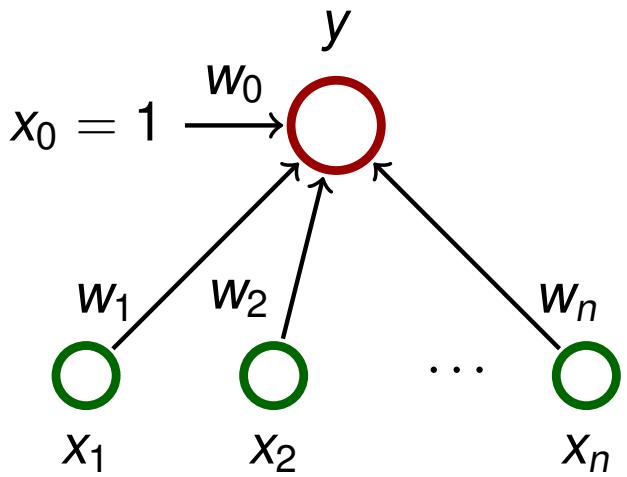
Perceptron a ADALINE

- Perceptron
- Perceptronové učící pravidlo
- Konvergence učení perceptronu
- ADALINE
- Učení ADALINE

Perceptron

Organizační dynamika:



$\vec{w} = (w_0, w_1, \ldots, w_n)$ a $\vec{x} = (x_0, x_1, \ldots, x_n)$ kde $x_0 = 1$.

Aktivní dynamika:

- vnitřní potenciál: $\xi = w_0 + \sum_{i=1}^{n} w_i x_i = \sum_{i=0}^{n} w_i x_i = \vec{w} \cdot \vec{x}$
- aktivační funkce: $\sigma(\xi) = \begin{cases} 1 & \xi \geq 0\,; \\ 0 & \xi < 0. \end{cases}$
- funkce sítě: $y[\vec{w}](\vec{x}) = \sigma(\xi) = \sigma(\vec{w} \cdot \vec{x})$

Perceptron

Adaptivní dynamika:

- Dána množina **tréninkových vzorů**

$$\mathcal{T} = \left\{\left(\vec{x}_1, d_1\right), \left(\vec{x}_2, d_2\right), \ldots, \left(\vec{x}_p, d_p\right)\right\}$$

Zde $\vec{x}_k = (x_{k0}, x_{k1} \ldots, x_{kn}) \in \mathbb{R}^{n+1}$, $x_{k0} = 1$, je vstup k-tého vzoru a $d_k \in \{0, 1\}$ je očekávaný výstup.

(d_k určuje, do které ze dvou kategorií dané $\vec{x}_k = (x_{k0}, x_{k1} \ldots, x_{kn})$ patří).

- Vektor vah $\vec{w} \in \mathbb{R}^{n+1}$ je **konzistentní s** $\mathcal{T}$ pokud $y[\vec{w}](\vec{x}_k) = \sigma(\vec{w} \cdot \vec{x}_k) = d_k$ pro každé $k = 1, \ldots, p$.

  Množina $\mathcal{T}$ je **vnitřně konzistentní** pokud existuje vektor $\vec{w}$, který je s ní konzistentní.

- Cílem je nalézt vektor $\vec{w}$, který je konzistentní s $\mathcal{T}$ za předpokladu, že $\mathcal{T}$ je vnitřně konzistentní.

Perceptron - adaptivní dynamika

Online učící algoritmus:

Idea: Cyklicky prochází vzory a adaptuje podle nich váhy, tj. otáčí dělící nadrovinu tak, aby se zmenšila vzdálenost špatně klasifikovaného vzoru od jeho příslušného poloprostoru.

Prakticky počítá posloupnost vektorů vah $\vec{w}^{(0)}, \vec{w}^{(1)}, \vec{w}^{(2)}, \ldots$.

- váhy v $\vec{w}^{(0)}$ jsou inicializovány náhodně blízko 0
- v kroku $t+1$ je $\vec{w}^{(t+1)}$ vypočteno takto:

$$\begin{aligned} \vec{w}^{(t+1)} &= \vec{w}^{(t)} - \varepsilon \cdot \left(y[\vec{w}^{(t)}](\vec{x}_k) - d_k \right) \cdot \vec{x}_k \\ &= \vec{w}^{(t)} - \varepsilon \cdot \left(\sigma\left(\vec{w}^{(t)} \cdot \vec{x}_k\right) - d_k \right) \cdot \vec{x}_k \end{aligned}$$

Zde $k = (t \mod p) + 1$ (tj. cyklické procházení vzorů) a $0 < \varepsilon \leq 1$ je **rychlost učení**.

Věta (Rosenblatt)

Jestliže je $\mathcal{T}$ vnitřně konzistentní, pak existuje t^* takové, že $\vec{w}^{(t^*)}$ je konzistentní s $\mathcal{T}$.

Důkaz Rosenblattovy věty

Pro zjednodušení budeme dále předpokládat, že $\varepsilon = 1$.

Nejprve si algoritmus přepíšeme do méně kompaktní formy:

- váhy v $\vec{w}^{(0)}$ jsou inicializovány náhodně blízko 0
- v kroku $t+1$ je $\vec{w}^{(t+1)}$ vypočteno takto:
  - **Jestliže** $\sigma(\vec{w}^{(t)} \cdot \vec{x}_k) = d_k$, **pak** $\vec{w}^{(t+1)} = \vec{w}^{(t)}$
  - **Jestliže** $\sigma(\vec{w}^{(t)} \cdot \vec{x}_k) \neq d_k$, **pak**
    - $\vec{w}^{(t+1)} = \vec{w}^{(t)} + \vec{x}_k$ pro $d_k = 1$
    - $\vec{w}^{(t+1)} = \vec{w}^{(t)} - \vec{x}_k$ pro $d_k = 0$

    (Řekneme, že nastala korekce.)

kde $k = (t \mod p) + 1$.

Důkaz Rosenblattovy věty

(Pro daný vektor $\vec{a} = (a_0, \ldots, a_n)$ označme $\left\|\vec{a}\right\|$ jeho eukleidovskou normu $\sqrt{\vec{a} \cdot \vec{a}} = \sqrt{\sum_{i=0}^{n} a_i^2}$)

Idea:

- Uvážíme *hodně dlouhý vektor* (spočítáme jak dlouhý) $\vec{w}^*$, který je konzistentní s $\mathcal{T}$.
- Ukážeme, že pokud došlo v kroku $t+1$ ke korekci vah (tedy buď $\vec{w}^{(t+1)} = \vec{w}^{(t)} + \vec{x}_k$ nebo $\vec{w}^{(t+1)} = \vec{w}^{(t)} - \vec{x}_k$), pak

$$\left\|\vec{w}^{(t+1)} - \vec{w}^*\right\|^2 \leq \left\|\vec{w}^{(t)} - \vec{w}^*\right\|^2 - \max_i \left\|\vec{x}_i\right\|$$

  Všimněte si, že $\max_i \left\|\vec{x}_i\right\| > 0$ *nezávisí* na t.
- Z toho plyne, že algoritmus nemůže udělat nekonečně mnoho korekcí.

Důkaz Rosenblattovy věty

Uvažme vektor $\vec{w}^*$ konzistentní s $\mathcal{T}$.

Búno předpokládejme, že $\vec{w}^* \cdot \vec{x}_k \neq 0$ pro $k = 1, \ldots, p$.

Předp., že v kroku $t+1$ došlo ke korekci, a že $k = (t \mod p) + 1$.

Ukážeme, že

$$\left\|\vec{w}^{(t+1)} - \vec{w}^*\right\|^2 \quad \leq \quad \left\|\vec{w}^{(t)} - \vec{w}^*\right\|^2 \quad + \quad \left\|\vec{x}_k\right\|^2 - 2|\vec{w}^* \cdot \vec{x}_k|$$

(Potom bude k důkazu věty stačit nahlédnout, že pro "dlouhý" vektor $\vec{w}^*$ je $|\vec{w}^* \cdot \vec{x}_k|$ "velké" kladné číslo.)

Rozlišíme dva případy: $\vec{d}_k = 1$ a $\vec{d}_k = 0$.

Důkaz Rosenblattovy věty

Předpokládejme $\vec{d}_k = 1$:

Došlo ke korekci, tedy $\vec{w}^{(t+1)} = \vec{w}^{(t)} + \vec{x}_k$ a tedy $\vec{w}^{(t+1)} - \vec{w}^* = (\vec{w}^{(t)} - \vec{w}^*) + \vec{x}_k$. Pak

$$\begin{aligned}
\left\|\vec{w}^{(t+1)} - \vec{w}^*\right\|^2 &= \left\|(\vec{w}^{(t)} - \vec{w}^*) + \vec{x}_k\right\|^2 \\
&= \left[(\vec{w}^{(t)} - \vec{w}^*) + \vec{x}_k\right]\left[(\vec{w}^{(t)} - \vec{w}^*) + \vec{x}_k\right] \\
&= \left\|(\vec{w}^{(t)} - \vec{w}^*)\right\|^2 + \left\|\vec{x}_k\right\|^2 + 2(\vec{w}^{(t)} - \vec{w}^*) \cdot \vec{x}_k \\
&= \left\|(\vec{w}^{(t)} - \vec{w}^*)\right\|^2 + \left\|\vec{x}_k\right\|^2 + 2\vec{w}^{(t)} \cdot \vec{x}_k - 2\vec{w}^* \cdot \vec{x}_k \\
&\leq \left\|(\vec{w}^{(t)} - \vec{w}^*)\right\|^2 + \left\|\vec{x}_k\right\|^2 - 2|\vec{w}^* \cdot \vec{x}_k|
\end{aligned}$$

Poslední nerovnost plyne z toho, že:

- došlo ke korekci při $\vec{d}_k = 1$, tedy muselo platit $\vec{w}^{(t)} \cdot \vec{x}_k < 0$,
- $\vec{w}^*$ je konzistentní s $\mathcal{T}$ a tedy $\vec{w}^* \cdot \vec{x}_k > 0$.

Důkaz Rosenblattovy věty

Předpokládejme $\vec{d}_k = 0$:

Došlo ke korekci, tedy $\vec{w}^{(t+1)} = \vec{w}^{(t)} - \vec{x}_k$ a tedy $\vec{w}^{(t+1)} - \vec{w}^* = (\vec{w}^{(t)} - \vec{w}^*) - \vec{x}_k$. Pak

$$\begin{aligned}
\left\|\vec{w}^{(t+1)} - \vec{w}^*\right\|^2 &= \left\|(\vec{w}^{(t)} - \vec{w}^*) - \vec{x}_k\right\|^2 \\
&= \left[(\vec{w}^{(t)} - \vec{w}^*) - \vec{x}_k\right]\left[(\vec{w}^{(t)} - \vec{w}^*) - \vec{x}_k\right] \\
&= \left\|(\vec{w}^{(t)} - \vec{w}^*)\right\|^2 + \left\|\vec{x}_k\right\|^2 - 2(\vec{w}^{(t)} - \vec{w}^*) \cdot \vec{x}_k \\
&= \left\|(\vec{w}^{(t)} - \vec{w}^*)\right\|^2 + \left\|\vec{x}_k\right\|^2 - 2\vec{w}^{(t)} \cdot \vec{x}_k + 2\vec{w}^* \cdot \vec{x}_k \\
&\leq \left\|(\vec{w}^{(t)} - \vec{w}^*)\right\|^2 + \left\|\vec{x}_k\right\|^2 - 2|\vec{w}^* \cdot \vec{x}_k|
\end{aligned}$$

Poslední nerovnost plyne z toho, že:

- došlo ke korekci při $\vec{d}_k = 0$, tedy muselo platit $\vec{w}^{(t)} \cdot \vec{x}_k \geq 0$,
- $\vec{w}^*$ je konzistentní s $\mathcal{T}$ a tedy $\vec{w}^* \cdot \vec{x}_k < 0$.

Důkaz Rosenblattovy věty

Máme dokázáno:

$$\left\|\vec{w}^{(t+1)} - \vec{w}^*\right\|^2 \leq \left\|\vec{w}^{(t)} - \vec{w}^*\right\|^2 + \left\|\vec{x}_k\right\|^2 - 2|\vec{w}^* \cdot \vec{x}_k|$$

Nechť nyní $\vec{w}^* = \alpha \cdot \vec{w}^+$ kde $\alpha > 0$. Pak

$$\left\|\vec{w}^{(t+1)} - \vec{w}^*\right\|^2 \leq \left\|\vec{w}^{(t)} - \vec{w}^*\right\|^2 + \left\|\vec{x}_k\right\|^2 - 2\alpha|\vec{w}^+ \cdot \vec{x}_k|$$

Nyní stačí uvážit $\alpha = \frac{\max_k \left\|\vec{x}_k\right\|^2}{\min_k |\vec{w}^+ \cdot \vec{x}_k|}$ a dostaneme

$$\left\|\vec{x}_k\right\|^2 - 2\alpha|\vec{w}^+ \cdot \vec{x}_k| \leq \left\|\vec{x}_k\right\|^2 - 2\max_k \left\|\vec{x}_k\right\|^2 \frac{|\vec{w}^+ \cdot \vec{x}_k|}{\min_k |\vec{w}^+ \cdot \vec{x}_k|} \leq -\max_k \left\|\vec{x}_k\right\|^2$$

Což dá

$$\left\|\vec{w}^{(t+1)} - \vec{w}^*\right\| \leq \left\|\vec{w}^{(t)} - \vec{w}^*\right\|^2 - \max_k \left\|\vec{x}_k\right\|^2$$

kdykoliv došlo ke korekci.

Z toho plyne, že nemůže dojít k nekonečně mnoha korekcím. □

Perceptron - adaptivní dynamika

Dávkový učící algoritmus:

Vypočte posloupnost $\vec{w}^{(0)}, \vec{w}^{(1)}, \vec{w}^{(2)}, \ldots$ váhových vektorů.

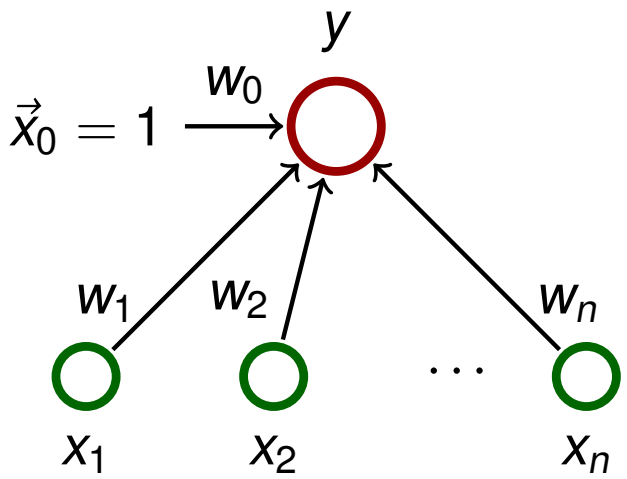
- váhy v $\vec{w}^{(0)}$ jsou inicializovány náhodně blízko 0
- v kroku $t+1$ je $\vec{w}^{(t+1)}$ vypočteno takto:

$$\vec{w}^{(t+1)} = \vec{w}^{(t)} - \varepsilon \cdot \sum_{k=1}^{p} (\sigma(\vec{w}^{(t)} \cdot \vec{x}_k) - d_k) \cdot \vec{x}_k$$

Zde $k = (t \mod p) + 1$
a $0 < \varepsilon \leq 1$ je **rychlost učení**.

ADALINE

Organizační dynamika:



$\vec{w} = (w_0, w_1, \ldots, w_n)$ a $\vec{x} = (x_0, x_1, \ldots, x_n)$ kde $x_0 = 1$.

Aktivní dynamika:

- vnitřní potenciál: $\xi = w_0 + \sum_{i=1}^{n} w_i x_i = \sum_{i=0}^{n} w_i x_i = \vec{w} \cdot \vec{x}$
- aktivační funkce: $\sigma(\xi) = \xi$
- funkce sítě: $y[\vec{w}](\vec{x}) = \sigma(\xi) = \vec{w} \cdot \vec{x}$

ADALINE

Adaptivní dynamika:

- Dána množina **tréninkových vzorů**

$$\mathcal{T} = \left\{\left(\vec{x}_1, d_1\right), \left(\vec{x}_2, d_2\right), \ldots, \left(\vec{x}_p, d_p\right)\right\}$$

Zde $\vec{x}_k = (x_{k0}, x_{k1} \ldots, x_{kn}) \in \mathbb{R}^{n+1}$, $x_{k0} = 1$, je vstup k-tého vzoru a $d_k \in \mathbb{R}$ je očekávaný výstup.

Intuice: chceme, aby síť počítala afinní aproximaci funkce, jejíž (některé) hodnoty nám předepisuje tréninková množina.

- **Chybová funkce:**

$$E(\vec{w}) = \frac{1}{2}\sum_{k=1}^{p}\left(\vec{w}\cdot\vec{x}_k - d_k\right)^2 = \frac{1}{2}\sum_{k=1}^{p}\left(\sum_{i=0}^{n} w_i x_{ki} - d_k\right)^2$$

- Cílem je nalézt $\vec{w}$, které minimalizuje $E(\vec{w})$.

Gradient chybové funkce

Uvažme **gradient** chybové funkce:

$$\nabla E(\vec{w}) = \left(\frac{\partial E}{\partial w_0}(\vec{w}), \ldots, \frac{\partial E}{\partial w_n}(\vec{w}) \right)$$

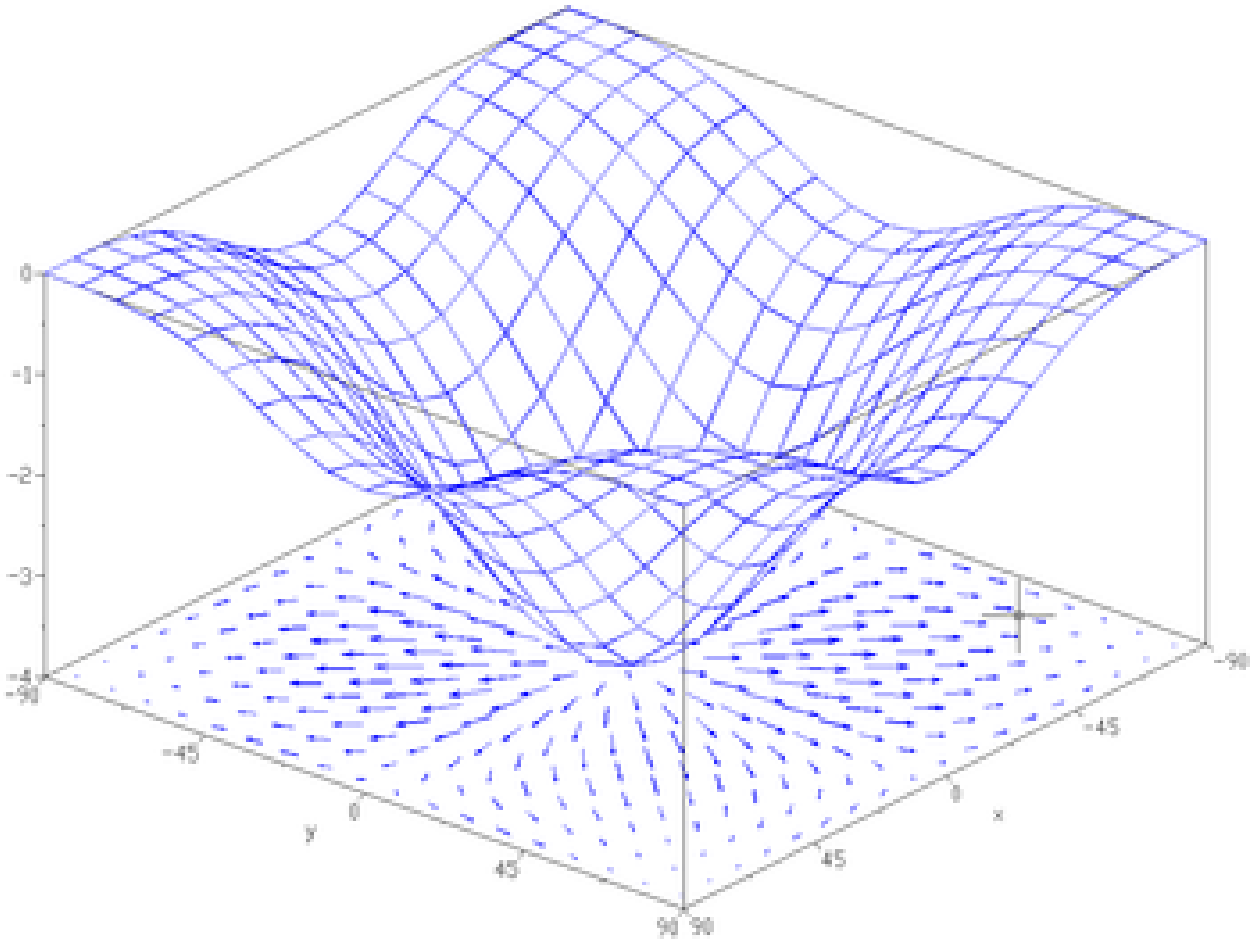
Intuice: $\nabla E(\vec{w})$ je vektor ve **váhovém prostoru**, který ukazuje směrem nejstrmějšího „růstu“ funkce $E(\vec{w})$. Vektory $\vec{x}_k$ zde slouží pouze jako parametry funkce $E(\vec{w})$ a jsou tedy fixní!

Fakt

Pokud $\nabla E(\vec{w}) = \vec{0} = (0, \ldots, 0)$, *pak* $\vec{w}$ *je globální minimum funkce* E.

Námi uvažovaná chybová funkce $E(\vec{w})$ má globální minimum, protože je konvexním paraboloidem.

Gradient - ilustrace



Pozor! Tento obrázek pouze ilustruje pojem gradientu, nezobrazuje chybovou funkci $E(\vec{w})$

Gradient chybové funkce ADALINE

$$
\begin{aligned}
\frac{\partial E}{\partial w_\ell}(\vec{w}) &= \frac{1}{2}\sum_{k=1}^{p}\frac{\delta E}{\delta w_\ell}\left(\sum_{i=0}^{n} w_i x_{ki} - d_k\right)^2 \\
&= \frac{1}{2}\sum_{k=1}^{p} 2\left(\sum_{i=0}^{n} w_i x_{ki} - d_k\right)\frac{\delta E}{\delta w_\ell}\left(\sum_{i=0}^{n} w_i x_{ki} - d_k\right) \\
&= \frac{1}{2}\sum_{k=1}^{p} 2\left(\sum_{i=0}^{n} w_i x_{ki} - d_k\right)\left(\sum_{i=0}^{n}\left(\frac{\delta E}{\delta w_\ell} w_i x_{ki}\right) - \frac{\delta E}{\delta w_\ell} d_k\right) \\
&= \sum_{k=1}^{p}\left(\vec{w}\cdot\vec{x}_k - d_k\right) x_{k\ell}
\end{aligned}
$$

Tedy

$$
\nabla E(\vec{w}) = \left(\frac{\partial E}{\partial w_0}(\vec{w}), \ldots, \frac{\partial E}{\partial w_n}(\vec{w})\right) = \sum_{k=1}^{p}\left(\vec{w}\cdot\vec{x}_k - d_k\right)\vec{x}_k
$$

ADALINE - adaptivní dynamika

Dávkový algoritmus (gradientní sestup):

- váhy v $\vec{w}^{(0)}$ jsou inicializovány náhodně blízko 0
- v kroku $t+1$ je $\vec{w}^{(t+1)}$ vypočteno takto:

$$\begin{aligned}\vec{w}^{(t+1)} &= \vec{w}^{(t)} - \varepsilon \cdot \nabla E(\vec{w}^{(t)}) \\ &= \vec{w}^{(t)} - \varepsilon \cdot \sum_{k=1}^{p} \left(\vec{w}^{(t)} \cdot \vec{x}_k - d_k\right) \cdot \vec{x}_k\end{aligned}$$

Zde $k = (t \mod p) + 1$
a $0 < \varepsilon \leq 1$ je rychlost učení.

(Všimněte si, že tento algoritmus je téměř stejný jako pro perceptron, protože $\vec{w}^{(t)} \cdot \vec{x}_k$ je hodnota funkce sítě (tedy $\sigma(\vec{w}^{(t)} \cdot \vec{x}_k)$ kde $\sigma(\xi) = \xi$).)

Tvrzení

Pro dostatečně malé $\varepsilon > 0$ posloupnost $\vec{w}^{(0)}, \vec{w}^{(1)}, \vec{w}^{(2)}, \ldots$ konverguje (po složkách) ke globálnímu minimu funkce E (tedy k vektoru $\vec{w}$, který splňuje $\nabla E(\vec{w}) = \vec{0}$).

ADALINE - adaptivní dynamika

Online algoritmus (Delta-rule, Widrow-Hoff rule):

- váhy v $\vec{w}^{(0)}$ jsou inicializovány náhodně blízko 0
- v kroku $t+1$ je $\vec{w}^{(t+1)}$ vypočteno takto:

$$\vec{w}^{(t+1)} = \vec{w}^{(t)} - \varepsilon(t) \cdot \left(\vec{w}^{(t)} \cdot \vec{x}_k - d_k\right) \cdot \vec{x}_k$$

kde $k = t \mod p + 1$
a $0 < \varepsilon(t) \leq 1$ je rychlost učení v kroku $t+1$.

Všimněte si, že tento algoritmus nepracuje s celým gradientem, ale jenom s jeho částí, která přísluší právě zpracovávanému vzoru!

Věta (Widrow & Hoff)

Pokud $\varepsilon(t) = \frac{1}{t}$ pak $\vec{w}^{(0)}, \vec{w}^{(1)}, \vec{w}^{(2)}, \ldots$ konverguje ke globálnímu minimu chybové funkce E.

ADALINE - klasifikace

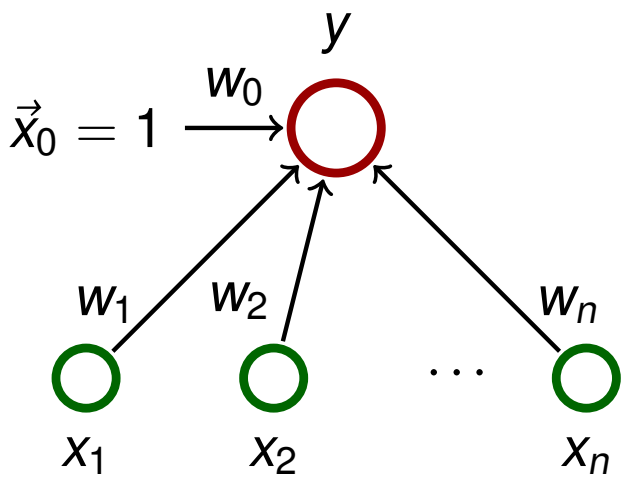
- Množina **tréninkových vzorů** je

$$\mathcal{T} = \left\{\left(\vec{x}_1, d_1\right), \left(\vec{x}_2, d_2\right), \ldots, \left(\vec{x}_p, d_p\right)\right\}$$

  kde $\vec{x}_k = (x_{k0}, x_{k1}, \ldots, x_{kn}) \in \mathbb{R}^{n+1}$ a $d_k \in \{1, -1\}$.
- Síť se natrénuje ADALINE algoritmem.
- Očekáváme, že bude platit následující:
  - jestliže $d_k = 1$, pak $\vec{w} \cdot \vec{x}_k \geq 0$
  - jestliže $d_k = -1$, pak $\vec{w} \cdot \vec{x}_k < 0$
- To nemusí vždy platit, ale často platí. Výhoda je, že se ADALINE algoritmus postupně stabilizuje i v neseparabilním případě (na rozdíl od perceptronového algoritmu).

ADALINE

Organizační dynamika:



$\vec{w} = (w_0, w_1, \ldots, w_n)$ a $\vec{x} = (x_0, x_1, \ldots, x_n)$ kde $x_0 = 1$.

Aktivní dynamika:

funkce sítě: $y[\vec{w}](\vec{x}) = \vec{w} \cdot \vec{x}$

ADALINE

Adaptivní dynamika:

- Dána množina **tréninkových vzorů**

$$\mathcal{T} = \left\{\left(\vec{x}_1, d_1\right), \left(\vec{x}_2, d_2\right), \ldots, \left(\vec{x}_p, d_p\right)\right\}$$

Zde $\vec{x}_k = (x_{k0}, x_{k1} \ldots, x_{kn}) \in \mathbb{R}^{n+1}$, $x_{k0} = 1$, je vstup k-tého vzoru a $d_k \in \mathbb{R}$ je očekávaný výstup.

Intuice: chceme, aby síť počítala afinní aproximaci funkce, jejíž (některé) hodnoty nám předepisuje tréninková množina.

- **Chybová funkce:**

$$E(\vec{w}) = \frac{1}{2}\sum_{k=1}^{p}\left(\vec{w}\cdot\vec{x}_k - d_k\right)^2 = \frac{1}{2}\sum_{k=1}^{p}\left(\sum_{i=0}^{n} w_i x_{ki} - d_k\right)^2$$

- Cílem je nalézt $\vec{w}$, které minimalizuje $E(\vec{w})$.

Dimenze $n = 1$

Dále budeme uvažovat pouze $n = 1$.

Hodnota sítě pro daný vstup $(1, x_1)$ bude $w_0 + w_1 x_1$

Tedy množina **tréninkových vzorů**

$$\mathcal{T} = \left\{\left(\vec{x}_1, d_1\right), \left(\vec{x}_2, d_2\right), \ldots, \left(\vec{x}_p, d_p\right)\right\}$$

splňuje $\vec{x}_k = (1, x_{k1}) \in \mathbb{R}^2$ a $d_k \in \mathbb{R}$

Zjednodušíme si notaci a **budeme předpokládat**

$$\mathcal{T} = \left\{(x_1, d_1), \ldots, (x_p, d_p)\right\}$$

kde $x_k \in \mathbb{R}$ a $d_k \in \mathbb{R}$ pro $k = 1, \ldots, p$.

Hodnota sítě s váhami w_0, w_1 pro k-tý vzor bude $w_0 + w_1 x_k$.

Chybová funkce pro $n = 1$

$$E(w_0, w_1) = \frac{1}{2} \sum_{k=1}^{p} (w_0 + w_1 x_k - d_k)^2$$

Minimalizujeme E vzhledem k w_0 a w_1:

$$\frac{\delta E}{\delta w_0} = 0 \quad \Leftrightarrow \quad w_0 = \bar{d} - w_1 \bar{x} \quad \Leftrightarrow \quad \bar{d} = w_0 + w_1 \bar{x}$$

kde $\bar{x} = \frac{1}{p} \sum_{k=1}^{p} x_k$ a $\bar{d} = \frac{1}{p} \sum_{k=1}^{p} d_k$

$$\frac{\delta E}{\delta w_1} = 0 \quad \Leftrightarrow \quad w_1 = \frac{\frac{1}{p} \sum_{k=1}^{p} (d_k - \bar{d})(x_k - \bar{x})}{\frac{1}{p} \sum_{k=1}^{p} (x_k - \bar{x})^2}$$

(tj. $w_1 = cov(d, x)/var(x)$)

Normální rozdělení pravděpodobnosti

Rozdělení spojité náhodné veličiny (tj. s hodnotami v $\mathbb{R}$)

Hustota pravděpodobnosti

$$f(x) \quad = \quad \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \quad =: \quad N[\mu, \sigma^2](x)$$

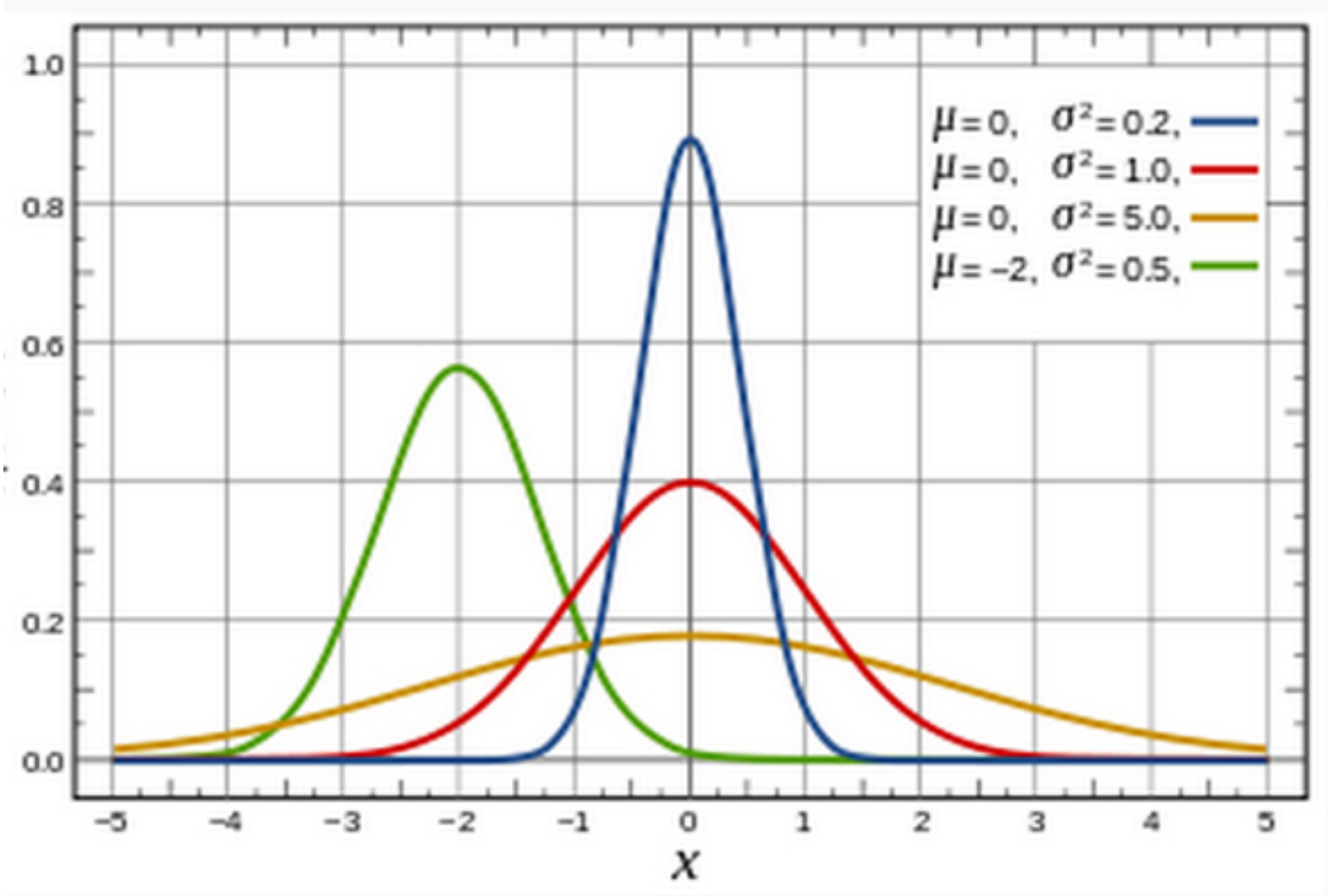
μ je střední hodnota, σ^2 rozptyl

Pokud má náhodná veličina X normální rozdělení, pak

$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} p(x)$$

Často se používá k vyjádření náhodné chyby, např. chyby měření, způsobené velkým počtem neznámých a vzájemně nezávislých příčin.

Normální rozdělení pravděpodobnosti



Věrohodnost (likelihood)

Fixujme $\mathcal{T} = \left\{(x_1, d_1), (x_2, d_2), \ldots, \left(x_p, d_p\right)\right\}$

Předpokládejme, že d_k bylo vygenerováno *náhodně* takto

$$d_k = w_0 + w_1 x_k + \epsilon_k$$

Zde

- w_0, w_1 jsou **neznámé konstanty**
- ϵ_k jsou generována náhodně s hustotou pravděpodobnosti $N[0, \sigma^2]$ kde σ^2 je **neznámý rozptyl**

Snadno se ukáže, že hustota pravděpodobnosti, se kterou je vygenerováno d_k splňuje

$$p(d_k \mid w_0, w_1, \sigma^2) = N[w_0 + w_1 x_k, \sigma^2](d_k)$$

Předpokládejme, že pro fixní w_0, w_1, σ^2 jsou $\epsilon_1, \ldots, \epsilon_p$ generována **nezávisle**. Pak hustota pravděpodobnosti, se kterou jsou vygenerována všechna $d_1, \ldots, d_p$ splňuje

$$p(d_1, \ldots, d_p \mid w_0, w_1, \sigma^2) = \prod_{k=1}^{p} N[w_0 + w_1 x_k, \sigma^2](d_k)$$

Maximální věrohodnost (maximum likelihood)

Chceme nalézt w_0, w_1, σ^2, která maximalizují

$$L(w_0, w_1, \sigma^2) := p(d_1, \ldots, d_p \mid w_0, w_1, \sigma^2)$$

Z technických důvodů budeme raději maximalizovat

$$\log(L(w_0, w_1, \sigma^2))$$

kde $\log(y)$ je přirozený logaritmus, tedy funkce inverzní k e^x.

Zřejmě

$$w_0, w_1, \sigma^2 \text{ maximalizují } L(w_0, w_1, \sigma^2)$$
$$\Leftrightarrow$$
$$w_0, w_1, \sigma^2 \text{ maximalizují } \log(L(w_0, w_1, \sigma^2))$$

Maximální log-věrohodnost (log-likelihood)

Ukážeme, že

$$\log(L(w_0, w_1, \sigma^2)) \quad = \quad -\frac{p}{2}\log 2\pi - p\log\sigma - \frac{1}{2\sigma^2}\sum_{k=1}^{p}(d_k - w_0 - w_1 x_k)^2$$

a tedy pro každé σ^2

$$w_0, w_1 \text{ maximalizují } \quad L(w_0, w_1, \sigma^2)$$
$$\Leftrightarrow$$
$$w_0, w_1 \text{ maximalizují } \quad \log(L(w_0, w_1, \sigma^2))$$
$$\Leftrightarrow$$
$$w_0, w_1 \text{ minimalizují } \quad E(w_0, w_1)$$

Tj. maximalizující w_0, w_1 nezávisí na σ^2.

Maximalizujeme-li vzhledem k σ^2, dostaneme

$$\sigma^2 = \frac{1}{p}\sum_{k=1}^{p}(d_k - w_0 - w_1 x_k)^2$$

(tj. průměrná čtvercová odchylka od žádaných hodnot d_k, jak se dalo čekat)

Věrohodnost (likelihood) - libovolná dimenze dat

Fixujme

$$\mathcal{T} = \left\{\left(\vec{x}_1, d_1\right), \left(\vec{x}_2, d_2\right), \ldots, \left(\vec{x}_p, d_p\right)\right\}$$

kde $\vec{x}_k \in \mathbb{R}^{n+1}$ a $d_k \in \mathbb{R}$ pro $k = 1, \ldots, p$.

Předpokládejme, že d_k bylo vygenerováno *náhodně* takto

$$d_k = \vec{w} \cdot \vec{x}_k + \epsilon_k = \sum_{i=0}^{n} w_k x_{ki} + \epsilon_k$$

Zde

- $\vec{w}$ je vektor **neznámých vah**
- ϵ_k jsou generována náhodně s hustotou pravděpodobnosti $N[0, \sigma^2]$ kde σ^2 je **neznámý rozptyl**

Pro fixní $\vec{w}, \sigma^2$ jsou $\epsilon_1, \ldots, \epsilon_p$ generována **nezávisle**. Pak $d_1, \ldots, d_p$ jsou generována s hustotou

$$p(d_1, \ldots, d_p \mid \vec{w}, \sigma^2) = \prod_{k=1}^{p} N[\vec{w} \cdot \vec{x}_k, \sigma^2](d_k)$$

Maximální log-věrohodnost (log-likelihood)

Pro

$$L(\vec{w}, \sigma^2) := p(d_1, \ldots, d_p \mid \vec{w}, \sigma^2) = \prod_{k=1}^{p} N[\vec{w} \cdot \vec{x}_k, \sigma^2](d_k)$$

platí

$$\log(L(\vec{w}, \sigma^2)) \quad = \quad -\frac{p}{2} \log 2\pi \; - \; p \log \sigma \; - \; \frac{1}{2\sigma^2} \sum_{k=1}^{p} (d_k - \vec{w} \cdot \vec{x}_k)^2$$

a tedy pro každé σ^2

$$\vec{w} \text{ maximalizuje } \quad L(\vec{w}, \sigma^2)$$
$$\Leftrightarrow$$
$$\vec{w} \text{ maximalizuje } \quad \log(L(\vec{w}, \sigma^2))$$
$$\Leftrightarrow$$
$$\vec{w} \text{ minimalizuje } \quad E(\vec{w})$$

Tj. maximalizující $\vec{w}$ nezávisí na σ^2.

Max. σ^2 splňuje $\sigma^2 = \frac{1}{p} \sum_{k=1}^{p} (d_k - \vec{w} \cdot \vec{x}_k)^2$