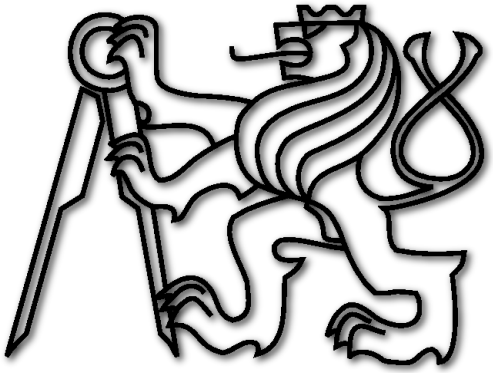
Negative Correlation Learning

Tomáš Siegl

sieglt@fel.cvut.cz

http://cig.felk.cvut.cz



Computational Intelligence Group

Department of Computer Science and Engineering

Faculty of Electrical Engineering

Czech Technical University in Prague

Obsah

- Teoretický rámec ensemblování (Regrese)
  - Bias - variance – covariance,
  - Ambuiguity teorém
  - Diverzita modelů
- Účení s Negativní korelací
  - Zasazení do teoretického rámce
- Klasifikace, budoucí vývoj

Motivační příklad

- Máme neznámou funkci $f = f(x)$, kterou chceme aproximovat

- Trénovací data: $D = \{(x_1,d_1),..,(x_N,d_N)\}$, kde

  $d = f + \varepsilon$, $E\{\varepsilon\} = 0$, střední hodnota ε *(šum)* rovná se nula

- Naučíme libovolný model g aproximovat funkci f

  $y = g(\boldsymbol{x},\boldsymbol{w})$, w – vektor vah

Příklad graficky

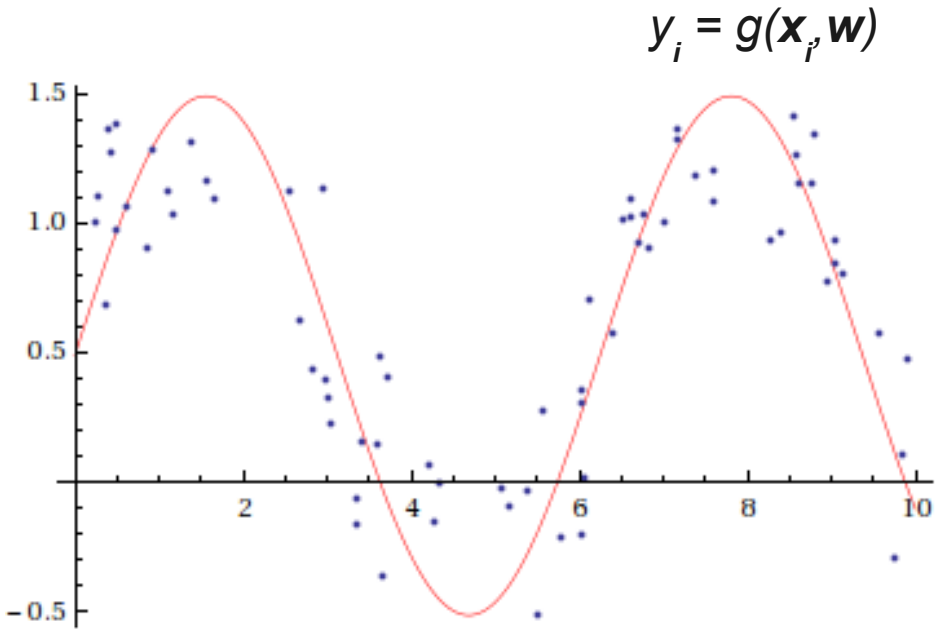
- Střední kvadratická chyba

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

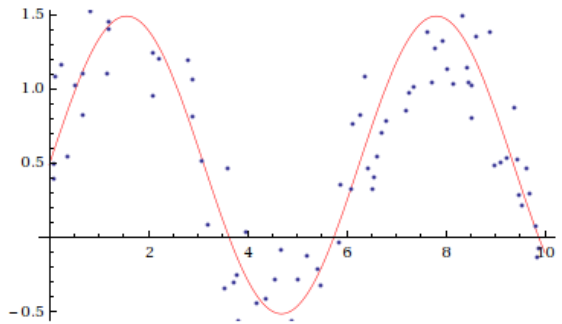
- Trénovací data

  N = 100,

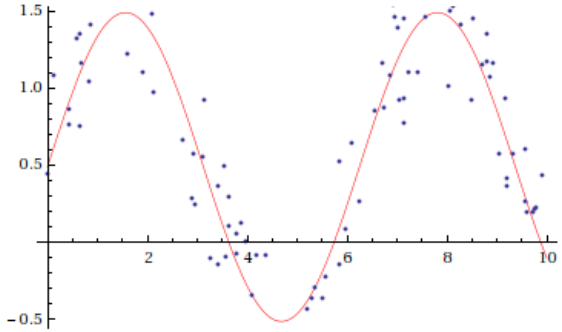
  MSE = 0.0799417



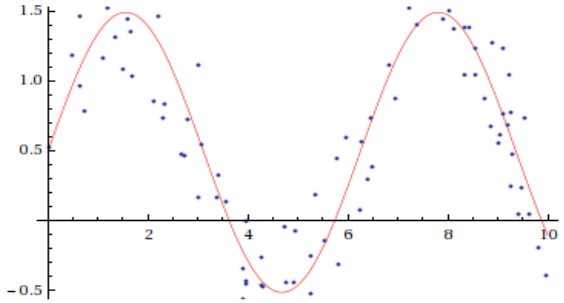
Různé testovací data z *f*



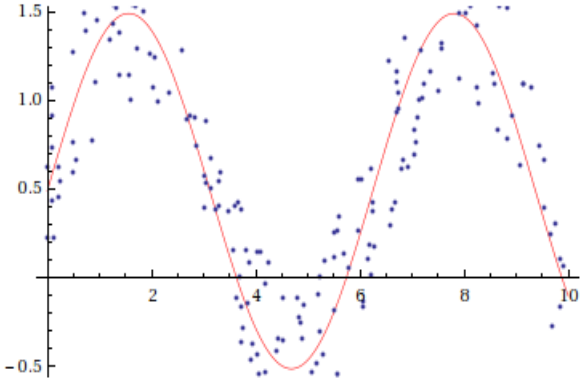
MSE = 0.0935569



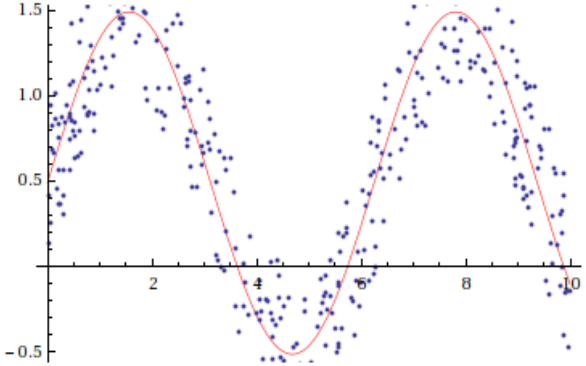
MSE = 0.0789607



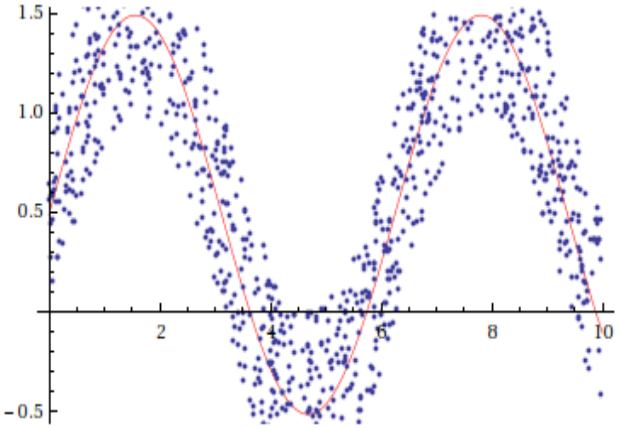
MSE = 0.0781363



MSE = 0.0905944



MSE = 0.0784731



MSE = 0.0852322

Jak se mění MSE

- Zkusíme odvodit jak se bude měnit MSE

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

- Připomenutí ze statistiky

Nechť X je náhodná proměná nabývající hodnoty x_i , i=1..N s rozdělením pravděpobnosti P(X)

Střední hodnota X: $$E\{X\} = \sum_{i=1}^{N} x_i P(x_i) \approx \frac{1}{N}\sum_{i=1}^{N} x_i$$

Variance X: $Var(x) = E\{(X - E\{X\})^2\} = E\{X^2\} - (E\{X\})^2$

Připomenutí ze statistiky

Přepíšeme vztah pro varianci

- $Var(x) = E\{(X - E\{X\})^2\} = E\{X^2\} - (E\{X\})^2$
- $E\{X^2\} = (E\{X\})^2 + Var(x)$

Další vlastnosti:

- $E\{c\} = c$, *když je c konstanta*
- $E\{cX\} = cE\{X\}$
- $E\{X + Y\} = E\{X\} + E\{Y\}$
- $E\{XY\} = E\{X\}\,E\{Y\}$, *pokud X, Y jsou nezávislé náhodné veličiny*

Odvození MSE

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

- *Nejprve odvodíme:*

f(x) – je skutečná funkce (neznáme)

$$E\{(y_i - d_i)^2\} = E\{(y_i - f_i)^2\} + E\{\varepsilon^2\}$$

- *A následně analogickým postupem získáme:*

$$E\{(y_i - f_i)^2\} = E\{(y_i - E\{y_i\})^2\} + E\{(E\{y_i\} - f_i)^2\} = Var(y_i) + Bias(y_i)^2$$

- Dáme vše dohromady

$$E\{MSE\} = Var(y_i) + Bias(y_i)^2 + E\{\varepsilon^2\}$$

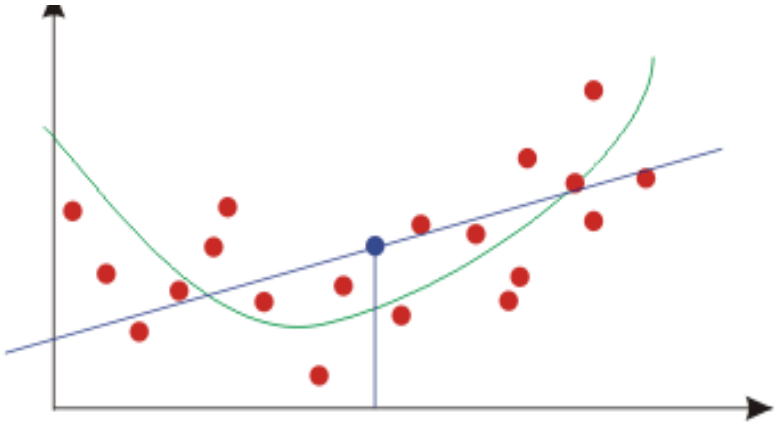
Neredukovatelný šum

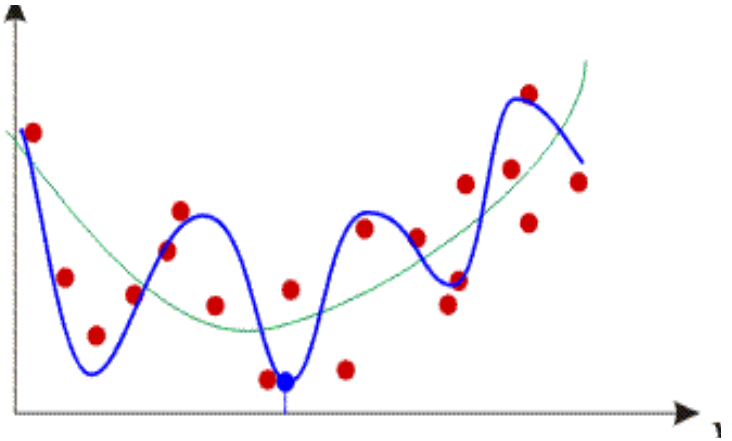
Bias Variance a Šum prakticky

- Hodnotíme učicí algoritmus
  - Bias: $E\{(E\{y_i\} - f_i)^2\}$ určuje jak přesný jsou modely v průměru od skutečné funkce, naučené přes všechny možné trenovací data
  - Variance: $E\{(y_i - E\{y_i\})^2\}$ měří jak jsou modely citlivý na změnu trenovacích dat
- Když chceme průměrné MSE mít minimální, musíme minimalizovat jak bias tak variance

Bias Variance a Šum prakticky

- Bias: $E\{(E\{y_i\} - f_i)^2\}$ Variance: $E\{(y_i - E\{y_i\})^2\}$.
- Pokud Model y = g(x), bude konstantní funkce s hodnotou c, tak Variance bude nulová, ale Bias bude veliký.
- Pokud Model přeučíme, že bude procházet každým bodem skutečné funkce f, tak Bias bude nulový, ale variance bude obrovská

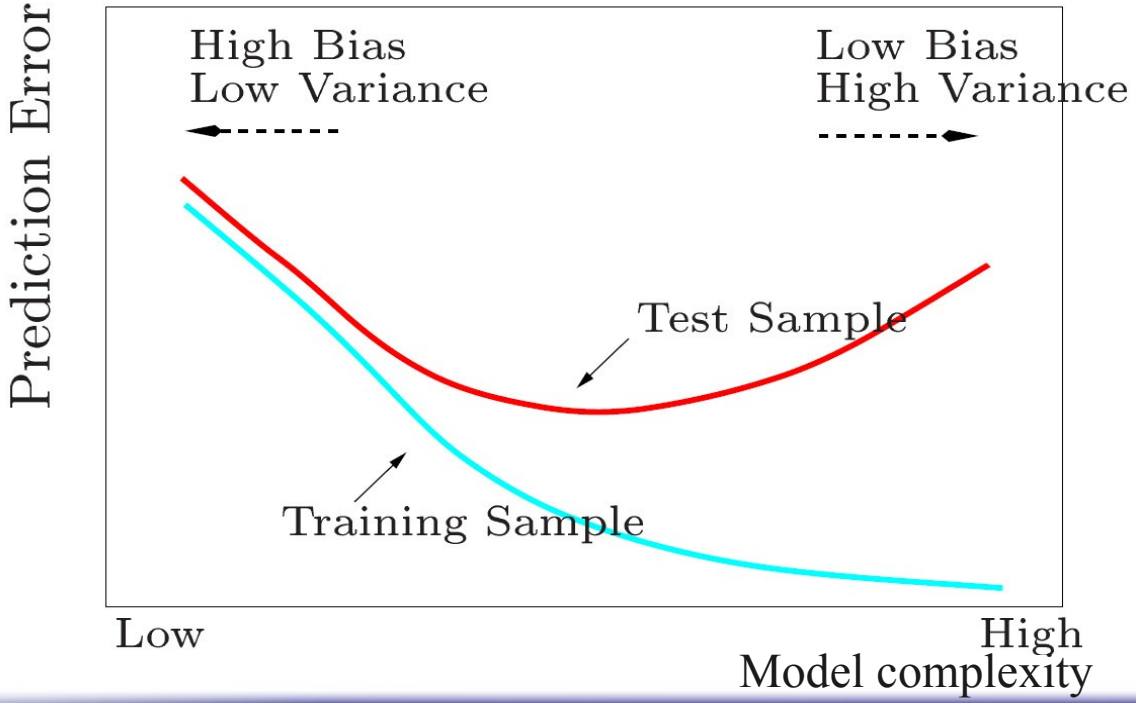




Učení s minimální chybou

- Pokud chceme mít výslednou chybu minimální, musíme při učení vyvážit bias a variance
- To není až tak triviální problém

.



Chyba u Ensemblu

- Mějme skupinu modelů $g_1(x,w_1),..g_m(x,w_m)$
- Každý model naučíme separátně. Chybová funkce bude MSE
- Posléze modely zkombinujeme

$$\bar{g}(x; w_1, .., w_M) = \frac{1}{M} \sum_{i=1}^{M} g_i(x; w_i)$$

- Výsledný ensemble můžeme považovat za samostatný model

Chyba u Ensemblu

- Chyba ensemblu tedy může být také dekomponováná na Bias-Variance složky
- Navíc chyba u ensemblu může být také dekomponovaná:

$$\bar{bias} = \frac{1}{M}\sum_{i=1}^{M}\left(E\{g_i\} - d\right)$$

$$\bar{var} = \frac{1}{M}\sum_{i=1}^{M} E\{(g_i - E\{g_i\})^2\}$$

$$\bar{covar} = \frac{1}{M(M-1)}\sum_{i=1}\sum_{i\neq j}^{M} E\{(g_i - E\{g_i\})(g_j - E\{g_j\})\}$$

Chyba u Ensemblu

- Celková chyba ensemblu

$$E\{(\bar{g}-d)^2\} = \bar{bias}^2 + \frac{1}{M}\bar{var} + \frac{1}{M(M-1)}\bar{covar}$$

- Závisí kromě na biasu a variance individuálních modelů také na kovarianci mezi jednotlivými modely
- **Při učení ensemblu, je tedy výhodné brát tuto kovarianci v úvahu**

Chyba ensemblu také jinak

- Ambiguity decomposition (Krogh, Vedelsby 1995)

$$(g_{ens} - d)^2 = \sum_{i=1}^{M} c_i (g_i - d)^2 - \sum_{i}^{M} c_i (g_i - g_{ens})^2$$

- V libovolném bodě kvadratická chyba ensemblu je **menší nebo rovná** vážené průměrné chybě jeho jednotlivých modelů.
- Druhá část vyjadřuje variabilitu jednotivých modelů, ze kterých je ensemble utvořen. **Vždy >= 0 !**

Optimální chyba ensemblu

- Minimalizace průměrnou chybu jednotlivých modelů
- Maximalizace rozdílu jednotlivých modelů a celkového ensemblu
- Tyto dvě složky musíme vyvážit stejně jako u jediného modelu: bias – variance

Zkombinujeme obě chyby

$$E\{(g_{ens}-d)^2\}=\bar{bias}^2+\frac{1}{M}\bar{var}+\frac{1}{M(M-1)}\bar{covar}$$

$$(g_{ens}-d)^2=\sum_{i=1} c_i(g_i-d)^2-\sum_i c_i(g_i-g_{ens})^2 \qquad c_i = 1/M$$

$$E\{\frac{1}{M}\sum_{i=1}(g_i-d)^2-\frac{1}{M}\sum_i(g_i-g_{ens})\}=\bar{bias}^2+\frac{1}{M}\bar{var}+\frac{1}{M(M-1)}\bar{covar}$$

- A budeme hledat gradient jako u Back propagation, ve výstupní vrstvě.

$$\frac{\partial e_i}{\partial g_i}=\frac{1}{M}[(g_i-d)-K(g_i-g_{ens})]$$

Když K = 0, gradient chybová funkce je úměrný gradientu chyby jednoho modelu

Když K = 1, g_i část se vyruší a ensemble se chová jako jeden model

Negativní korelace

- Ensemblovací metoda, která při učení jednotlivých modelů, bere v úvahu vzájemnou podobnost jednotlivých modelů
- Původně heuristická metoda

$$e_i = \frac{1}{2}(g_i - d)^2 + \lambda\, p_i$$

Penalizační člen

$$p_i = (g_i - d)\sum_{j=1}^{i-1}(g_j - d) \qquad p_i = (g_i - d)\sum_{j \neq i}(g_j - d) \qquad p_i = (g_i - g_{ens})\sum_{j \neq i}(g_j - g_{ens})$$

- Učí jednotlivé modely paralelně, $\lambda \in \langle 0,1 \rangle$

Negativní korelace

- Během učení explicitně vyvažuje jednotlivé části chyby: bias variance kovariance
  - Vyšší hodnota λ způsobuje snižování hodnoty kovariance
- Dá se odvodit, že gradient chyby jednotlivého modelu je uměrný gradientu chyby celého ensemblu

Negativní korelace

- Všimneme si, že penalizační člen, můžeme přepsat, protože $\sum_i (g_i - g_{ens}) = 0$

$$p_i = (g_i - g_{ens})[-(g_i - g_{ens})] = -(g_i - g_{ens})^2$$

- Chybová funkce jednoho modelu

$$e_i = \frac{1}{2}(g_i - d)^2 - \lambda (g_i - g_{ens})^2$$

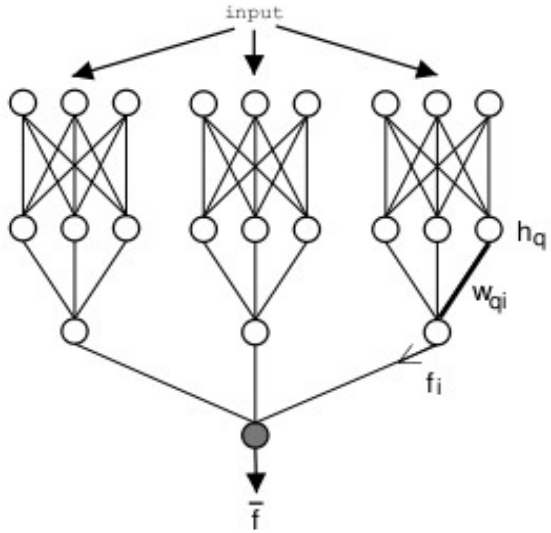
- Porovnáme s ambiguity theorémem

λ = 1/2

$$\frac{1}{2}(g_{ens} - d)^2 = \frac{1}{M}\sum_{i=1}^{M}(\frac{1}{2}(g_i - d)^2) - \sum_{i}^{M}(\frac{1}{2}(g_i - g_{ens})^2)$$

Ensemble jako samostatný model

- Prohledávání stavového prostoru v závislosti na hodnotě λ
- Ensemble jako tři samostatné sítě učené NCL
  - Pro: λ = 0 a λ = 1
- Ensemble jako jedna veliká sít



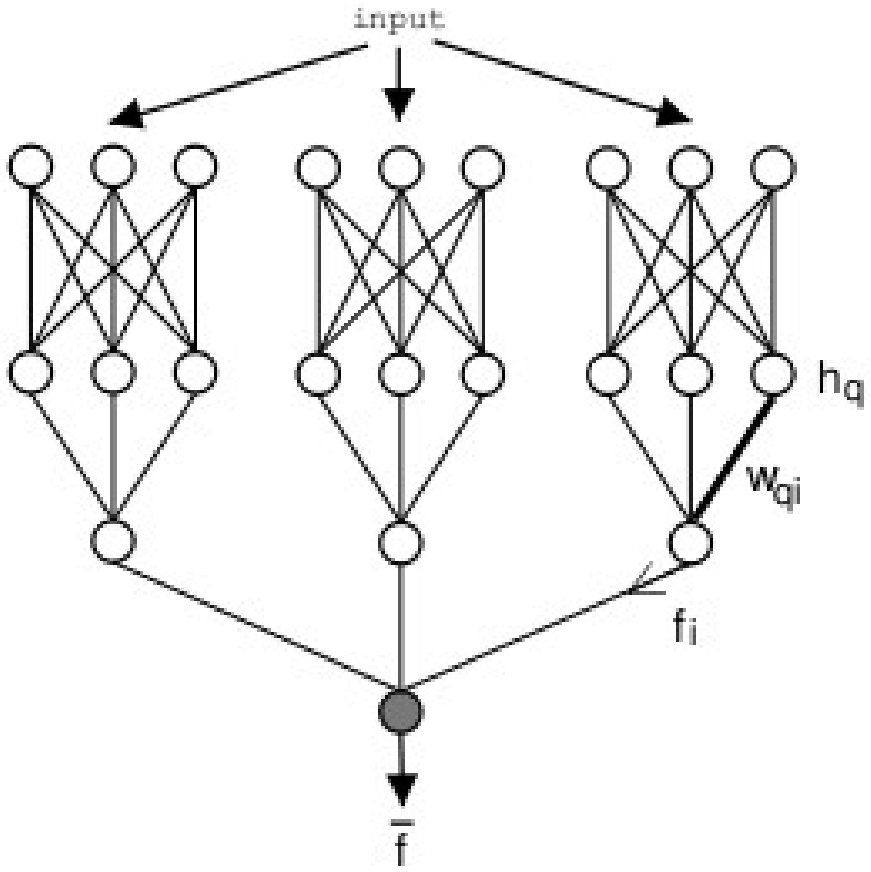
Ensemble jako samostatný model

- 3 samostatné sítě, ensemble průměr
- Aktualizace W_{qi}
  - $\lambda = 0$

$$\frac{\partial e_i}{\partial w_{gi}} = [(f_i - d)][f_i(1 - f_i)][h_q]$$

  - $\lambda = 1$

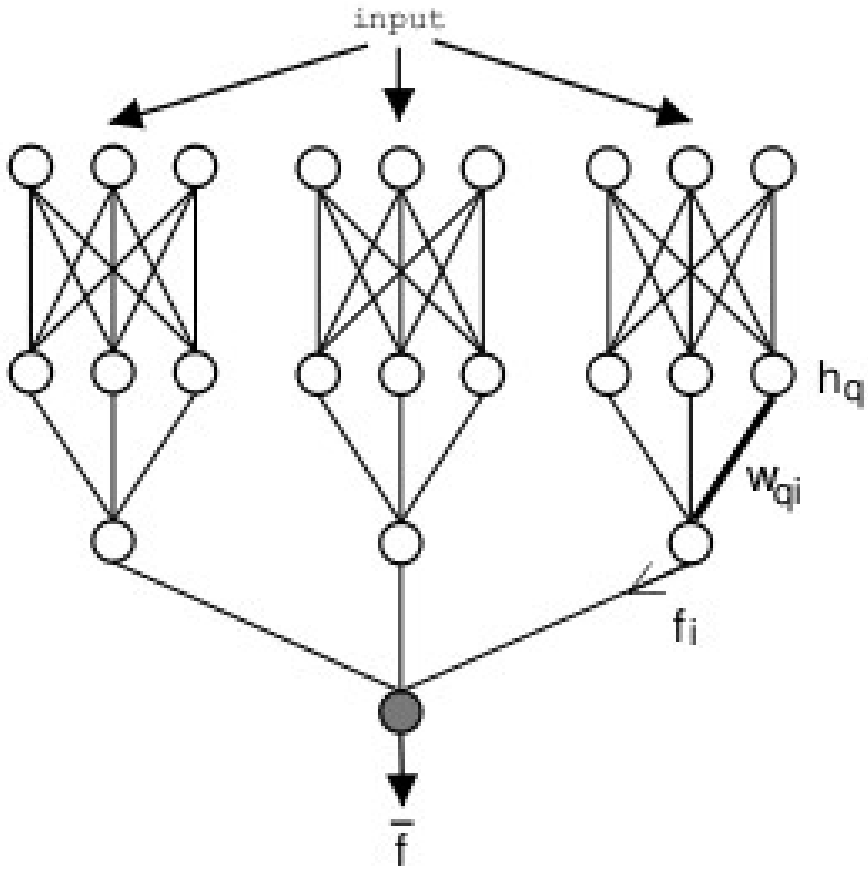
$$\frac{\partial e_i}{\partial w_{gi}} = [(\bar{f}_i - d)][f_i(1 - f_i)][h_q]$$



Ensemble jako samostatný model

- samostatná sít
  - výstup lineární funkce
  - Aktivační funkce
  - $a_i = \frac{1}{M} \sum_{i}^{M} f_i$
- Aktualizace W_{qi}

$$\frac{\partial e}{\partial w_g} = [(\bar{f} - d)][1][\frac{1}{M}][f_i(1 - f_i)][h_q]$$



Porovnání gradientu

- 3 samostatné sítě $\frac{\partial e_i}{\partial w_{gi}} = [(\bar{f}_i - d)][f_i(1-f_i)][h_q]$
- Ensemble jako jedna síť $\frac{\partial e}{\partial w_g} = [(\bar{f} - d)][1][\frac{1}{M}][f_i(1-f_i)][h_q]$
- Minima na stejných místech
- Ale gradientní plocha je M krát mělkční
  - Stejný efekt u BP s malou učící rychlostí
- Škálování λ, škáluje mezi jednou sítí s lineární výstupní vrstvou a paralelním ensemble systémem

Experimentální výsledky

- UCI
  - Friedman

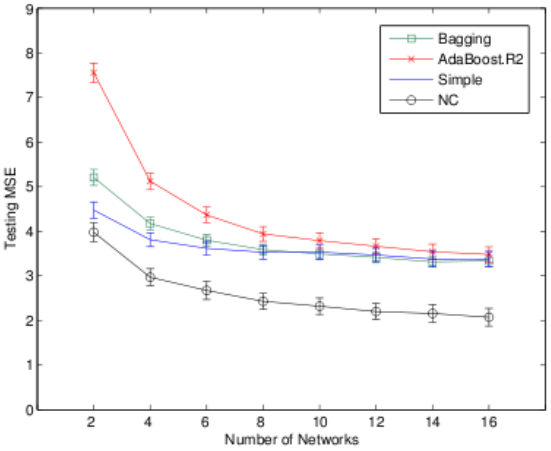


Figure 19: Friedman, varying number of networks (6 hidden nodes in each)

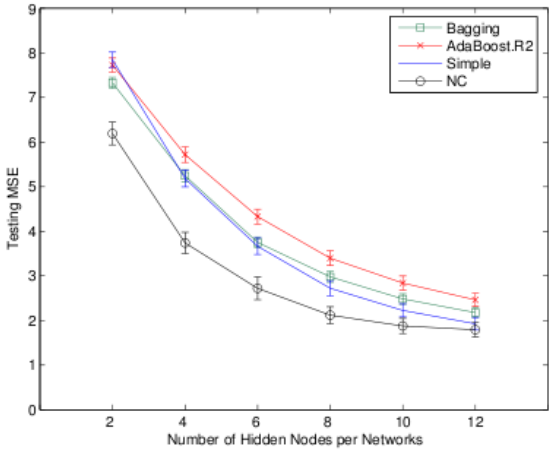
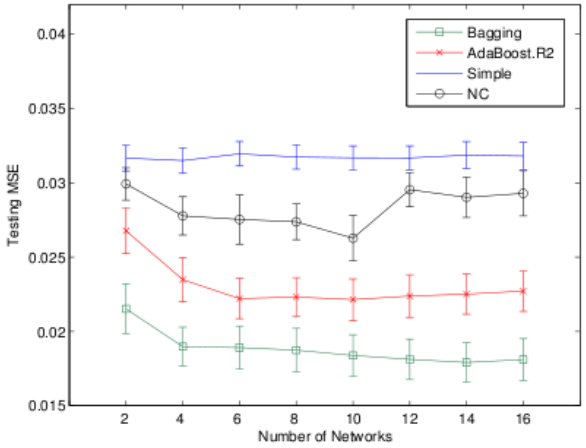
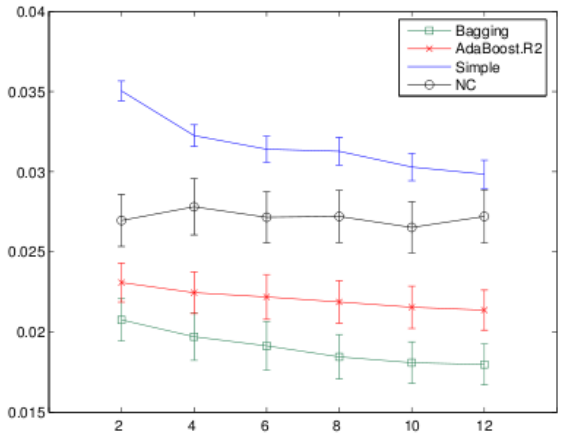


Figure 20: Friedman, varying number of hidden nodes per network (ensemble of 6 networks)

  - Boston





Ensemble klasifikátorů

- Chybová funkce je většinou definovaná na oboru hodnot: <0,1> (Zero – One loss)
- Chen 2009: Diversity and Regularization in Neural Network Ensembles. PhD thesis
  - Ambiguity dekompozice i pro 0 - 1 chybovou funkci.
  - Porovnání koreluje s 10 jinymi mírami pro měření diverzity klasifikátorů

Budoucí výzkum

- Teoreticky prozkoumát navržený ambiquity teorém pro zero – one loss funkci
  - Prozkoumat dolní a horní mez λ v závislosti na počtu modelů v ensemblu
- Najít „míru“, která bude během učení korelovat s variancí ensemblu
  - Eliminace náročného počítání přes všechny možné treninkové množiny
  - Explicitní kontrola nad Variancí i kovariancí během učení ensemblu

Závěr

Děkuji za pozornost