji algoritam obučavanja koji je u svim situacijama najbolji
- Generisati skup osnovnih ML sistema (base-learners – BL) koji nakon odgovarajućeg kombinovanja daju veću tačnost od pojedinačnih
- Različiti BL koriste različite
  - Algoritme
  - Hiperparametere
  - Reprezentaciju /Modalitete/Poglede
  - Obučavajuće skupove
  - Podprobleme
- Diverzitet vs tačnost

Glasanje

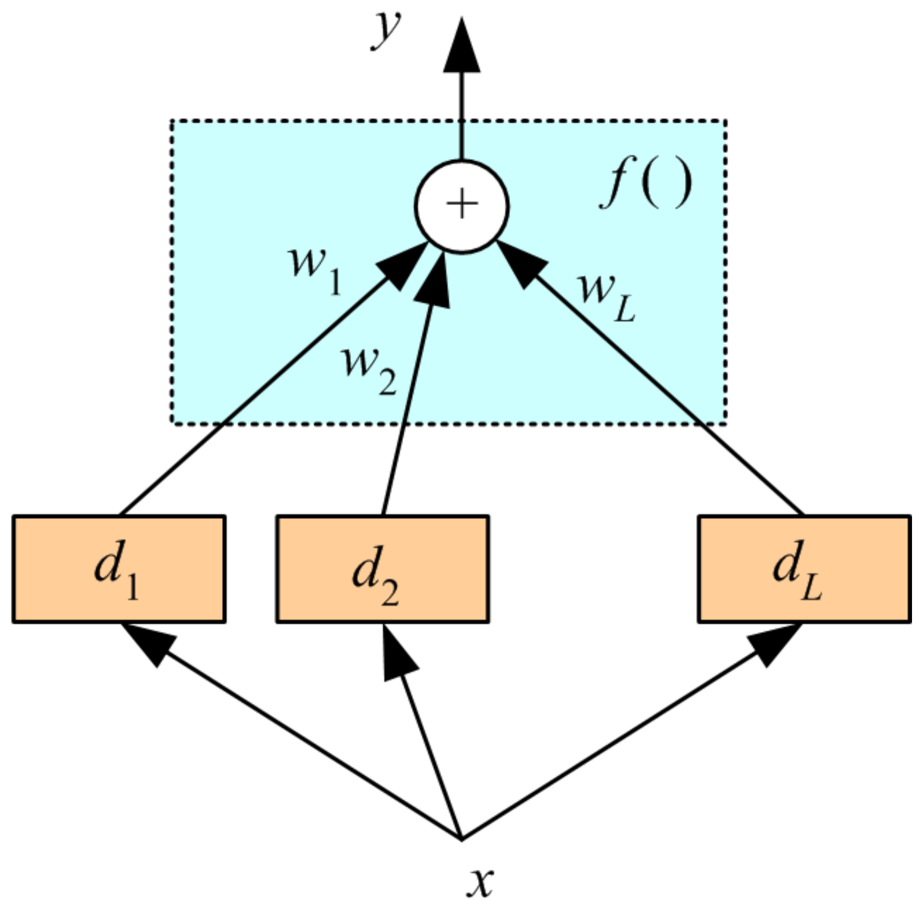
- Linearno kombinovanje

$$y = \sum_{j=1}^{L} w_j d_j$$

$$w_j \geq 0 \text{ and } \sum_{j=1}^{L} w_j = 1$$

- Klasifikacija

$$y_i = \sum_{j=1}^{L} w_j d_{ji}$$



- Bajesovska interpretacija:

$$P(C_i \mid x) = \sum_{\text{all models } \mathcal{M}_j} P(C_i \mid x, \mathcal{M}_j) P(\mathcal{M}_j)$$

- Ako je d_j iid

$$E[y] = E\left[\sum_j \frac{1}{L} d_j\right] = \frac{1}{L} L \cdot E[d_j] = E[d_j]$$

$$\mathrm{Var}(y) = \mathrm{Var}\left(\sum_j \frac{1}{L} d_j\right) = \frac{1}{L^2} \mathrm{Var}\left(\sum_j d_j\right) = \frac{1}{L^2} L \cdot \mathrm{Var}(d_j) = \frac{1}{L} \mathrm{Var}(d_j)$$

Bajas (pomeraj) se ne menje, a varijansa opada sa L.

- Ako su zavisni, greška se povećava sa pozitivnom korelacijom

$$\mathrm{Var}(y) = \frac{1}{L^2} \mathrm{Var}\left(\sum_j d_j\right) = \frac{1}{L^2}\left[\sum_j \mathrm{Var}(d_j) + 2 \sum_j \sum_{i<j} Cov(d_i, d_j)\right]$$

Fiksirana pravila kombinovanja

Rule	Fusion function $f(\cdot)$
Sum	$y_i = \frac{1}{L}\sum_{j=1}^{L} d_{ji}$
Weighted sum	$y_i = \sum_j w_j d_{ji}, w_j \geq 0, \sum_j w_j = 1$
Median	$y_i = \text{median}_j d_{ji}$
Minimum	$y_i = \min_j d_{ji}$
Maximum	$y_i = \max_j d_{ji}$
Product	$y_i = \prod_j d_{ji}$

	C_1	C_2	C_3
d_1	0.2	0.5	0.3
d_2	0.0	0.6	0.4
d_3	0.4	0.4	0.2
Sum	0.2	**0.5**	0.3
Median	0.2	**0.5**	0.4
Minimum	0.0	**0.4**	0.2
Maximum	0.4	**0.6**	0.4
Product	0.0	**0.12**	0.032

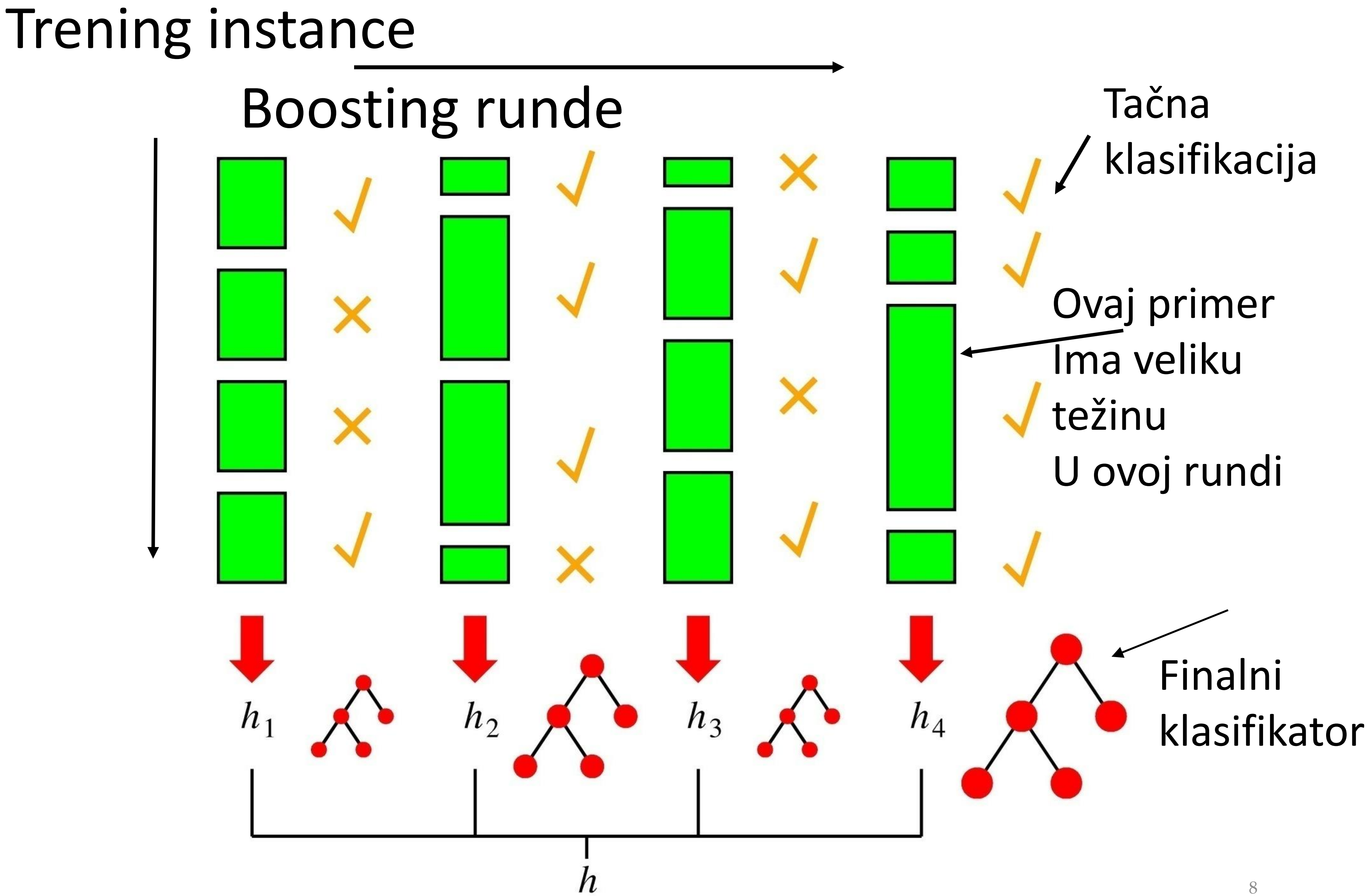
Bagging (Bootstrap aggregating)

- Pomoću butstrepovanja generisati *L* obučavajućih skupova, a zatim obučiti po jedan bazni klasifikator na svakom od njih (Breiman, 1996)
- Objediniti odluke glasanjem (Srednja vrednost ili medijana sa regresijom)
- Nestabilni algoritmi se po pravilu poboljšavaju baging-om.

Boosting

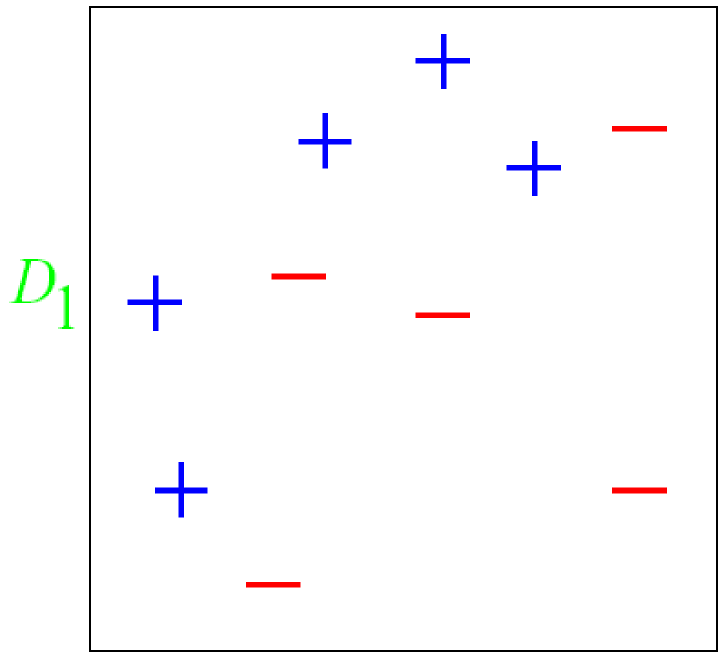
- Obučavanje klasifikatora (npr. Stabala odlučivanja) u nizu.
- Svaki naredni se fokusira na one primere koji su pogrešno klasifikovani od strane prethodnog klasifikatora u sekvenci.
- Finalna klasifikacija se dobija na osnovu glasanja svih klasifikatora u datoj sekvenci, slično kao kod bagging-a.
- Svaki pojedinačni bazni klasifikator je "slab" ali je zato ansambl "jak"
- AdaBoost je jedan specifičan algoritam iz ove klase.

Boosting - ilustracija



AdaBoost primer

Taken from "**A Tutorial on Boosting**" **by** Yoav Freund and Rob Schapire

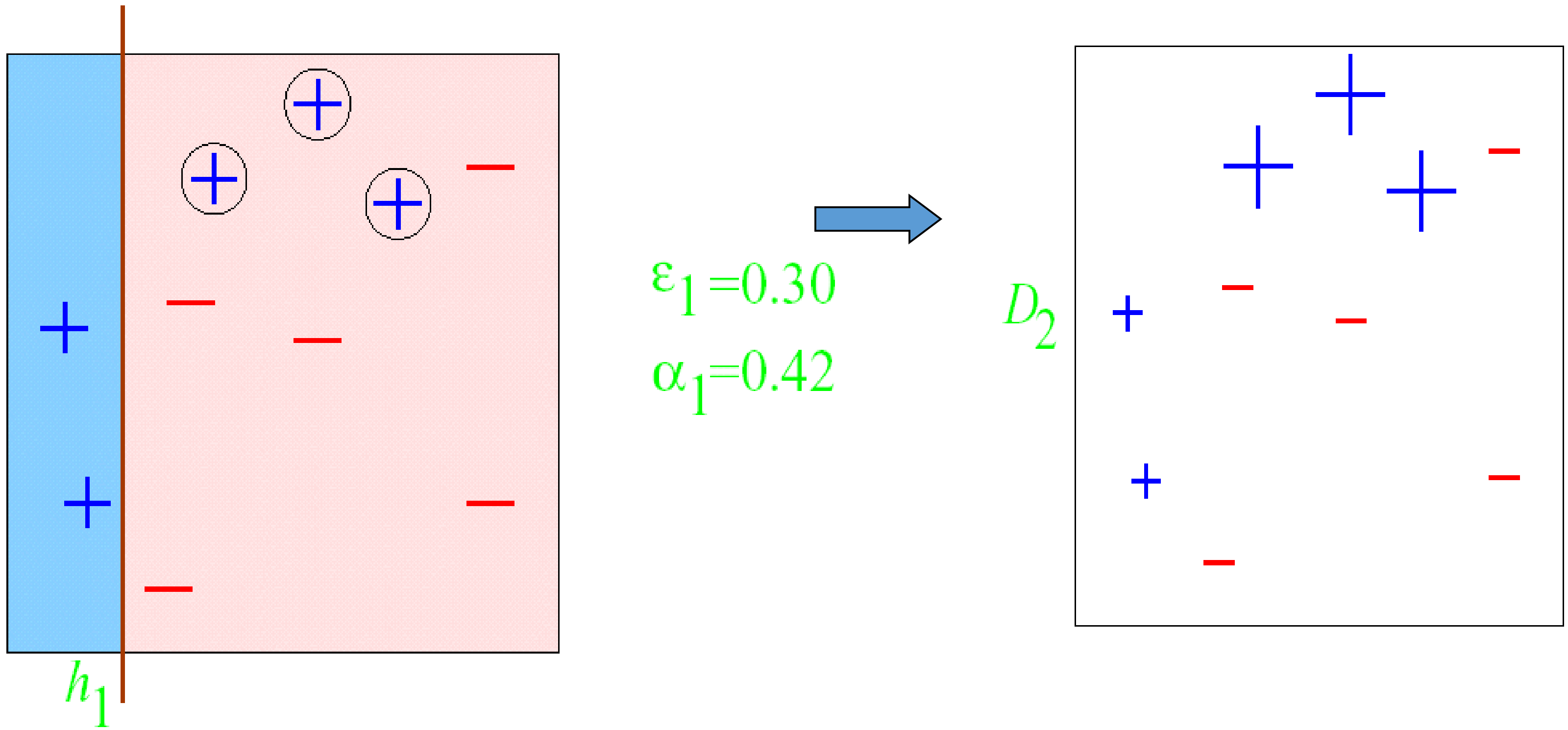


Početni trening skup: sve instance imaju jednake težine

AdaBoost Primer

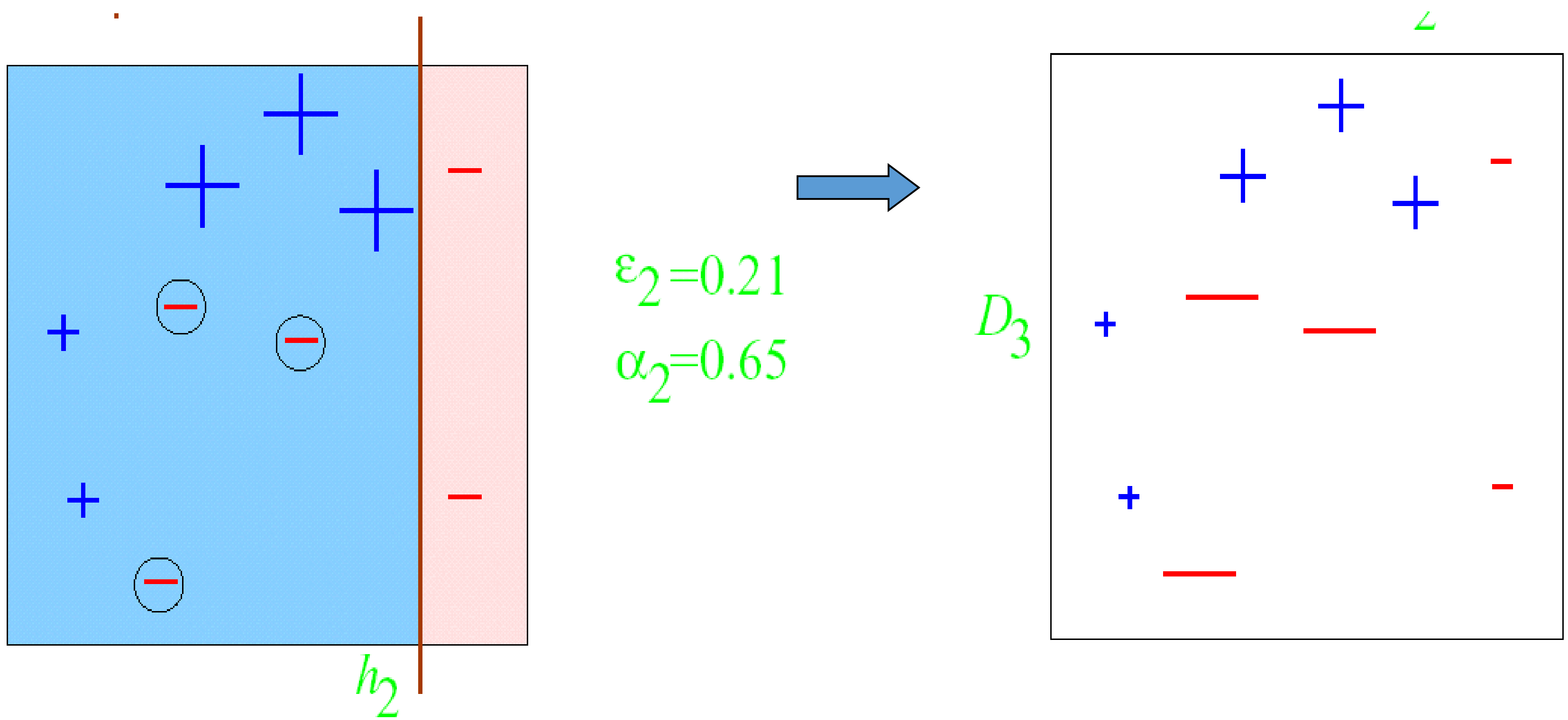
ε = greška klasifikatora
α = težinski faktor klasifikatora

Runda 1



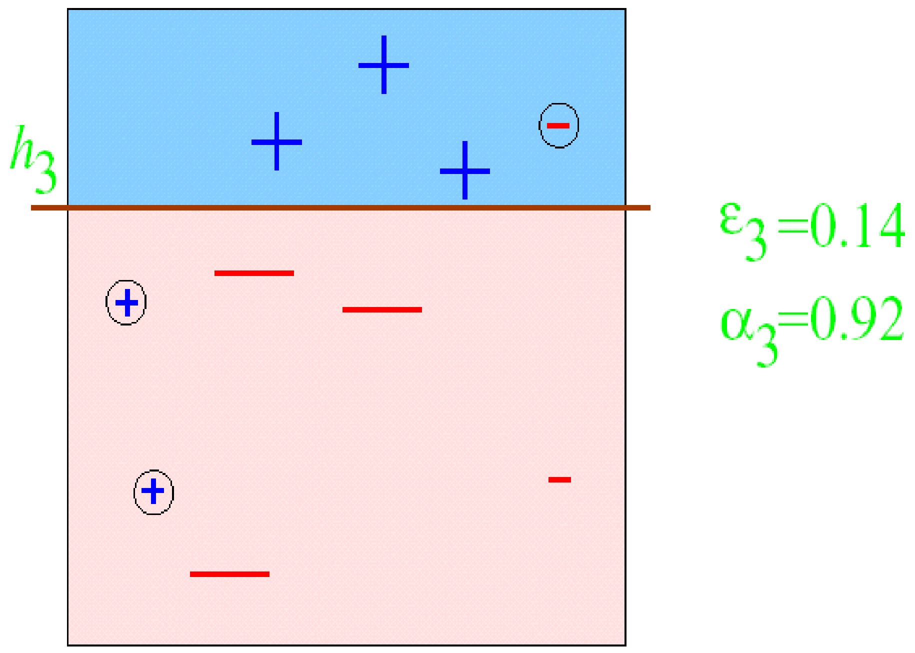
AdaBoost Primer

Runda 2

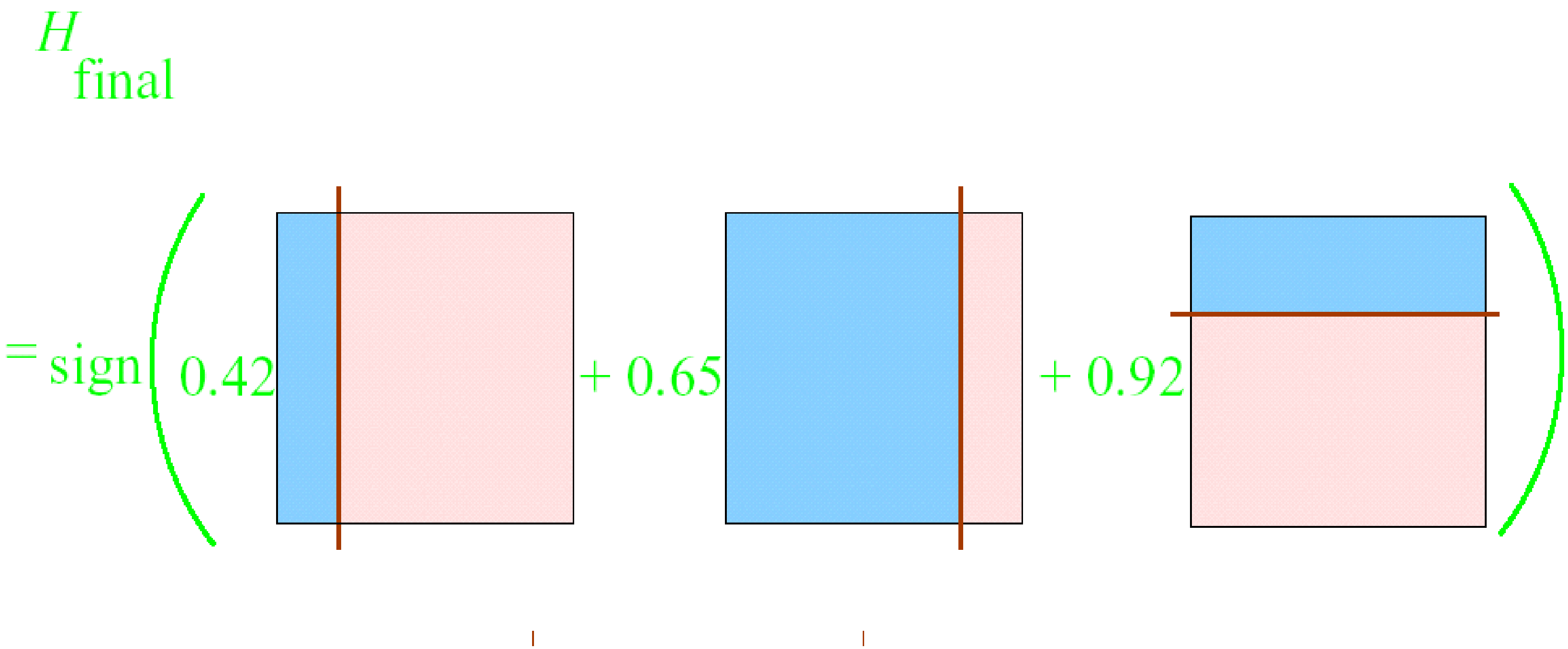


AdaBoost Primer

Runda 3



AdaBoost Primer



AdaBoost

Generiše sekvencu baznih klasifikatora od kojih se svaki fokusira na greške prethodnog

(Freund and Schapire, 1996)

Training:

For all $\{x^t, r^t\}_{t=1}^N \in \mathcal{X}$, initialize $p_1^t = 1/N$

For all base-learners $j = 1, \ldots, L$

Randomly draw $\mathcal{X}_j$ from $\mathcal{X}$ with probabilities p_j^t

Train d_j using $\mathcal{X}_j$

For each (x^t, r^t), calculate $y_j^t \leftarrow d_j(x^t)$

Calculate error rate: $\epsilon_j \leftarrow \sum_t p_j^t \cdot 1(y_j^t \neq r^t)$

If $\epsilon_j > 1/2$, then $L \leftarrow j - 1$; stop

$\beta_j \leftarrow \epsilon_j / (1 - \epsilon_j)$

For each (x^t, r^t), decrease probabilities if correct:

If $y_j^t = r^t$ $p_{j+1}^t \leftarrow \beta_j p_j^t$ Else $p_{j+1}^t \leftarrow p_j^t$

Normalize probabilities:

$Z_j \leftarrow \sum_t p_{j+1}^t; \quad p_{j+1}^t \leftarrow p_{j+1}^t / Z_j$

Testing:

Given x, calculate $d_j(x), j = 1, \ldots, L$

Calculate class outputs, $i = 1, \ldots, K$:

$$y_i = \sum_{j=1}^L \left(\log \frac{1}{\beta_j}\right) d_{ji}(x)$$

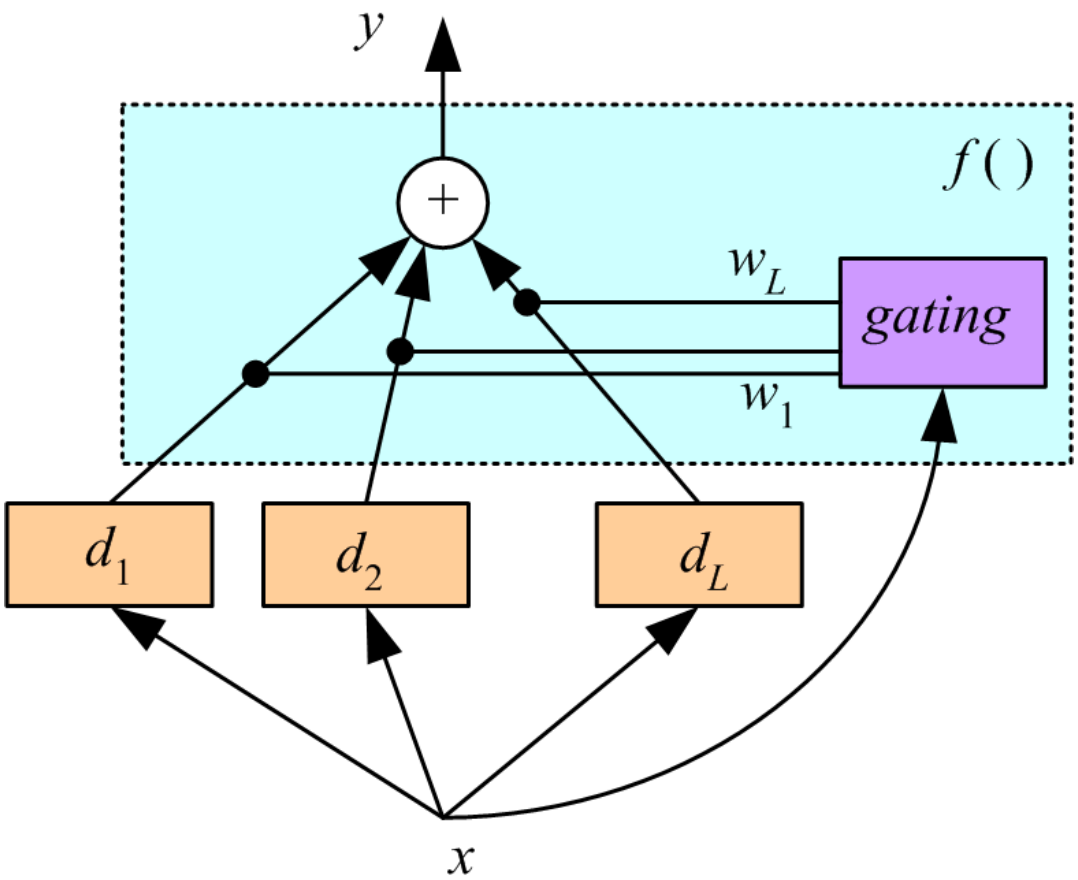
Grupa eksperata

Glasanje sa težinama koje zavise od ulaza (gating)

$$y = \sum_{j=1}^{L} w_j d_j$$

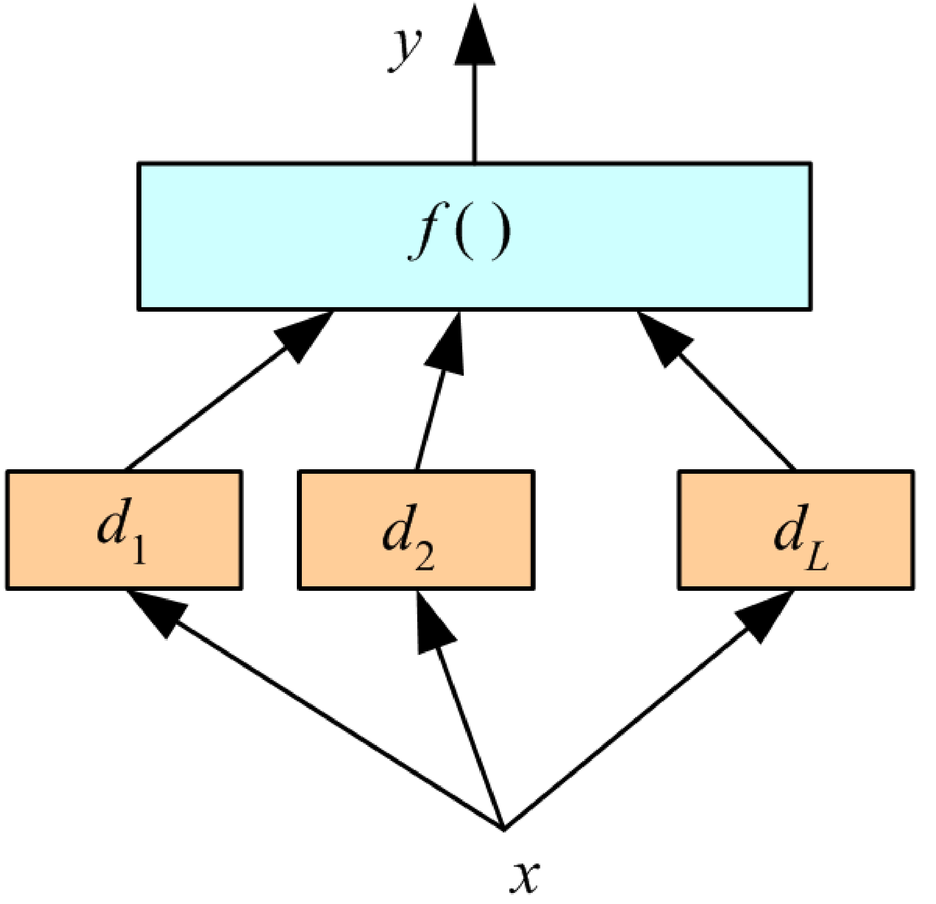
(Jacobs et al., 1991)

Eksperti i gejting mogu biti nelinearni



Stacking

- Kombajner *f* () je zaseban learner (Wolpert, 1992)



Fino podešavanje ansambla

- Ako je dat ansambl zavisnih klasifikatora, ne koristiti ih kao takve, već pokušati da se postigne nezavisnost

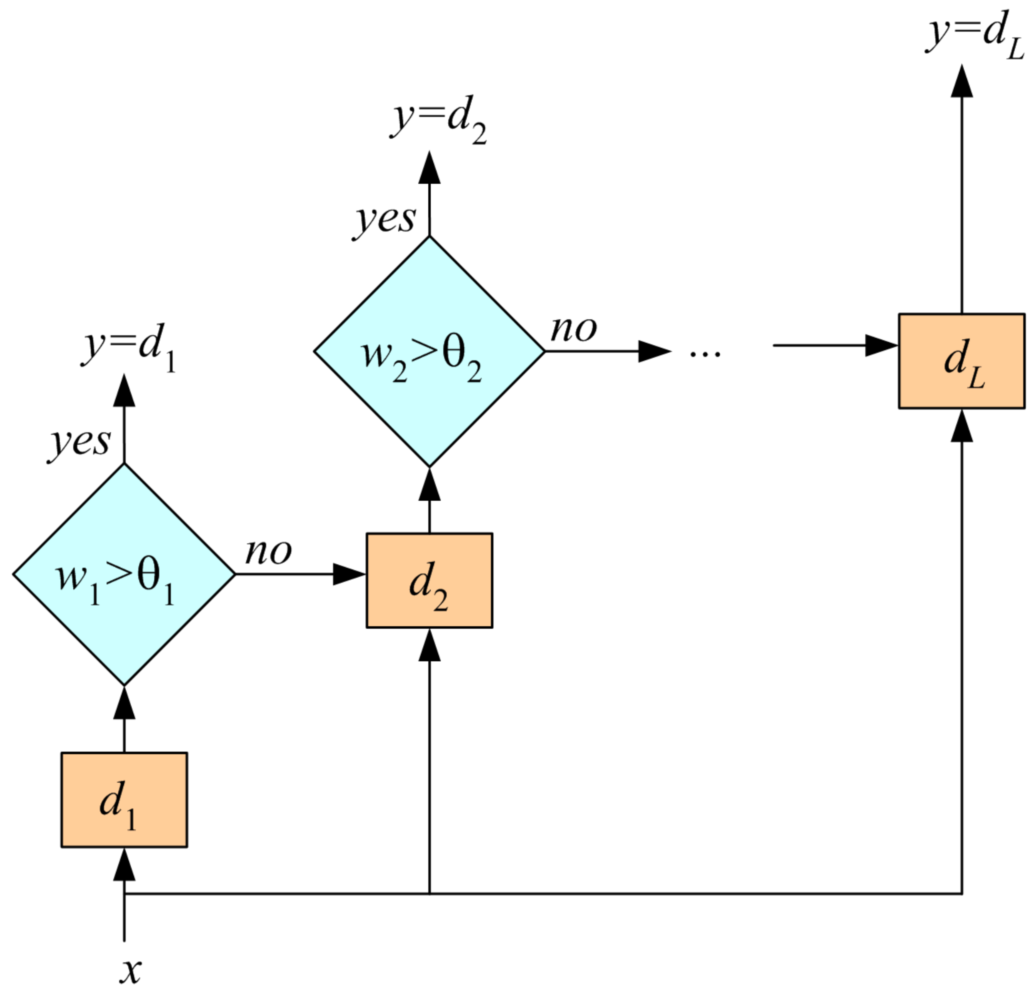
1. Selekcija podskupova: Pristup Unapred (rastući)/Unazad (kresanje) u cilju poboljšanja tačnosti/diverziteta/nezavisnosti
2. Obučavanje metaklasifikatora: od izlaza korelisanih klasifikatora, ekstrahovati nekorelisanu kombinaciju. Ako koristimo PCA, dobijamo “eigenlearners.”

- Slično sa selekcija vs ekstrakcija obeležja

Kaskade

Koristiti d_j samo ako prethodni klasifikator nije dovoljno dobar.

Kaskade se formiraju u redosledu povećanja kompleksnosti klasifikatora.



Kombinovanje više izvora/pogleda

- Rana integracija: Konkatenirati sva obeležja i obučiti jedinstven klasifikator.
- Kasna integracija: Na svakom skupu obeležja obučiti jedan klasifikator, a zatim objediniti njihove odluke na osnovu nekog fiksnog pravila ili stakingom.
- Intermedijarna integracija: Na svakom skupu obeležja izračunati jezgro (kernel), a zatim formirati jedinstveni SVM klasifikator sa višestrukim jezgrima.
- Kombinovanje obeležja vs klasifikacija vs jezgra.