Vícevrstvé sítě II

Algoritmus BP - odvození

Opakování: co je to energie sítě?

Energie = míra naučenosti, přesněji nenaučenosti. Vzpomeň si:

$$E = \frac{1}{2}\sum_{i=1}^{n} (y_i - d_i)^2$$

kde n je počet výstupů sítě, y_i je i-tý výstup a d_i je i-tý požadovaný výstup.

$E = min$ znamená: síť je naučena

FAQ 1: odkud spadla tato definice energie?

FAQ 2: mají všechny neuronové sítě stejnou (tuhle) energetickou funkci?

FAQ 3: čeho (kterých proměnných) je tahle energetická funkce funkcí?

Ad 3:

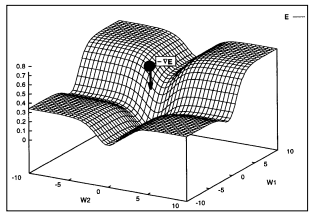
$$E = g(\vec{W}, \vec{\theta}, \vec{X})$$

$\vec{W}$ $\vec{\theta}$ považujme v tomto případě za proměnné,

$\vec{X}$ za konstantní.

Dovedete si představit geometrickou interpretaci energetické funkce?

Energetická funkce - geometrická interpretace



Problémy? Mnoho! O nich za chvilku.

„Lidská“ interpretace

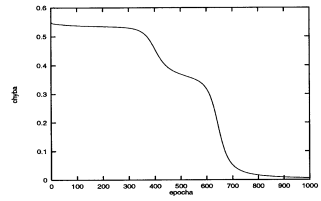
Tuto funkci si lze představit jako zakřivenou plochu v hyperprostoru připomínající pohoří. Okamžitá hodnota vah a prahů je souřadnicí, energie bodem na této ploše. Z tohoto bodu se snažíme postupně dosáhnout minima (energie).

Tuto situaci můžeme připodobnit chůzi po horském masívu za husté mlhy, kdy vidíme jen několik metrů před sebe. Když budeme chtít dojít do údolí (minimum energie) půjdeme logicky s kopce, tedy proti směru největšího spádu (gradientu). Stejným způsobem tento problém řeší i metoda BP.

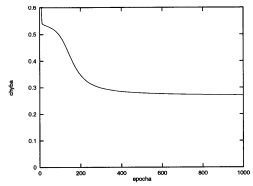
Zpět k problémům

- vysoký stupeň dimenze prostoru vah a prahů,
- složitost a zborcenost energetické plochy,
- tvar plochy závisí na vstupním vektoru, plocha je tedy různá,
- představit si ji prakticky nelze,
- atd.

Typický průběh globální chyby během učení



Dokážete vysvětlit, co se stalo v tomto případě?



Uváznutí v lokálním minimu.

Vyřešíme později.

Cíl učení a metoda k jeho dosažení:

Pro dané hodnoty na vstupech sítě vykazuje energetická funkce určitou hodnotu energie. Naším cílem je tuto hodnotu minimalizovat, a to modifikací váhových koeficientů a prahů. Proto musíme určit o kolik a s jakým znaménkem se změní hodnota energie, když změníme určitou váhu o Δw.

To je možné snadno určit výpočtem z parciální derivace $\partial E/\partial w$. Určujeme tedy citlivost energie (chyby) na změny vah, popř. prahů. Postupujeme vlastně **gradientní metodou**.

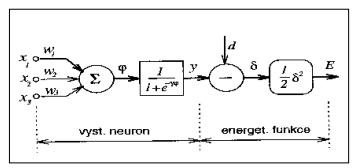
Jak určit gradient energetické funkce?

Vypočítáme-li parciální derivace

$$\partial E \,/\, \partial w_{ij}^{o}$$

pro všechny váhy, dostaneme vektor, který určuje **lokální gradient** energetického prostoru.

Studie: neuron ve výstupní vrstvě



Energetická funkce g je funkcí složenou. Proto můžeme parciální derivaci rozepsat

$$\frac{\partial E}{\partial w_{ij}^{o}} = \frac{\partial E}{\partial y_{i}^{o}} \frac{\partial y_{i}^{o}}{\partial \varphi_{i}} \frac{\partial \varphi_{i}}{\partial w_{ij}^{o}}$$

Jednotlivé položky v předchozím vztahu přitom mají fyzikální význam:

$\frac{\partial E}{\partial y_i^o}$ ukazuje závislost energie na výstupu sítě,

$\frac{\partial y_i^o}{\partial \varphi_i^o}$ je derivací sigmoidy a

$\frac{\partial \varphi_i^o}{\partial w_{ij}^o}$ odráží závislost vnitřního potenciálu na vahách.

Už víte, proč musí mít BP neuron sigmoidu?

- Aby tato derivace existovala, musí být tato nelineární funkce spojitá a diferencovatelná: tedy nesmí to být skoková nelinearita (ale může se k ní libovolně blížit).
- Tu to vlastnost má právě sigmoida.

Takže konkrétně:

1. Derivace energie: $E = \frac{1}{2}\sum_{i=1}^{n}(y_i - d_i)^2$ $\quad \frac{\partial E}{\partial y_i^o} = (y_i^o - d_i)$

2. Derivace sigmoidy $S(\varphi) = \frac{1}{1+e^{-\gamma\varphi}}$ $\quad \frac{\partial y_i^o}{\partial \varphi_i^o} = \gamma y_i^o(1 - y_i^o)$.

3. Derivace vnitř. potenciálu $\varphi_i^o = \sum_{j=1}^{n} w_{ij}^o y_j^h + \Theta_i^o$ $\quad \frac{\partial \varphi_i^o}{\partial w_{ij}^o} = y_j^h$

Shrneme-li tyto dílčí výsledky, dostaneme

$$\frac{\partial E}{\partial w_{ij}^o} = (y_i^o - d_i)\gamma y_i^o(1 - y_i^o) y_j^h = \delta_i^o y_j^h \quad .$$

Symbolem δ označujeme lokální gradient.

Co tento vztah znamená?

- Určuje gradient energetického prostoru, podle kterého budeme upravovat váhy.

Váhy proto můžeme upravovat podle vztahu (η určuje koeficient učení, postupujeme proti gradientu):

$$\Delta \vec{W} = -\eta \nabla E$$

Změny ve výstupní vrstvě:

$$\Delta w_{ij}^o(t) = \eta \delta_i^o(t) y_j^h(t)$$

$$\Delta \Theta_i^o(t) = \eta \delta_i^o(t)$$

kde

$$\delta_i^o(t) = (y_i^o - d_i).S'(\varphi)$$

Váhy a prahy tedy při učení upravujeme podle vztahů

$$w_{ij}^{o}(t+1) = w_{ij}^{o}(t) + \Delta\ w_{ij}^{o}(t)$$

$$\Theta_{i}^{o}(t+1) = \Theta_{i}^{o}(t) + \Delta\ \Theta_{i}^{o}(t)$$

a máme zaručeno, že směřujeme do minima energetické funkce.

Můžeme analogicky postupovat dále?

- Čeho tím dosáhneme?
- Chceme určit,
  - Jak je energetická funkce závislá na vahách (a prazích) skryté a dalších vrstev?
- Ve skutečnosti ale chceme podobně jako výše zjistit, jak upravovat váhy a prahy skrytých a dalších vrstev.

Studie (viz slajd)

Nejedná se vlastně o nic jiného, než o technické rozepsání vztahu pro zjištění závislosti energie na vahách skryté/dalších vrstvy.

Po odvození (provedeme později) dostaneme

$$\Delta w_{ij}^{h}(t) = \eta\ \delta_{i}^{h}(t) x_{j}(t)$$

$$\Delta\Theta_{i}^{h}(t) = \eta\delta_{i}^{h}(t)$$

kde

$$\delta_{i}^{h} = y_{i}^{h}(1 - y_{i}^{h}) \sum_{k=1}^{n} w_{ki}^{o} \delta_{k}^{o}$$

a změny vah a prahů provedeme takto:

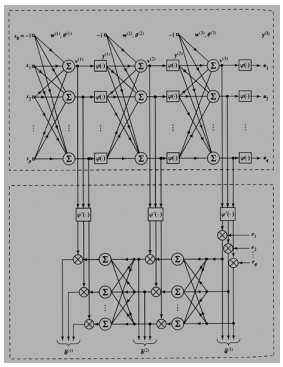
$$w_{ij}^{h}(t+1) = w_{ij}^{h}(t) + \Delta\ w_{ij}^{h}(t)$$

$$\Theta_{i}^{h}(t+1) = \Theta_{i}^{h}(t) + \Delta\ \Theta_{i}^{h}(t)$$

Výpočet δ lze pro celou síť zobecnit následovně:

$$\delta_{i}^{h-1} = y_{i}^{h-1}(1 - y_{i}^{h-1}) \sum_{k=1}^{n} w_{ki}^{h} \delta_{k}^{h}$$

Postup výpočtu nových vah a prahů shrnuje následující obrázek



- Je z něj patrné, jak se chyba z výstupu šíří celou sítí na vstup.
- Odtud název učicího algoritmu - **zpětné šíření**.

Modifikace standardního algoritmu

- Setrvačnost

$$\Delta\ w_{ij}^{o}(t) = \eta\ \ \delta_i^{o}(t)\, y_j^{h}(t) + \alpha\Delta\ \ w_{ij}^{o}(t-1)$$

$$\Delta\ \Theta_i^{o}(t) = \eta\ \ \delta_i^{o}(t) + \alpha\Delta\ \Theta_i^{o}(t-1)$$

Co se v těchto vztazích změnilo?

Modifikace standard. algoritmu II

- Normalizované kumulované delta pravidlo,
  - změny až najednou,
- Delta-bar-delta pravidlo,
  - nejen současný, ale i minulý gradient,
- rozšířené,
  - obojí plus gradient,
- Quickprop, Maxprop,
- rychlé učení.