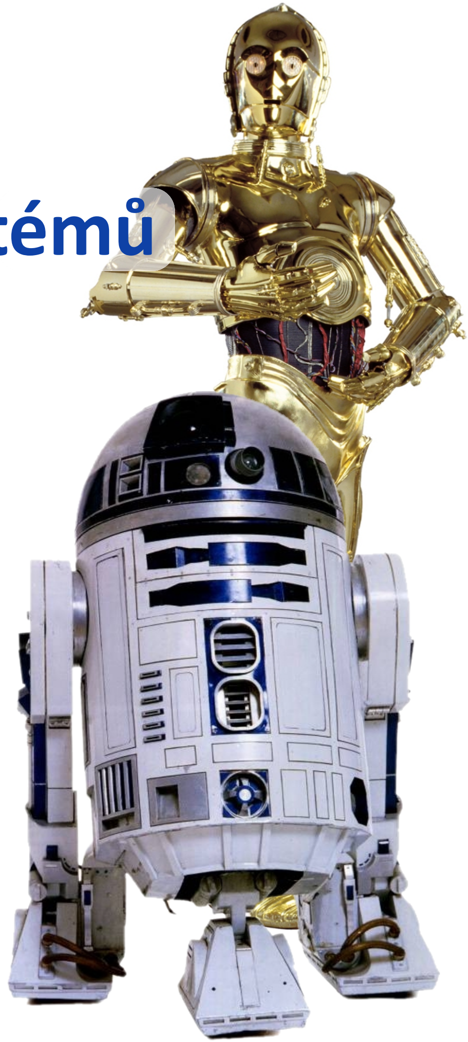
Teorie kognitivních systémů

6 Regularizace



- Přeučení – jeho projevy
- Techniky potlačení přeučení
- Regularizace – definice
- Naivní odvození regularizace
- Algoritmus regularizace
- Regularizovaná LR a LogR

Regularizace
Definice

Regularizace (*Regularization*) – matem. postup zavedení **další informace** do výpočtu, aby bylo možné řešit **špatně podmíněnou** nebo **přeurčenou úlohu**.

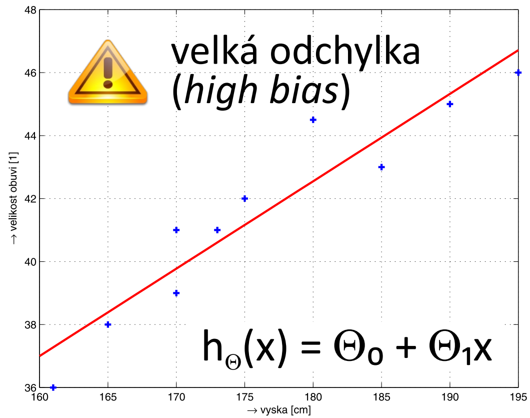
Přidaná informace je nejčastěji:

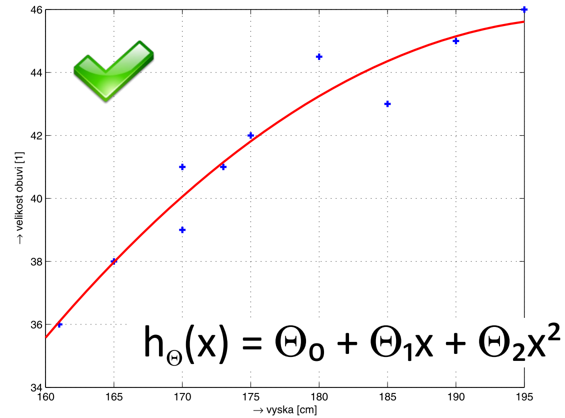
- penalizace složitosti (např. omezením hladkosti nebo normy vektorového prostoru)
- výběr parametrů modelu z nějaké distribuce, tj. jejich apriorní odhad

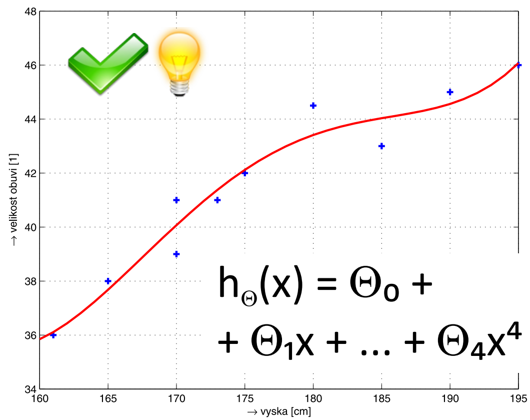
Regularizace se využívá při výběru modelu: Provádí se implicitní nebo explicitní **penalizace modelu podle počtu jeho parametrů**.

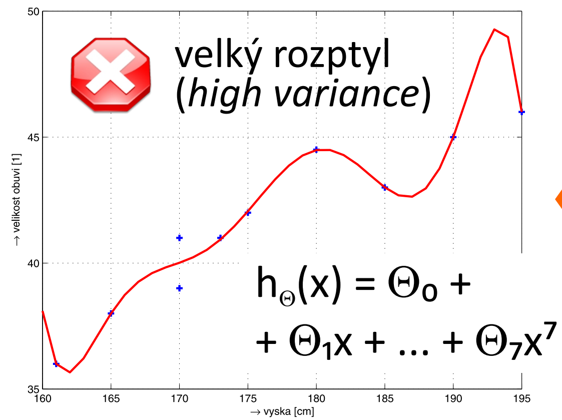
Problém: Přeučení (*Overfitting*)
Lineární regrese velikosti obuvi

UNDER-FITTING



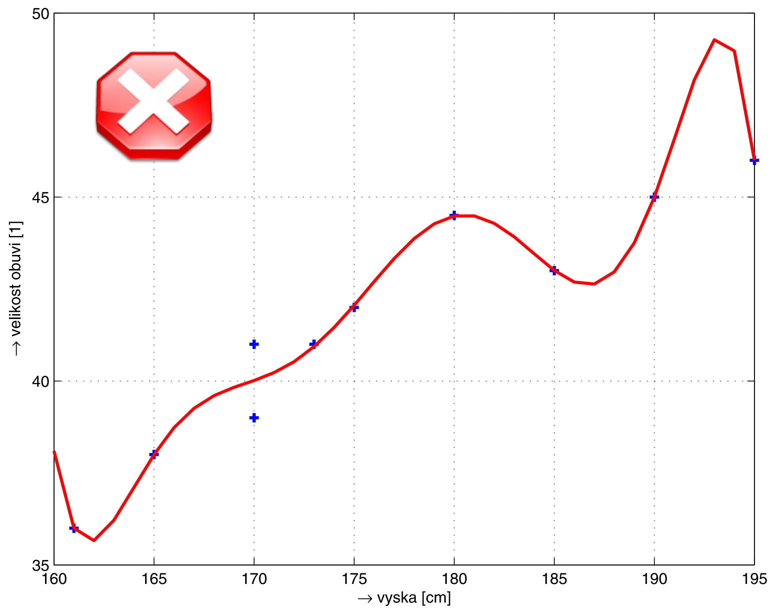






OVER-FITTING

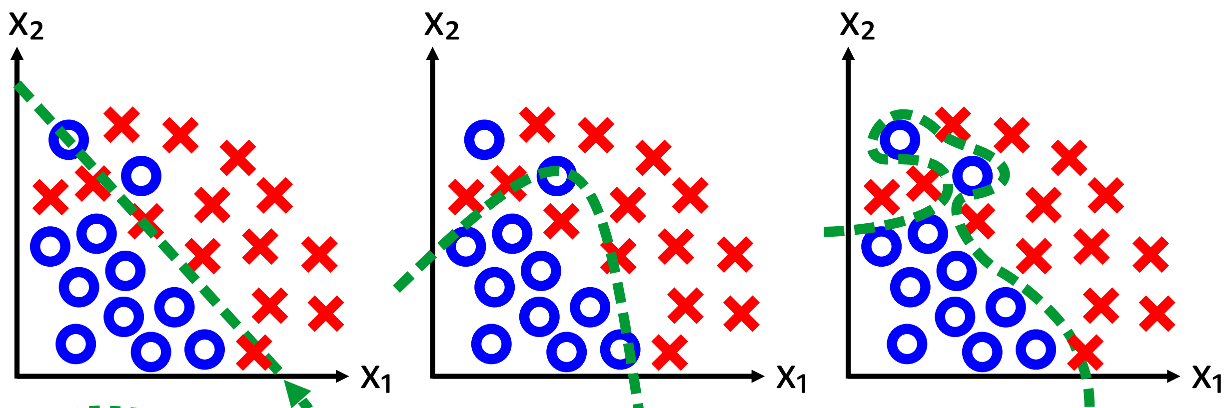
Přeučení (*Overfitting*)
Popis problému



$h_\Theta(x) = \Theta_0 + \Theta_1 x + \ldots + \Theta_7 x^7$

Přeučení – máme-li příliš složitý model, natrénovaná hypotéza dokonale prochází body z trénovací množiny (tj. $J(\Theta) \to 0$), ale neposkytuje dobrou aproximaci modelované úlohy: selhává při **generalizaci**, tj. predikci modelované veličiny v zadaném bodě.

Problém: Přeučení (*Overfitting*)
Logistická regrese



$h_\Theta(x) = P(\Theta_0 + \Theta_1 x_1 + \Theta_2 x_2)$

$h_\Theta(x) = P(\Theta_0 + \Theta_1 x_1 + \Theta_2 x_2 + \Theta_3 x_1^2 + \Theta_4 x_2^2 + \Theta_5 x_1 x_2)$

$h_\Theta(x)$ = P(<*polynom vyššího řádu*>)

Tento model zjevně **nemá žádnou generalizační schopnost**, přestože na trénovací množině $J(\Theta) = 0$...

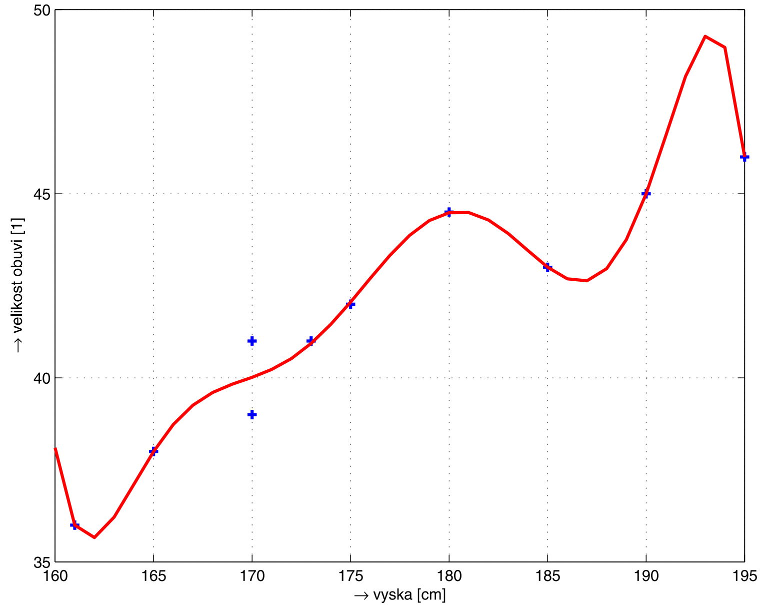
Potlačení přeučení
Techniky aplikovatelné v případě přeučení

- křížová validace (*Cross-validation*)
- včasné zastavení (*Early Stopping*)
- prořezávání (*Pruning*)
- apriorní modelování distribuce parametrů (*Bayesian Priors*)
- porovnávání modelů na ověřovací množině
- **regularizace** –
  aplikace **Occamovy břitvy** (*Occam's Razor / Lex parsimoniae*) na výsledné hypotézy: „Máme-li stejně dobré konkurenční hypotézy, vyberme tu, která vyžaduje nejméně premis".

Toho lze dosáhnout porovnáním konkurenčních hypotéz (modelů) na ověřovací množině nebo penalizací „zesložiťování" modelu...

Potlačení přeučení
Naivní přístup

Když průběh funkce hypotézy $h_\Theta(\mathbf{x})$ **vykreslíme**, lze si utvořit rámcovou představu o jejím generalizačním potenciálu. Je-li průběh funkce příliš komplikovaný, snížíme počet parametrů → **to ale není vždy možné**...



Pokud např. $\mathbf{x} \in \mathbf{R}^{100}$, pak je vizualizace prakticky nemožná.

Potlačení přeučení
Možnosti

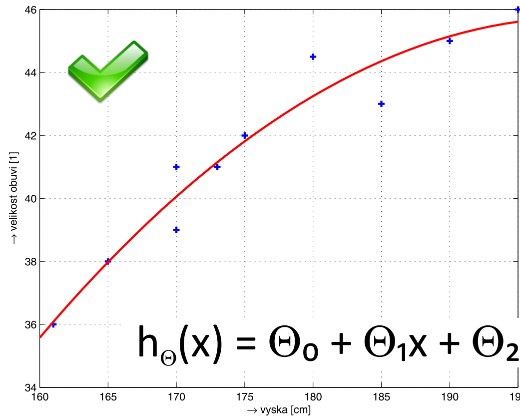
(i) **snížení počtu příznaků** (a tím parametrů modelu, protože každá složka příznakového vektoru je váhována parametrem Θ_j)

- ruční výběr příznaků, které se **zdají být důležité**, a proto je rozumné je ponechat (ostatní vyhodit)
- aplikace **algoritmu výběru modelu** (později)

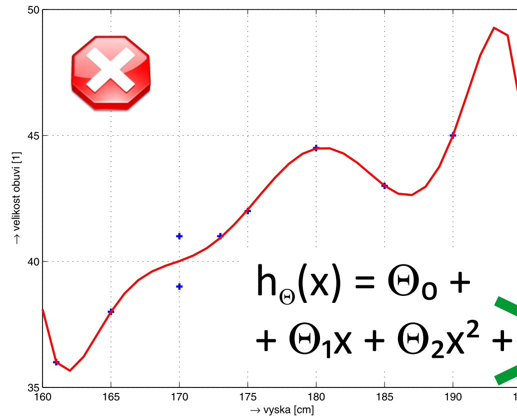
(ii) **regularizace**

- ponecháme všechny příznaky, ale snížíme jejich váhu – parametry Θ_j budou malá čísla tak, aby křivka hypoté- příliš neoscilovala
- technika funguje dobře v případě, že máme mnoho příznaků, z nichž každý „trochu" přispívá k predikci výsledné hodnoty

Regularizace
Intuitivní odvození



$h_\Theta(x) = \Theta_0 + \Theta_1 x + \Theta_2 x^2$



$h_\Theta(x) = \Theta_0 + \Theta_1 x + \Theta_2 x^2 + \cancel{\dots + \Theta_7 x^7}$

Nastavíme „pokuty” tak, abychom trénovací algorimus přinutili zvolit hodnoty parametrů Θ_3 až Θ_7 velmi malé...

$$\min_{\Theta} \frac{1}{2m} \sum_{i=1}^{m} \left(h_\Theta\left(\mathbf{x}^{(i)}\right) - y^{(i)} \right)^2 + \underbrace{\boxed{1000 \cdot \left(\Theta_3^2 + \ldots + \Theta_7^2\right)}}_{\Theta_1 \ldots \Theta_7 \to 0}$$

Základní myšlenka regularizace
Regularizovaná verze cenové funkce

Požadujeme **malé hodnoty parametrů** $\Theta_0, \Theta_1, \ldots, \Theta_n$:
- „jednodušší" (hladší) průběh hypotézy
- odolnější vůči přeučení

Vezměme opět příkad s predikcí velikost obuvi:
Příznaky: $\mathbf{x} = [x_1, x_2, \ldots, x_n]^T$
Parametry modelu: $\Theta_0, \Theta_1, \ldots, \Theta_n$

nevíme, které potlačit, tak potlačíme všechny „trochu"

Upravíme cenovou funkci tak, aby penalizovala všechny parametry modelu, tj. přidáme **regularizační člen** (*Reg. Term*):

$$J(\Theta) = \frac{1}{2m}\left[\sum_{i=1}^{m}\left(h_\Theta(\mathbf{x}^{(i)}) - y^{(i)}\right)^2 + \boxed{\lambda \sum_{j=1}^{n} \Theta_j^2}\right]$$

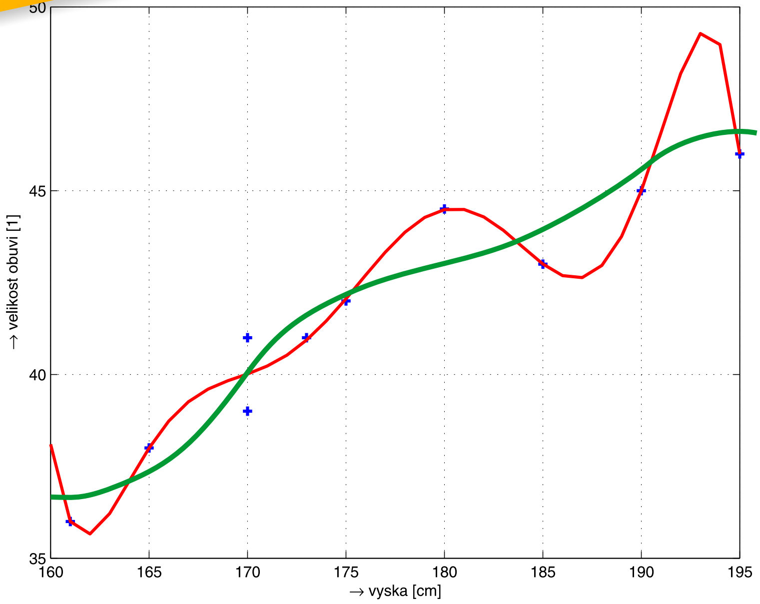
Regularizovaná cenová funkce
Důsledky v projektivním prostoru

$$J(\Theta) = \frac{1}{2m}\left[\sum_{i=1}^{m}\left(h_\Theta\left(\mathbf{x}^{(i)}\right) - y^{(i)}\right)^2 + \lambda \sum_{j=1}^{n}\Theta_j^2\right]$$

regularizační parametr

Θ_0 se nepenalizuje!

Minimalizace modifikované J(Θ) vede k hladšímu průchodu křivky hypotézy trénovacími daty...



Regularizovaná lineární regrese
Nastavení parametru λ

Regularizovaná lineární regrese – vybíráme Θ_j tak, aby minimalizovaly cenovou funkci ve tvaru

$$J(\Theta) = \frac{1}{2m}\left[\sum_{i=1}^{m}\left(h_\Theta\left(\mathbf{x}^{(i)}\right) - y^{(i)}\right)^2 + \lambda\sum_{j=1}^{n}\Theta_j^2\right]$$

Co když zvolíme λ příliš velké, např. $\lambda = 10^{10}$?

- algoritmus stále funguje – nastavení λ na libovolnou hodnotu nemůže způsobit havárii
- ovšem nedochází k potlačení přeučení
- výsledkem je „nedoučení” (*Underfitting*) – hypotéza nedokáže spolehlivě predikovat ani trénovací data
- gradientní sestup **nekonverguje**

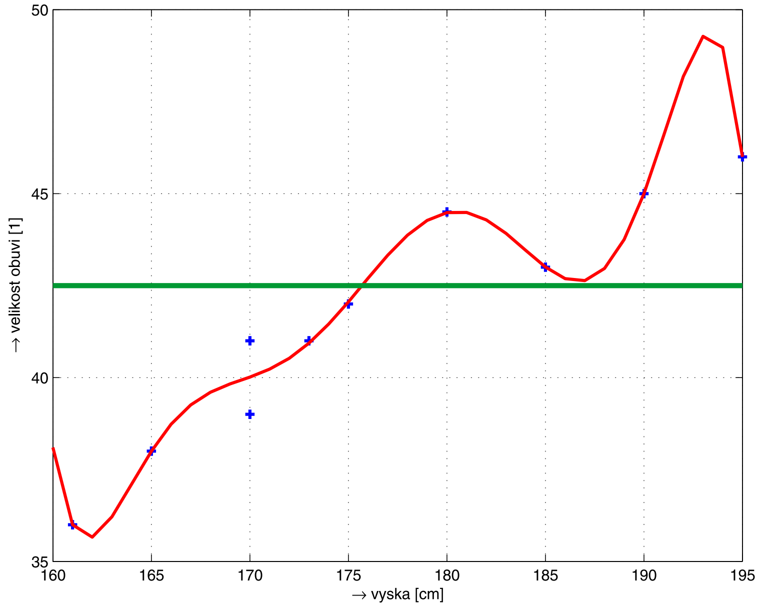
Regularizovaná lineární regrese
Příliš velká hodnota parametru λ

$$h_\Theta(x) = \Theta_0 + \underbrace{\Theta_1 x + \Theta_2 x^2 + \Theta_3 x^3 + \ldots + \Theta_n x^n}$$

pro velké λ se členy $\Theta_j \rightarrow 0$, tj.
hypotéza se redukuje na $h_\Theta(x) = \Theta_0$

Model je pak zřejmě „nedoučený” (*Under-fitted*), tzn. má velkou systematickou chybu.

$h_\Theta(x) = \Theta_0$



Regularizovaná lineární regrese
Modifikace rekurzivního sestupu

```
while not converged() do {
```

$$\Theta_0 \leftarrow \Theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\Theta\left(\mathbf{x}^{(i)}\right) - y^{(i)} \right) \cdot x_0^{(i)}$$

```
  for j = 1 .. n do
```

$$\Theta_j \leftarrow \Theta_j - \alpha \overbrace{\left[\frac{1}{m} \sum_{i=1}^{m} \left(h_\Theta\left(\mathbf{x}^{(i)}\right) - y^{(i)} \right) \cdot x_j^{(i)} + \frac{\lambda}{m} \Theta_j \right]}^{\frac{\partial J(\Theta)}{\partial \Theta_j}}$$

```
}
```

$$\Theta_j \leftarrow \Theta_j \overbrace{\left(1 - \alpha \frac{\lambda}{m}\right)}^{< 1,\ \approx 0.99} - \overbrace{\alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\Theta\left(\mathbf{x}^{(i)}\right) - y^{(i)} \right) \cdot x_j^{(i)}}^{\text{pův. verze GS (bez regularizace)}}$$

Regularizovaná lineární regrese
Modifikace normální rovnice

TM **příznaky pozorovaných vzorků** **odpovědi učitele**

$$\mathbf{x}^{(i)} = \begin{bmatrix} x_0^{(i)} \\ x_1^{(i)} \\ x_2^{(i)} \\ \vdots \\ x_n^{(i)} \end{bmatrix} \in \mathbb{R}^{n+1}, i \in \langle 1, m \rangle \qquad \mathbf{y} = \begin{bmatrix} y^{(i)} \\ y^{(i)} \\ y^{(i)} \\ \vdots \\ y^{(i)} \end{bmatrix} \in \mathbb{R}, i \in \langle 1, m \rangle$$

$$\mathbf{\Theta} = \left(\mathbf{X}^{\mathsf{T}}\mathbf{X} + \lambda \begin{bmatrix} 0 & & & & \\ & 1 & & & \\ & & 1 & & \\ & & & \ddots & \\ & & & & 1 \end{bmatrix}_{(n+1) \times (n+1)} \right)^{-1} \mathbf{X}^{\mathsf{T}} y$$

Regularizovaná lineární regrese
Singulární případ normální rovnice

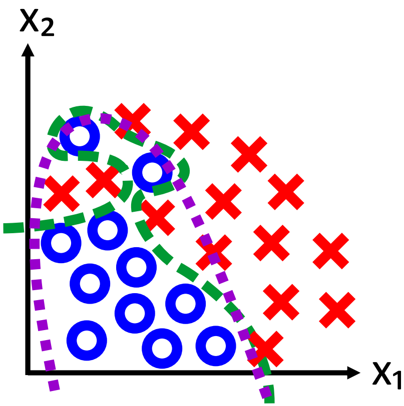
Je-li méně trénovacích vzorků než příznaků, tj. $m < n$, pak ne-regularizovaná normální rovnice **nelze spočítat**, jelikož výraz $(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}$ není invertibilní, je **singulární**.

Pseudoinverze (v MATLABu pinv) sice proběhne bez chyb, ale výsledky jsou nesmyslné → **regularizace**:

vždy invertibilní

$$\boldsymbol{\Theta} = \left(\boxed{\mathbf{X}^{\mathsf{T}}\mathbf{X} + \lambda \underbrace{\begin{bmatrix} 0 & & & & \\ & 1 & & & \\ & & 1 & & \\ & & & \ddots & \\ & & & & 1 \end{bmatrix}}_{(n+1)\times(n+1)}} \right)^{-1} \mathbf{X}^{\mathsf{T}} y$$

Regularizovaná logistická regrese
Modifikace rekurzivního sestupu



Omezením hodnoty parametrů Θ regularizací dosáhneme hladšího průběhu rozhodovací hranice...

$$h_\Theta = P(\Theta_0 + \Theta_1 x_1 + \Theta_2 x_1^2 + \Theta_3 x_1^2 x_2 + \Theta_4 x_1^2 x_2^2 + \Theta_5 x_1^2 x_2^3 + \ldots)$$

Upravená podoba cenové funkce LogR s regularizací:

$$J(\mathbf{\Theta}) = -\left[\frac{1}{m}\sum_{i=1}^{m} y^{(i)} \log h_\Theta\left(\mathbf{x}^{(i)}\right) + \left(1 - y^{(i)}\right) \log\left(1 - h_\Theta\left(\mathbf{x}^{(i)}\right)\right)\right] + \boxed{\frac{\lambda}{2m}\sum_{j=1}^{n} \Theta_j^2}$$

Regularizovaná logistická regrese
Modifikace rekurzivního sestupu

```
while not converged() do {
```

$$\Theta_0 \leftarrow \Theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} \Big(h_\Theta\big(\mathbf{x}^{(i)}\big) - y^{(i)} \Big) \cdot x_0^{(i)}$$

```
  for j = 1 .. n do
```

$$\Theta_j \leftarrow \Theta_j - \alpha \overbrace{\boxed{\left[\frac{1}{m} \sum_{i=1}^{m} \Big(h_\Theta\big(\mathbf{x}^{(i)}\big) - y^{(i)} \Big) \cdot x_j^{(i)} + \frac{\lambda}{m} \Theta_j \right]}}^{\frac{\partial J(\Theta)}{\partial \Theta_j}}$$

```
}
```

Algoritmus vypadá stejně jako u lineární regrese, ovšem liší se v použitém tvaru hypotézy:

$$\boxed{h_\Theta(\mathbf{x}) = \frac{1}{1 + e^{-\mathbf{\Theta}^\mathsf{T}\mathbf{x}}}}$$

Pokročilá optimalizace
pomocí MATLABu/Octave

Využijeme algoritmy pokročilé optimalizace dostupné v knihovnách či MATLABu/Octave: Naprogramujeme regularizovanou cenovou funkce a tu minimalizujeme...

`fminunc(@costfunc, 0, opts);`

`function [J, grad] = costfunc(theta)`

$$J = -\left[\frac{1}{m}\sum_{i=1}^{m} y^{(i)} \log h_\Theta\left(\mathbf{x}^{(i)}\right) + \left(1-y^{(i)}\right) \log\left(1-h_\Theta\left(\mathbf{x}^{(i)}\right)\right)\right] + \frac{\lambda}{2m}\sum_{j=1}^{n}\Theta_j^2$$

`grad = zeros(n + 1);`

$$\mathrm{grad}(1) = \frac{1}{m}\sum_{i=1}^{m}\left(h_\Theta\left(\mathbf{x}^{(i)}\right) - y^{(i)}\right) \cdot x_0^{(i)}$$

`for i = 2 : n + 1`

$$\mathrm{grad}(i) = \frac{1}{m}\sum_{i=1}^{m}\left(h_\Theta\left(\mathbf{x}^{(i)}\right) - y^{(i)}\right) \cdot x_j^{(i)} + \frac{\lambda}{m}\Theta_j$$

`end`

`end`