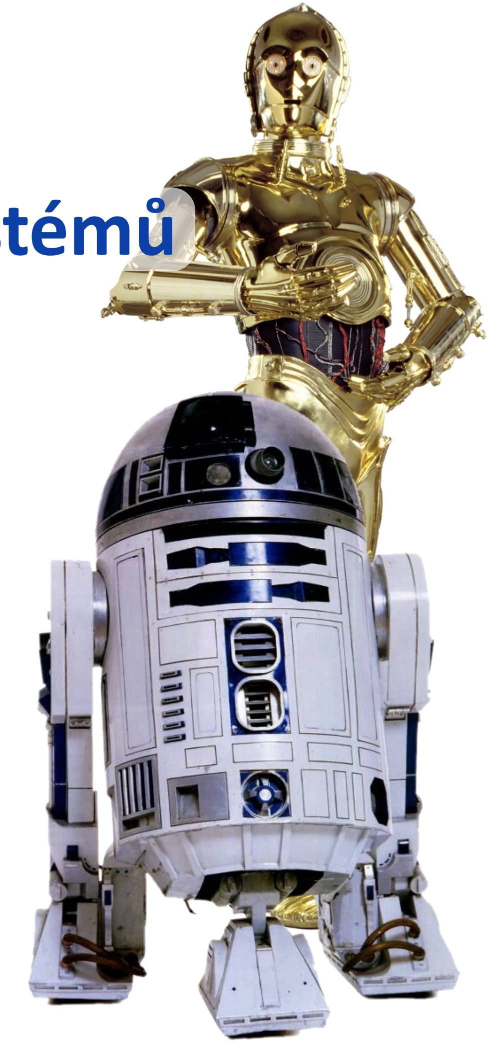
Teorie kognitivních systémů

4 Logistická regrese



- Logistická regrese
- Model hypotézy pro LogR
- Interpretace výsledků
- Rozhodovací hranice
- Cenová funkce pro LogR
- Klasifikace do více tříd
- Algoritmus One vs All

Logistická regrese
Regrese pro predikci diskrétních hodnot

- predikovaná proměnná *y* = $h_\Theta(\mathbf{x})$ nabývá diskrétních hodnot: buď 0 nebo 1 – **binomická** či **binární logistická regrese**, nebo 0 až K – **multinomická logistická regrese** (MNL) nebo také logistická regrese s vícekategoriální vysvětlovanou pro-měnnou, tj. jedná se o **klasifikační úlohu**

- predikovaná proměnná (čili modelovaná veličina) má tzv. **alternativní rozdělení p-sti** (*Bernoulli Distribution*) – jde o variantu binomického rozdělení *Bi(n, p)*, kde počet pokusů *n* = 1 a *p* je p-st úspěchu (čili proměnná nabyde hodnoty 1)

- **jeden z nejpopulárnějších a nejrozšířenějších učících se algoritmů**, řada dalších technik z oblasti UI a strojového učení je na ní založena

Logistická regrese
Příklady klasifikačních úloh

Binární logistická regrese –

$y = h_\Theta(\mathbf{x}) \in \{0; 1\}$

0: **negativní třída**

1: **pozitivní třída**

- **E-mail**: spam / ne-spam
- **Lékařská diagnostika nádorů**: zhoubný / nezhoubný
- **Bezpečnost (IS)**: autentizovaný / neautentizovaný přístup

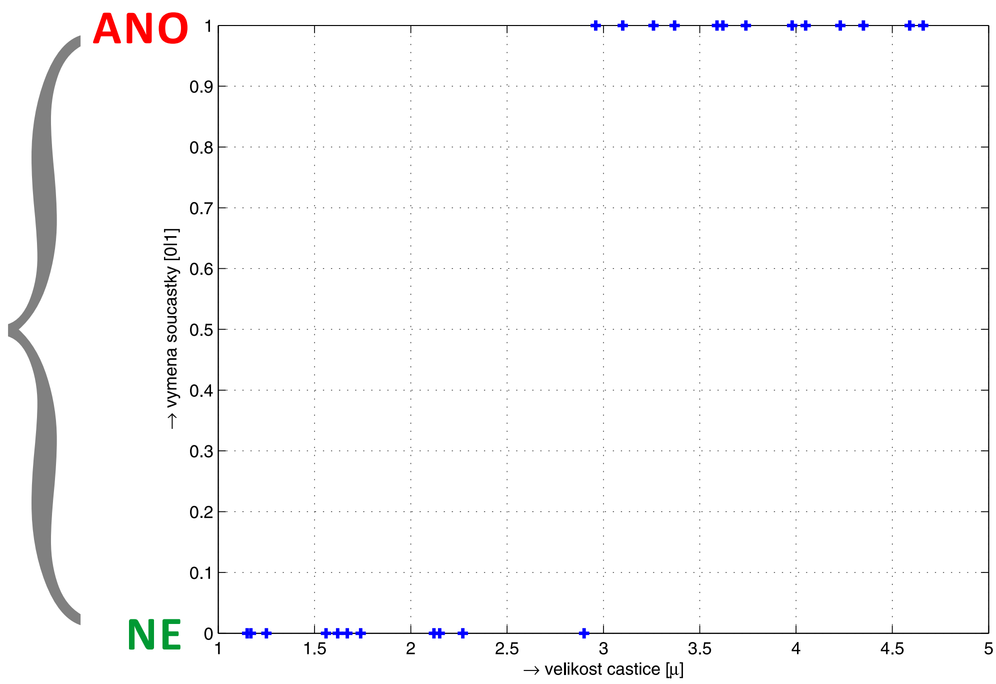
***n*-ární (multinomická) lineární regrese** –

$y = h_\Theta(\mathbf{x}) \in \mathbf{N} \cup \{0\}$

- **Lékařská diagnostika**: angína / chřipka / rýmička / kašel / ...
  např. existující systém DXplain (http://dxplain.org/dxp/dxp.pl)

Logistická regrese
Motivační příklad – tribodiagnostika

Vyměnit součástku?



Co se stane při aplikaci již známého aparátu **lineární regrese**?

Zkusme proložit data přímkou, tj. hypotéza $h_\Theta(x) = \Theta_0 + \Theta_1 x$

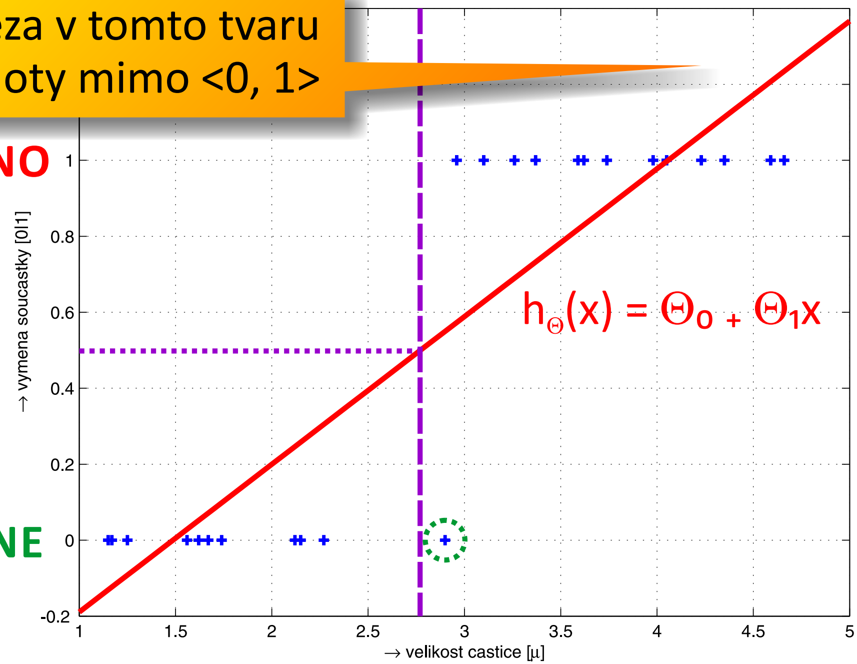
Logistická regrese
Motivační příklad – tribodiagnostika

Problém: Hypotéza v tomto tvaru předpovídá hodnoty mimo <0, 1>

Vyměnit součástku?

ANO

NE



Klasifikaci proveďme prahováním hypotézy $h_\Theta(x)$ hodnotou 0,5:

pro $h_\Theta(x) \geq 0.5$: $y \leftarrow 1$; pro $h_\Theta(x) < 0.5$: $y \leftarrow 0$

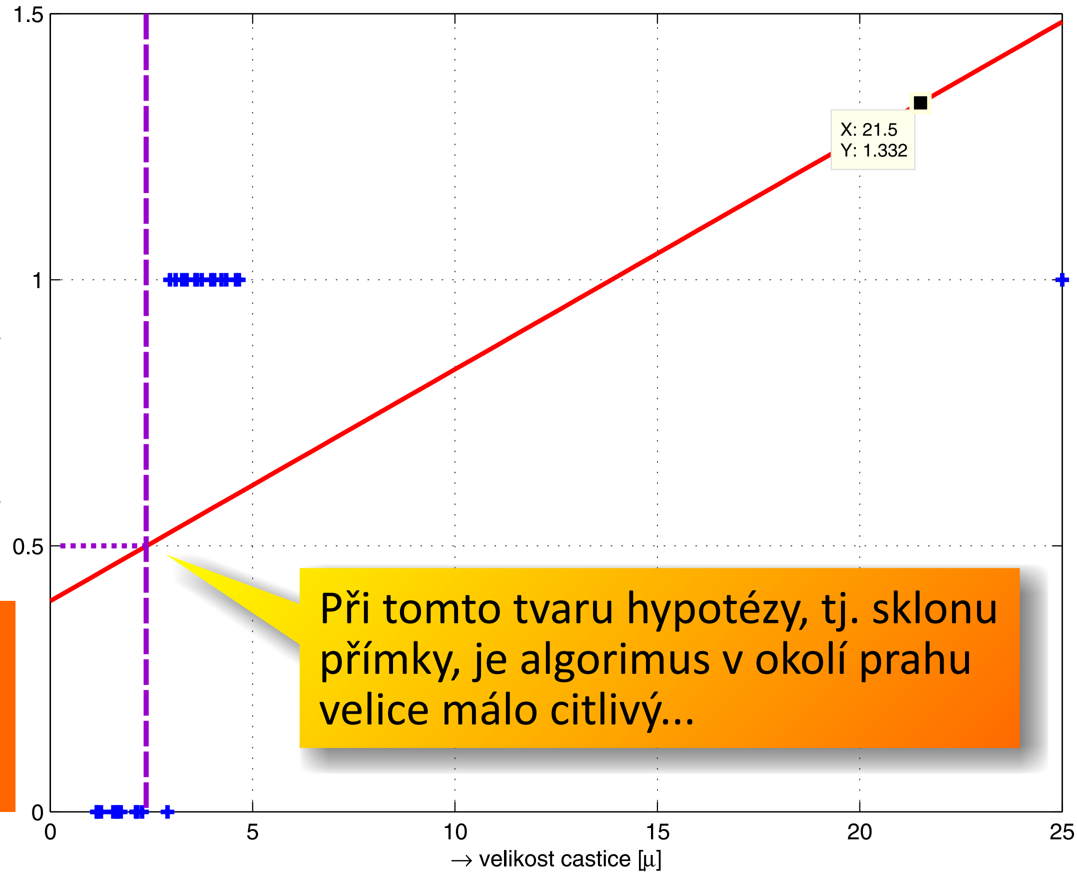
Logistická regrese
Motivační příklad – tribodiagnostika

Můžeme mít štěstí a bude to nějak fungovat, ale mnohem častěji to **fungovat nebude**, a to zejména proto, že:

- byť jsou data (odpovědi učitele) v trénovací množině pouze z množiny {0; 1}, algoritmus předpovídá spojité hodnoty na intervalu $(-\infty; \infty)$ → nevíme, co dělat s hodnotami mimo interval <0; 1>...
- pro hodnoty $h_\Theta(x) \in$ <0; 1> stanovíme práh na hodnotě 0.5 – to ale nemusí být správná hodnota prahu, prahem určená oddělující nadplocha neseparuje data z trénovací množiny správně podle pokynů učitele
- vyskytnou-li se v datech extrémní marginální hodnoty, pak takto navržený mechanismus klasifikace selže úplně (viz následující obr.)

Logistická regrese

Motivační příklad – tribodiagnostika



PROBLÉM

Při tomto tvaru hypotézy, tj. sklonu přímky, je algorimus v okolí prahu velice málo citlivý...

Logistická regrese
Vlastnosti

Klasifikační úloha: $y \in \{0; 1\}$

Lineární regrese: $y = h_\Theta(x) \in (-\infty; \infty)$
- omezení hypotézy na interval <0; 1> je problematické → není jednoznačné, nelze algoritmem LR určit/natrénovat, je velmi citlivé na extrémní marginální hodnoty v trénovací množině, atd.

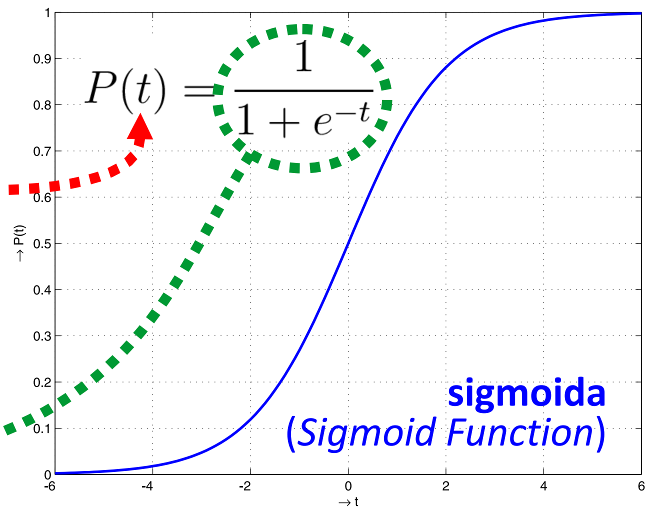
Logistická regrese: $y = h_\Theta(x) \in \langle 0; 1 \rangle$
- ač se to z historických důvodů nazývá regrese, jedná se o **klasifikaci** (protože hypotéza logistické regrese predikuje z daného vstupu **x** pravděpodobnost, že y = 1)

Logistická regrese
Reprezentace hypotézy

Požadujeme $0 \leq h_{\Theta}(\mathbf{x}) \leq 1$,

tj. $h_{\Theta}(\mathbf{x}) = P(\Theta^{T}\mathbf{x})$

reálné číslo



Hypotéza má tedy tvar:

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

– **sigmoida** je speciálním případem tzv. **logistické funkce** (*Logistic Function*) pro $a = 1$, $m = 0$, $n = 1$, $\tau = 1$

$$f(t; a, m, n, \tau) = a\frac{1 + me^{-t/\tau}}{1 + ne^{-t/\tau}}$$

Logistická regrese
Interpretace výsledků hypotézy

$h_\Theta(\mathbf{x})$ – odhad **pravděpodobnosti**, že modelovaná veličina nabyde hodnoty 1 (tj. např. jev nastane), je-li vstupem vektor příznaků **x**

Příklad: Při tribodiagnostice byla naměřena průměrná velikost částic kovu v oleji 3.27 μ – tzn. vstupem hypotézy bude

$$\mathbf{x} = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} = \begin{bmatrix} 1 \\ 3.27 \end{bmatrix}$$

nutné, aby to šlo zapsat maticově jako soustavu

$$h_\Theta(\mathbf{x}) = \frac{1}{1 + e^{-(\Theta_0 + \Theta_1 \cdot 3.27)}} = 0.79$$

Hypotéza říká, že při velikosti částic 3.27 μ je **s pravděpodobností 79% potřeba ložisko vyměnit**.

Logistická regrese
Statistická interpretace výsledků hypotézy

$h_\Theta(\mathbf{x})$ – pravděpodobnost, že y = 1 za podmínky **x**, parametrizovaná Θ:

$$h_\Theta(\mathbf{x}) = P(y = 1 \mid \mathbf{x}; \Theta)$$

Z toho plynou některé zřejmé závěry:

$P(y = 0 \mid \mathbf{x}; \Theta) + P(y = 1 \mid \mathbf{x}; \Theta) = 1$, z čehož

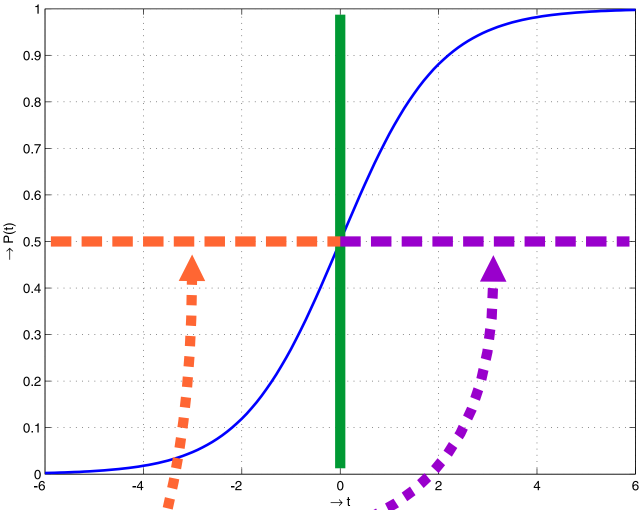
$P(y = 0 \mid \mathbf{x}; \Theta) = 1 - P(y = 1 \mid \mathbf{x}; \Theta)$

Tzn. v předchozím příkladu odhad pravděpodobnosti, že ložisko není třeba měnit, je 1 - 0.79 = 0.21, tj. 21%

Logistická regrese
Chování modelu

Hypotéza $h_\Theta(\mathbf{x}) = P(\Theta^T\mathbf{x})$, kde

$$P(t) = \frac{1}{1+e^{-t}}$$



Model predikuje:

„y = 1” $\Leftrightarrow h_\Theta(\mathbf{x}) \geq 0.5$, tj.

$P(t) \geq 0.5$

$\Theta^T\mathbf{x} \geq 0$

„y = 0” $\Leftrightarrow h_\Theta(\mathbf{x}) < 0.5$, tj.

$P(t) < 0.5$

$\Theta^T\mathbf{x} < 0$

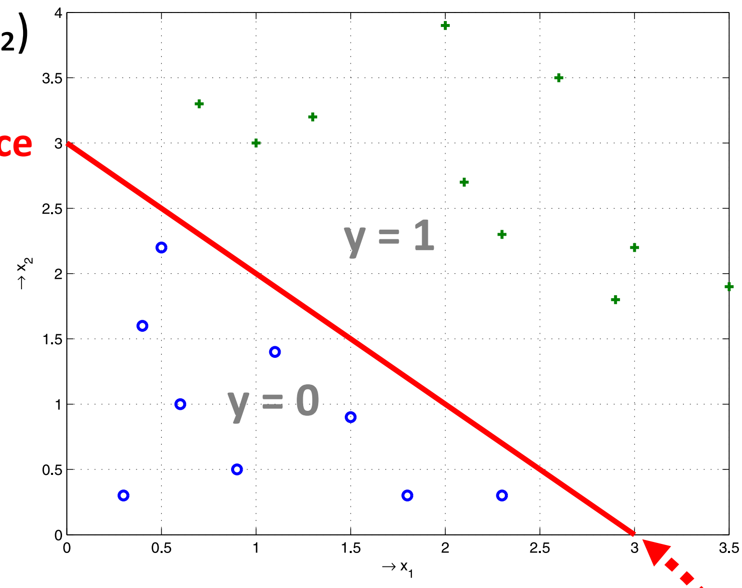
Rozhodovací hranice (*Decision Boundary*)
Chování LR v obrazovém prostoru

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

rozhodovací hranice

Zvolme $\Theta = \begin{bmatrix} -3 \\ 1 \\ 1 \end{bmatrix}$

(zatím neřešíme, jak
Θ „nastavit”)



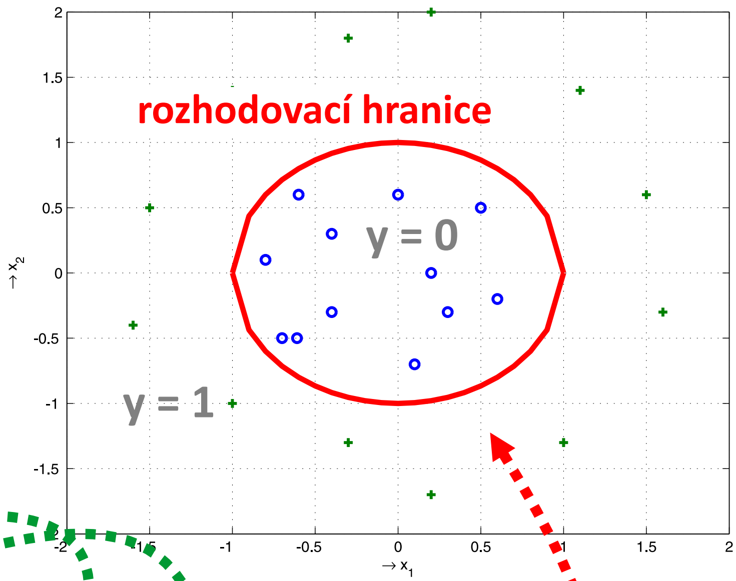
Model predikuje „y = 1” $\Leftrightarrow \underbrace{\boxed{-3 + x_1 + x_2}}_{\Theta^T\mathbf{x}} \geq 0$, čili když $x_1 + x_2 \geq 3$.

Naopak když $x_1 + x_2 < 3$, model predikuje „y = 0”.

Pokud právě $x_1 + x_2 = 3 \Leftrightarrow h_\Theta(\mathbf{x}) = 0.5$

Rozhodovací hranice
Nelineární rozhodovací hranice

Volbou tvaru a parametrů hypotézy můžeme vytvářet libovolně složité rozhodovací hranice:



Zvolme $\Theta = \begin{bmatrix} -1 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}$

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

Model predikuje „y = 1” $\Leftrightarrow -1 + x_1^2 + x_2^2 \geq 0$

$x_1^2 + x_2^2 \geq 1$

(ne)rovnice kružnice

Cenová funkce logistické regrese
Odvození – výchozí podmínky

Trénovací množina o *m* vzorcích

$$T = \{(\mathbf{x}^{(1)}, y^{(1)}); (\mathbf{x}^{(2)}, y^{(2)}); \ldots; (\mathbf{x}^{(m)}, y^{(m)})\}$$

$$\mathbf{x}^{(i)} = \begin{bmatrix} x_0^{(i)} \\ x_1^{(i)} \\ \vdots \\ x_n^{(i)} \end{bmatrix} \in \mathbb{R}^{n+1}, \forall i : x_0^{(i)} = 1; y \in \{0, 1\}$$

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

Problém: Jak nastavit parametry Θ, aby algoritmus logistické regrese klasifikoval neznámé vzorky s nejmenší chybou?

Cenová funkce logistické regrese
Odvození

Lineární regrese – minimalizujeme cenovou funkci ve tvaru:

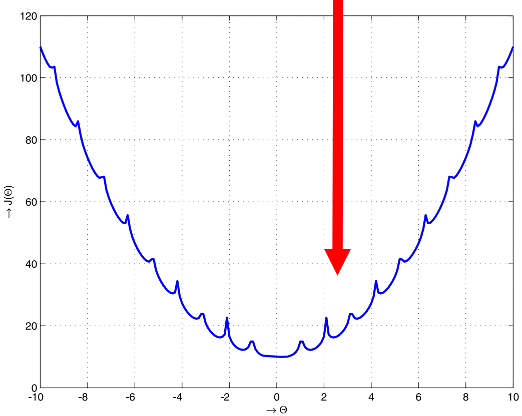
$$J(\Theta) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}\left(h_\Theta\left(\mathbf{x}^{(i)}\right) - y^{(i)}\right)^2$$

Logistická regrese – zavedeme funkci Cost:

$$J(\Theta) = \frac{1}{m}\sum_{i=1}^{m}\mathrm{Cost}\left(h_\Theta\left(\mathbf{x}^{(i)}\right), y^{(i)}\right)$$

$$\mathrm{Cost}\left(h_\Theta(\mathbf{x}), y\right) = \frac{1}{2}\left(h_\Theta(\mathbf{x}) - y\right)^2$$

$$\frac{1}{1+e^{-\Theta^\mathsf{T}\mathbf{x}}}$$

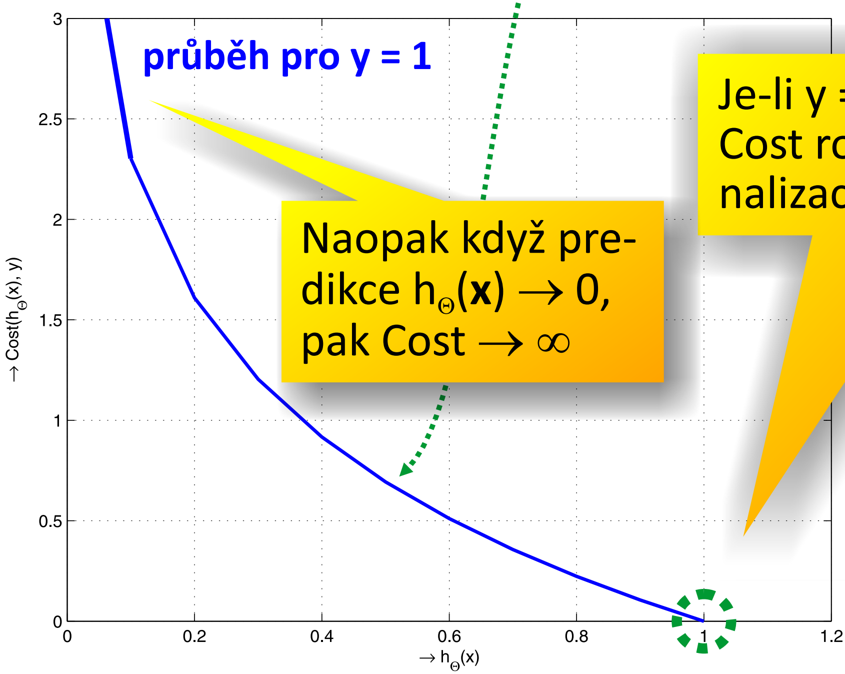


nekonvexní funkce – GS nenalezne glob. optimum

Je třeba nadefinovat cenovou funkci jinak, aby GS fungoval...

Cenová funkce logistické regrese
Odvození

$$\mathrm{Cost}\big(h_\Theta(\mathbf{x}), y\big) = \begin{cases} \boxed{-\log\big(h_\Theta(\mathbf{x})\big)} & \text{když } y = 1 \\ -\log\big(1 - h_\Theta(\mathbf{x})\big) & \text{když } y = 0 \end{cases}$$



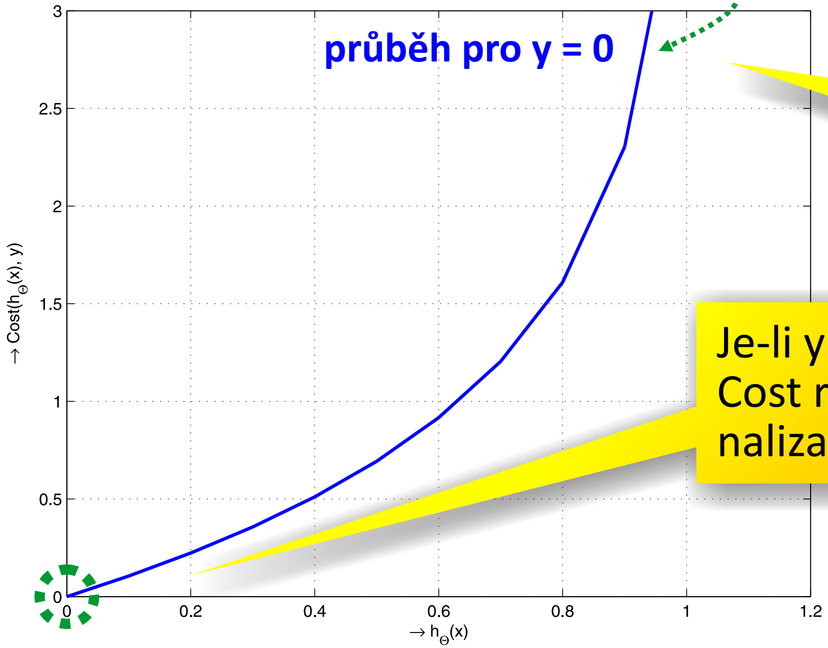
Naopak když predikce $h_\Theta(\mathbf{x}) \to 0$, pak Cost $\to \infty$

Je-li y = 1 a $h_\Theta(\mathbf{x}) = 1$, pak je Cost rovno nule (žádná penalizace)...

max. penalizace v případě, že $h_\Theta(\mathbf{x}) = 0$, tj. $P(y = 1 | \mathbf{x}; \Theta) = 0$, ale přitom y = 1...

Cenová funkce logistické regrese
Odvození

$$\mathrm{Cost}\left(h_\Theta(\mathbf{x}), y\right) = \begin{cases} -\log\left(h_\Theta(\mathbf{x})\right) & \text{když } y = 1 \\ \boxed{-\log\left(1 - h_\Theta(\mathbf{x})\right)} & \text{když } y = 0 \end{cases}$$



Když predikce $h_\Theta(\mathbf{x}) \to 0$, pak Cost $\to \infty$

Je-li y = 0 a $h_\Theta(\mathbf{x}) = 0$, pak je Cost rovno nule (žádná penalizace)...

Cenová funkce logistické regrese
Zjednodušený zápis vhodný pro GS

Cenová funkce $$J(\Theta) = \frac{1}{m}\sum_{i=1}^{m} \mathrm{Cost}\left(h_\Theta\left(\mathbf{x}^{(i)}\right), y^{(i)}\right)$$

$$\mathrm{Cost}\left(h_\Theta(\mathbf{x}), y\right) = \begin{cases} -\log\left(h_\Theta(\mathbf{x})\right) & \text{když } y = 1 \\ -\log\left(1 - h_\Theta(\mathbf{x})\right) & \text{když } y = 0 \end{cases}$$

Přičemž vždy $y \in \{0; 1\}$ – to je významný poznatek, umožňující zjednodušit zápis cenové funkce:

$$\boxed{\mathrm{Cost}\left(h_\Theta(\mathbf{x}), y\right) = -y\log\left(h_\Theta(\mathbf{x})\right) - (1-y)\log\left(1 - h_\Theta(\mathbf{x})\right)}$$

Pro y = 1: $\mathrm{Cost}\left(h_\Theta(\mathbf{x}), y\right) = -\log\left(h_\Theta(\mathbf{x})\right)$

Pro y = 0: $\mathrm{Cost}\left(h_\Theta(\mathbf{x}), y\right) = -\log\left(1 - h_\Theta(\mathbf{x})\right)$

Cenová funkce logistické regrese
Zjednodušený zápis vhodný pro GS

Cenová funkce $$J(\Theta) = \frac{1}{m}\sum^{m}\mathrm{Cost}\Big(h_\Theta\big(\mathbf{x}^{(i)}\big), y^{(i)}\Big) =$$

$$= -\frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)}\log\big(h_\Theta(\mathbf{x}^{(i)})\big) + (1-y^{(i)})\log\big(1-h_\Theta(\mathbf{x}^{(i)})\big)\right]$$

Tvar cenové funkce vychází z poznatků statistiky, z principů techniky zvané **metoda maximální věrohodnosti** (*Maximum Likelihood Estimation, MLE*):

- je **konvexní**

Odhad neznámých veličin z pozorovaných dat:
1. formulace pravděpodobnostního modelu,
2. ověření shody modelu s realitou.

Hledáme parametry Θ, tzn. **minimalizujeme** výše uvedenou cenovou funkci J(Θ)...

Gradientní sestup
v případě logistické regrese

Potřebujeme minimalizovat cenovou funkci J(Θ), abychom našli optimální nastavení parametrů Θ → aplikujeme GS:

```
while not converged() do {
    for j = 0 .. n do
```

$x_0^{(i)} = 1$

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

```
}
```

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

Algoritmus vypadá stejně (a taky **stejný je**) jako v případě lineární regrese, až na tvar hypotézy $h_\Theta(\mathbf{x})$...

Pokročilejší optimalizační algoritmy pro výpočet parametrů logistické regrese

GS je tzv. „baseline” – základní, nejméně sofistikovaná podoba optimalizačního algoritmu (viz přednáška Lineární regrese – asymptotická míra konvergence).

Ovšem implementujeme-li GS, musíme naprogramovat výpočet **cenové funkce** J(Θ) – pro potřeby sledování konvergence, a její **parciální derivace** – pro potřeby výpočtu kroku GS...

Toto lze využít při aplikaci pokročilejších algoritmů:

- **Konjugovaný GS** (*Conjugate GD*)
- **BFGS** (*Broyden-Fletcher-Goldfarb-Shanno*)
- **L-BFGS** (*Limited-memory BFGS*)

značně kompliko-vané

Výhody: mnohem rychlejší konvergence, není třeba volit α

Pokročilejší optimalizační algoritmy
Implementace

Není příliš rozumné implementovat pokročilejší optimalizační algoritmy (CGD, BFGS, L-BFGS) vlastními silami – existuje řada **knihoven** (JProGraM, netlib, SciPy, HANSO, ...)

V **MATLAB**u/**Octave**:

- funkce FMINUNC (Function Minimisation Unconstrained)
  – „nadopovaný” gradientní sestup

UKÁZKA

Klasifikace pomocí LogR do více tříd
Algoritmus „One vs All”

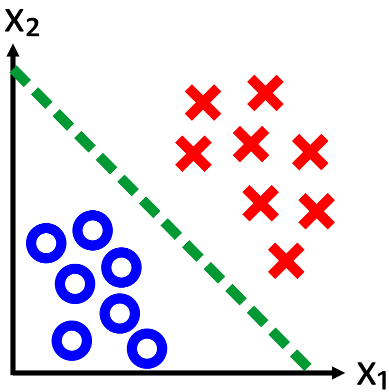
Klasifikace do více tříd (*Multiclass Classification*) –

Predikce počasí: jasno, polojasno, zataženo, déšť, sněžení

Učitel: y = 1 y = 2 y = 3 y = 4 y = 5

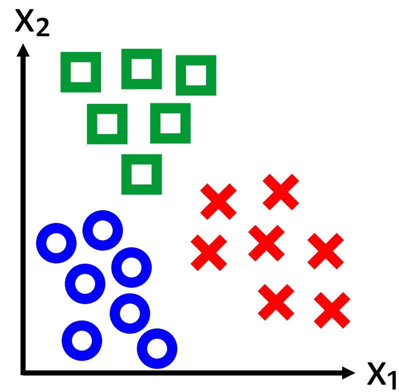
Lékařská diagnóza: zdráv(a), rýma, kašel, zácpa, smrt

Klasifikace článků: zprávy, kultura, sport, věda, politika

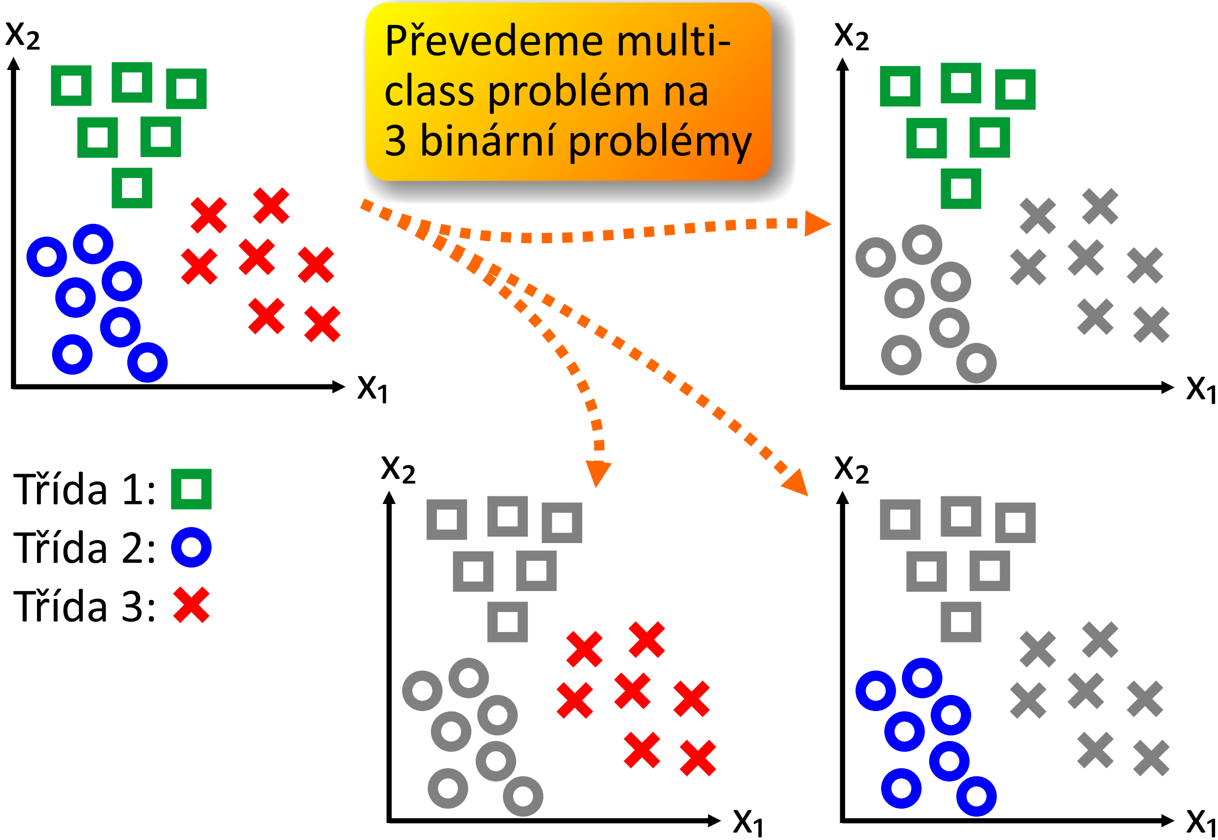


binární
umíme

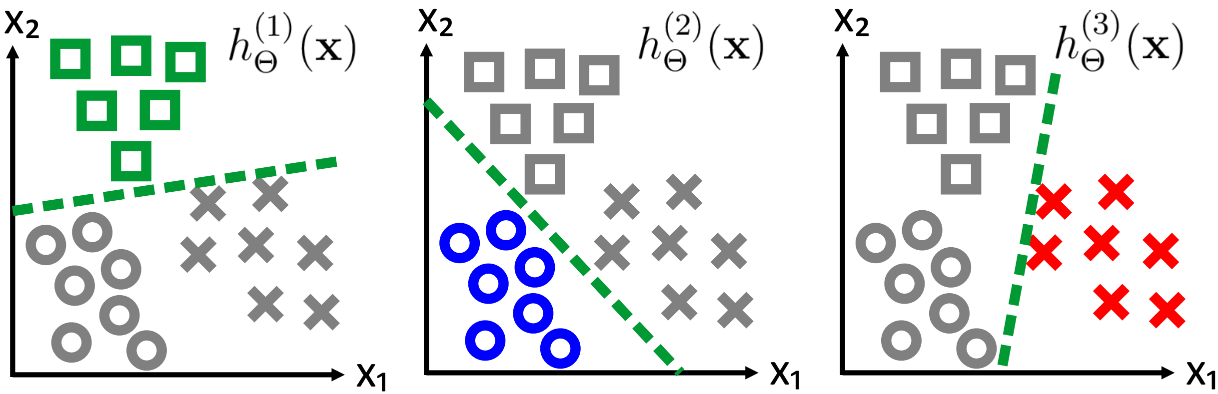
multiclass



One vs All (One vs Rest)
Algoritmus klasifikace do více tříd



One vs All (One vs Rest) Algoritmus klasifikace do více tříd



Obecně $h_\Theta^{(i)}(\mathbf{x}) = P(y = i|\mathbf{x}; \Theta)$, pro $i \in \{1, 2, 3\}$

Tj. natrénujeme *i* binárních logistických regresních klasifikátorů pro každou třídu *i* tak, aby predikovaly pravděpodobnost příslušnost vzorku k *i*-té třídě.

Při klasifikaci neznámého vzorku vybíráme cílovou třídu *i* tak, aby $\max_i h_\Theta^{(i)}(\mathbf{x})$