Linearna klasifikacija i perceptronsko učenje

- Linearna klasifikacija
  - Geometrijska interpretacija linearne klasifikacije
  - Perceptronsko učenje

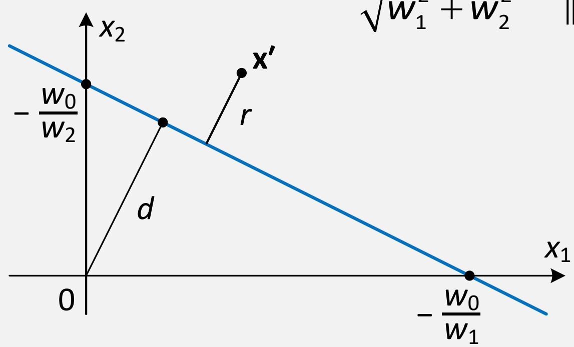
Jednačine prave i ravni (podsećanje)

Jednačina prave u ravni (2D)

$$w_1 x_1 + w_2 x_2 + w_0 = 0$$

$$\mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0 = 0$$

$$d = \frac{|w_0|}{\sqrt{w_1^2 + w_2^2}} = \frac{|w_0|}{\|\mathbf{w}\|}$$



Rastojanje proizvoljne tačke **x'** od prave $\mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0 = 0$:

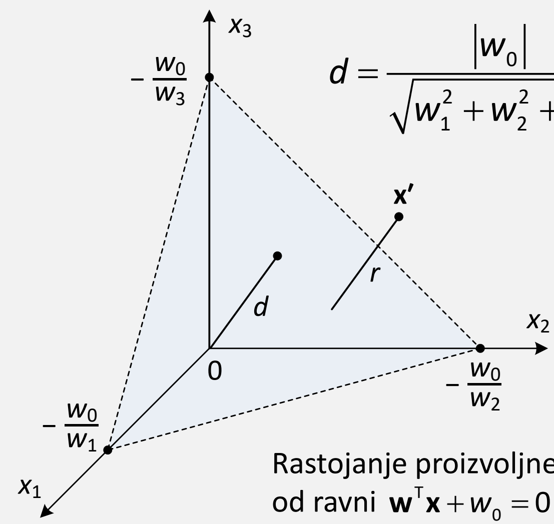
$$r = \frac{|\mathbf{w}^{\mathrm{T}}\mathbf{x}' + w_0|}{\sqrt{w_1^2 + w_2^2}} = \frac{|\mathbf{w}^{\mathrm{T}}\mathbf{x}' + w_0|}{\|\mathbf{w}\|}$$

Jednačina ravni u prostoru (3D)

$$w_1 x_1 + w_2 x_2 + w_3 x_3 + w_0 = 0$$

$$\mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0 = 0$$

$$d = \frac{|w_0|}{\sqrt{w_1^2 + w_2^2 + w_3^2}} = \frac{|w_0|}{\|\mathbf{w}\|}$$



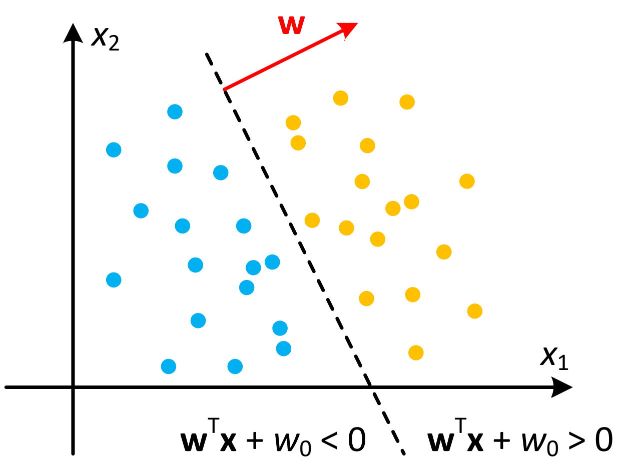
Rastojanje proizvoljne tačke **x'** od ravni $\mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0 = 0$:

$$r = \frac{|\mathbf{w}^{\mathrm{T}}\mathbf{x}' + w_0|}{\sqrt{w_1^2 + w_2^2 + w_3^2}} = \frac{|\mathbf{w}^{\mathrm{T}}\mathbf{x}' + w_0|}{\|\mathbf{w}\|}$$

Linearna klasifikacija

- Linearna klasifikacija u dve klase može se realizovati na osnovu linearne regresije tako što nepoznati uzorak dodeljujemo jednoj ili drugoj klasi u zavisnosti od toga da li je vrednost izlazne veličine pozitivna ili negativna

$$g(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0, \quad g(\mathbf{x}) \underset{\omega_2}{\overset{\omega_1}{\gtrless}} 0$$



- Uzorci dveju klasa su *linearno razdvojivi* ako postoji hiperravan $g(\mathbf{x}) = 0$ koja predstavlja granicu odlučivanja između njih
- Ako su $\mathbf{x}_1$ i $\mathbf{x}_2$ dve tačke na hiperravni odlučivanja, važi sledeće:

$$\mathbf{w}^{\mathrm{T}}\mathbf{x}_1 + w_0 = \mathbf{w}^{\mathrm{T}}\mathbf{x}_2 + w_0 = 0,$$

odakle se dobija:

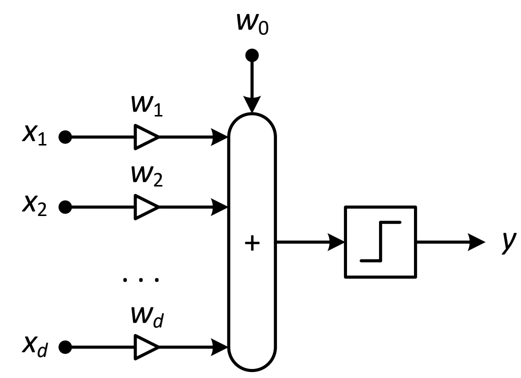
$$\mathbf{w}^{\mathrm{T}}(\mathbf{x}_1 - \mathbf{x}_2) = 0,$$

što znači da je vektor **w** *ortogonalan* na hiperravan odlučivanja

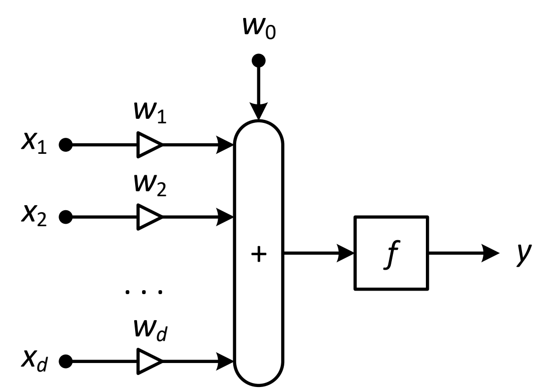
- Količnik $\dfrac{|g(\mathbf{x})|}{\|\mathbf{w}\|}$ predstavlja euklidsko rastojanje tačke **x** od hiperravni odlučivanja

Perceptron

- Jednostavan koncept mašine za učenje namenjene rešavanju problema linearne klasifikacije
  - Perceptron izračunava skalarnu funkciju više ulaznih promenljivih x_i pomnoženih težinskim faktorima w_i
    - U slučaju binarne klasifikacije, uzorak se klasifikuje u jednu od dve klase na osnovu znaka izlaza, što se može predstaviti step-funkcijom sa izlazima $y = \pm 1$
  - Perceptron je u stanju da automatski formira model, odnosno, odredi vrednosti težinskih koeficijenata $\mathbf{w}$ na osnovu uzoraka, koristeći određeni *algoritam učenja*



- Perceptron se često nadograđuje tako što se umesto step-funkcije na izlazu uvede neka druga, po pravilu neprekidna nelinearna funkcija
  - Ovako modifikovan perceptron naziva se *neuron* i predstavlja osnovu veštačkih neuronskih mreža
  - Pošto izlaz predstavlja nelinearnu funkciju linearne kombinacije ulaznih promenljivih (često sigmoid, i u tom slučaju je zapravo reč o *logističkoj regresiji*)



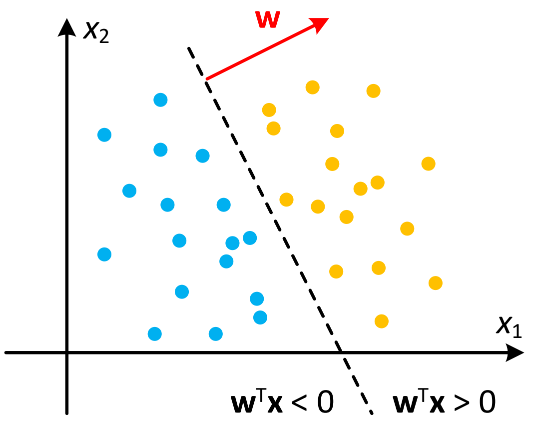
Perceptronsko učenje (problem dve klase)

- Neka skup za obuku $\mathbf{X} = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)},..., \mathbf{x}^{(N)}\}$ sadrži uzorke koji pripadaju klasama ω_1 i ω_2 i neka važi pretpostavka njihove linearne razdvojivosti
  - Prag w_0 podvodi se pod težinski vektor **w** uvođenjem pomoćnog obeležja $x_0 = 1$ (nadalje će se smatrati da je $g(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$, pri čemu vektor **x** sadrži i obeležje x_0)

Zadatak perceptronskog učenja:

Naći vektor **w** takav da važi:

$$\mathbf{w}^T\mathbf{x} > 0, \forall \mathbf{x} \in \omega_1$$
$$\mathbf{w}^T\mathbf{x} < 0, \forall \mathbf{x} \in \omega_2$$



- Ako se uvedu oznake klasa $y_1 = 1$ za klasu ω_1 i $y_2 = -1$ za klasu ω_2, problem se svodi na nalaženje vektora **w** takvog da $y^{(i)} \cdot \mathbf{w}^T\mathbf{x}^{(i)} > 0, \forall \mathbf{x}^{(i)}$
  - Ovo je ekvivalentno promeni znaka svih uzoraka iz klase ω_2 ($\mathbf{x} \leftarrow [-\mathbf{x}], \forall \mathbf{x} \in \omega_2$) nakon čega je dovoljno naći vektor **w** takav da $\mathbf{w}^T\mathbf{x} > 0, \forall \mathbf{x}$ (oznake klasa mogu se ignorisati)

Napomena: Uvođenje obeležja x_0 koje je uvek jednako 1 značajno komplikuje geometrijsku interpretaciju jer povećava dimenzionalnost problema za 1. Konkretno, u datom primeru i vektor **x** i vektor **w** zapravo su trodimenzionalni, pa se situacija više ne može adekvatno predstaviti u samo dve dimenzije.

Perceptronsko učenje (problem dve klase)

- Sada treba naći pogodnu funkciju greške $J(\mathbf{w})$ koju treba minimizovati
  - $J(\mathbf{w})$ treba da predstavlja meru greške klasifikacije na osnovu proizvoljne hiperravni $\mathbf{w}$
  - $J(\mathbf{w})$ treba da ima vrednost 0 ako su svi uzorci ispravno klasifikovani, a treba da bude tim veća što je više pogrešno klasifikovanih uzoraka i što su pogrešno klasifikovani uzorci udaljeniji od hiperravni $\mathbf{w}$
- Nakon promene znaka uzorcima iz klase ω_2, čime je uslov sveden na $\mathbf{w}^{\mathrm{T}}\mathbf{x} > 0, \forall \mathbf{x}$, za funkciju greške može se usvojiti tzv. perceptronska kriterijumska funkcija:

$$J_{\mathrm{P}}(\mathbf{w}) = \sum_{\mathbf{x} \in X_{\mathrm{M}}} (-\mathbf{w}^{\mathrm{T}}\mathbf{x}),$$

  gde je X_{M} skup uzoraka pogrešno klasifikovanih pomoću $\mathbf{w}$
  - $J_{\mathrm{P}}(\mathbf{w})$ je jednaka 0 kada su svi uzorci ispravno klasifikovani, jer je tada $X_{\mathrm{M}} = \varnothing$
  - U ostalim slučajevima $J_{\mathrm{P}}(\mathbf{w})$ je pozitivna jer za pogrešno klasifikovane uzorke važi $\mathbf{w}^{\mathrm{T}}\mathbf{x} < 0$
- $J_{\mathrm{P}}(\mathbf{w})$ se minimizuje metodom gradijentnog silaska
  - Ako su uzorci linearno separabilni, algoritam gradijentnog silaska dolazi do rešenja u konačno mnogo koraka

Perceptronsko učenje (problem dve klase)

- Gradijent perceptronske kriterijumske funkcije jednak je:

$$\nabla_{\mathbf{w}} J_{\mathrm{P}}(\mathbf{w}) = \nabla_{\mathbf{w}} \sum_{\mathbf{x}\in X_{\mathrm{M}}} (-\mathbf{w}^{\mathrm{T}}\mathbf{x}) = \sum_{\mathbf{x}\in X_{\mathrm{M}}} (-\mathbf{x})$$

  pa se vrednost **w** po metodi gradijentnog silaska obnavlja po pravilu:

$$\mathbf{w}(k+1) = \mathbf{w}(k) - \alpha \nabla_{\mathbf{w}} J_{\mathrm{P}}(\mathbf{w}(k)) = \mathbf{w}(k) + \alpha \sum_{\mathbf{x}\in X_{\mathrm{M}}} \mathbf{x}$$

- Ovo pravilo poznato je kao *grupno perceptronsko pravilo ažuriranja*
  - U obnavljanju vrednosti **w** učestvuju svi pogrešno klasifikovani uzorci, odnosno, njihova suma

- Početi od proizvoljne vrednosti $\mathbf{w}(0)$
- Odrediti skup pogrešno klasifikovanih uzoraka X_{M} za dato $\mathbf{w}(k)$, a zatim korigovati vektor $\mathbf{w}(k)$:

$$\mathbf{w}(k+1) = \mathbf{w}(k) - \alpha \nabla_{\mathbf{w}} J_{\mathrm{P}}(\mathbf{w}(k)) = \mathbf{w}(k) + \alpha \sum_{\mathbf{x}\in X_{\mathrm{M}}} \mathbf{x}$$

  gde je α brzina učenja (fiksni mali broj)
- Ponavljati prethodni korak do ispravne klasifikacije svih uzoraka (dok ne bude $X_{\mathrm{M}} = \emptyset$)

Perceptronsko učenje (problem dve klase)

- Težinski vektor **w** može obnavljati vrednost i na osnovu pojedinačnih uzoraka, i u tom slučaju uzorci $\mathbf{x}^{(i)}$ ulaze u algoritam jedan za drugim, najbolje slučajnim redosledom
  - Ako algoritam nije konvergirao nakon obrade svih uzoraka jednom, procedura se ponavlja dok se ne dostigne konvergencija (do ispravne klasifikacije svih uzoraka)
  - Perceptronsko pravilo ažuriranja svodi se na:

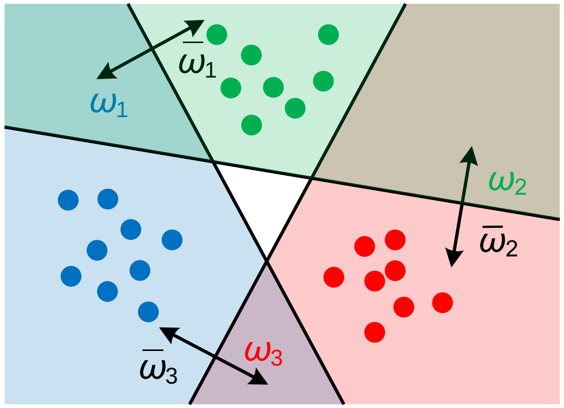
    $$\mathbf{w}(k+1) = \mathbf{w}(k) + \alpha \mathbf{x}^{(k)}$$

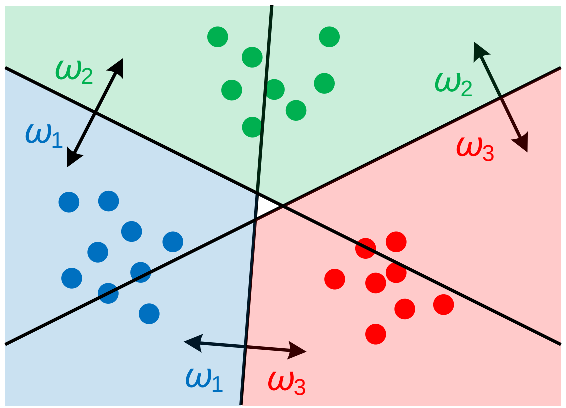
    gde je $\mathbf{x}^{(k)}$ neki od uzoraka koji su u k-tom koraku bili pogrešno klasifikovani pomoću $\mathbf{w}(k)$
- Ako su klase linearno razdvojive, obe verzije perceptronskog učenja (grupna i pojedinačna) konvergiraju ka ispravnom rešenju u konačno mnogo koraka
- Ako klase nisu linearno razdvojive, perceptronsko učenje neće konvergirati
  - Nijedan težinski vektor **w** ne može ispravno klasifikovati *svaki* uzorak, pa će nakon svakog koraka algoritma ponovo biti pogrešno klasifikovanih uzoraka, a samim tim i novih promena
  - Jedno rešenje je u usvajanju promenljive brzine učenja $\alpha(k)$ koja teži ka 0 kada $k \to \infty$
  - Varijanta algoritma koja sprečava oscilovanje između različitih rešenja bez mogućnosti zaustavljanja naziva se *algoritam džepa*
    - Ideja je da algoritam tokom iterativnog procesa pamti „u džepu“ najbolje rešenje koje je pronađeno do datog trenutka i ažurira ga samo ako se pojavi bolje rešenje
    - Pod određenim uslovima algoritam džepa konvergira ka suboptimalnom rešenju, odnosno, dolazi do minimalnog broja pogrešnih klasifikacija

Klasifikacija u više od dve klase

- U praksi retko postoji linearna separabilnost između više klasa, tako da se linearna klasifikacija u više klasa uglavnom i ne radi pomoću perceptrona
- U opštem slučaju, klasifikacija u više od dve klase (ω_1, ω_2,..., ω_K) može se uvek realizovati kao višestruka primena klasifikacije u dve klase:
  - ω_1/ne ω_1; ω_2/ne ω_2;... ; ω_K/ne ω_K (eng. *one versus rest* – OVR)
  - ω_i/ω_j za svako i i j (eng. *one versus one* – OVO)

  a zatim odlučivanjem po većinskom principu, ali se pokazuje da oba načina imaju nedostatke (mogu se javiti sporni regioni nedefinisane pripadnosti)





Klasifikacija u više od dve klase

- Pristup „jedan protiv svih" (OVR) u praksi se modifikuje tako da se klasifikacija u više klasa vrši na osnovu diskriminantnih funkcija, čime se prevazilazi problem spornih regiona nedefinisane pripadnosti
  - Ponovo se svakoj klasi ω_i pridružuje po jedan perceptron, čiji se težinski vektor $\mathbf{w}_i$ određuje perceptronskim učenjem na klasama ω_i i „ne ω_i", pri čemu taj vektor ujedno određuje i odgovarajuću diskriminantnu funkciju $g_i(\boldsymbol{x}) = \mathbf{w}_i^T\mathbf{x} + w_0$
  - Odlučivanje se ne vrši većinski, već klasifikator dodeljuje uzorak opisan vektorom obeležja $\boldsymbol{x}$ klasi ω_i ako je $g_i(\boldsymbol{x}) > g_j(\boldsymbol{x})$ za svako $j \neq i$, odnosno, uzorak se dodeljuje klasi čiji je perceptron dao najveći *odziv* na $\mathbf{x}$
  - Pošto su diskriminantne funkcije linearne, površi odlučivanja predstavljaju hiperravni