Strojové učení

Úvod, lineární regrese

Marta Vomlelová
marta@ktiml.mff.cuni.cz

References

[1] P. Berka. *Dobývání znalostí z databází*. Academia, 2003.

[2] T. Hastie, R. Tishirani, and J. Friedman. *The Elements of Statistical Learning, Data Mining, Inference and Prediction*. Springer Series in Statistics. Springer, 2003.

[3] T. Mitchell. *Machine Learning*. McGraw Hill, New York, 1997.

[4] S. Russel and P. Norwig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 2003.

[5] I.H. Witten and E.Frank. *Data Mining - Practical machine learning tools and techniques with Java implementation*. Accademic Press Pub., USA, 1999.

Strojové učení

Program se **učí** ze zkušenosti **data** vzhledem k nějaké **třídě úkolů T** a **míře úspěšnosti (chyby) U (resp. Err)**, pokud se jeho výkon na úkolech třídy T zlepšuje s přibývající zkušeností **data**.

Užití strojového učení – příklady

- Predikce, zda pacient hospitalizovaný s infarktem bude mít druhý infarkt. Predikci můžeme založit na demografických datech, stravě a zdravotím stavu (výsledcích vyšetření) pacienta.
- Predikce ceny akcií za 6 měsíců, na základě informací o společnosti a celkovém stavu ekonomiky.
- Rozpoznat ručně psané PSČ z digitalizovaného obrazu.
- Odhadnout množsví glukózy v krvi diabetického pacienta z infračerveného spektra krve pacienta.
- Identifikovat rizikové faktory rakoviny, dle klinických a demografických dat.

Dva přístupy tvorby modelu

- **Expertní –** expert vytvoří model
  ukázalo se jako těžko schůdné najít a zaplatit experta ochotného a schopného vytvořit model.
- **Sehnat data a naučit model z dat**
  daleko schůdnější cesta, otázka je, nakolik je model použitelný v praxi.
- **Spolupráce experta a strojového učení**
  (podle mě) ideální varianta, expert snáze kritizuje (opravuje) model vytvořený z dat než aby tvořil model celý sám.

Základní pojmy

- **data**

	A_1	A_j	A_n	Cílový atribut
$X =$ vektor	$\langle X_1$	X_j	$X_n\rangle$	Y nebo G
x_1				
$x_i =$ vektor	$\langle x_1$	x_j	$x_n\rangle$	y nebo g
x_N				

- **kvantitativní proměnné**
- **kvalitativní proměnné** – různé třídy (kategorie, diskrétní veličiny, faktory), dvě či více, uspořádané či neuspořádané

- **vstupní proměnné** Vstupní (nezávislé) proměnné značíme symbolem X, j–tou proměnnou odkazujeme X_j (alternativně A_j resp. velká písmena $A, B, \ldots$). Pozorovanou hodnotu značíme malým písmenem x_i i v případě, že jde o vektor. Index i znamená, že jde o i–té pozorování, $i = 1, \ldots, N$. Je–li X vektor, všechna pozorování dohromady tvoří matici $\mathbf{X}$ rozměrů $N \times n$. Tučně značíme pouze vektory přes všechny příklady (tj. rozměru N), jinak vektory zůstávají normálním písmem, tj. x_i je vektor i–tého příkladu, $\mathbf{x_j}$ je vektor pozorování j–té proměnné přes všechny příklady.
- **Cílová proměnná** Proměnná, kterou známe u trénovacích dat, ale ve výsledku chceme na nových datech tuto proměnnou predikovat na základě ostatních (vstupních) veličin. Kvantitativní cílovou proměnnou značíme Y, kvalitativní značíme G (group, skupina).

- **Úloha strojového učení** Cílem učení je vytvořit model (funkci), která pro každou hodnotu vstupních proměnných X vydá dobrou predikci $\hat{Y}$ výstupu Y, resp. $\hat{G}$ kategorie G pro diskrétní případ.
- **regrese** Predikujeme–li numerický atribut.
- **klasifikace** Predikujeme–li diskrétní atribut.

Příklady modelů

- uložená data
- lineární funkce
- nelineární funkce (např. báze funkcí a koeficienty jejich lineární kombinace, logistická regrese, SVM)
- rozhodovací strom (rozhodovací známka a jejich kombinace)
- množina pravidel (jen konstanty nebo i proměnné ILP)
- bayesovská síť
- neuronová síť
- funkce skrytá v algoritmu vytvoření predikce
- ...

Co vše je třeba

- připravit data – my trochu, jinak *data mining*
- naučit model
  - který typ modelu (záleží na problému)
  - který model daného typu (funkce odhadující chybu modelu)
- otestovat model – nejlépe na nových datech.

Software

- Weka http://www.cs.waikato.ac.nz/ ml/weka/
  GNU program v Java
- mnoho jiných

Navrhněte model (1)

DenVTýdnu	VýrobceMěřáku	MnožstvíSrážek
po	rr	2.0
po	zz	0
út	zz	1.1
st	zz	1.9
st	rr	0.0

Navrhněte model (2)

BarvaTrička	Snídal?
červená	ano
modrá	ne
zelená	ano
bílá	ano
bílá	ne

Navrhněte model (3)

Pohlaví	Výška
muž	183
muž	179
žena	168
žena	182
muž	165

Navrhněte model (4)

Výška	Pohlaví
183	muž
179	muž
168	žena
182	žena
165	muž

Navrhněte model (5)

Výška	Váha

Navrhněte model (6)

IDKlienta	ZůstatekNaÚčtu

Lineární modely

Lineární modely

n	počet atributů, tj. dimenze x
N	počet příkladů v datech
β	n, resp. $n+1$ rozměrný vektor parametrů modelu
$\hat{y}$	odpověďní veličina, tj. naše predikce cílové funkce $f(x)$
i	index procházející jednotlivé příklady
j	index procházející jednotlivé dimenze

Lineární regrese

- Cíl: aproximovat funkci $f(x)$, kde x je n–rozměrný vektor, pomocí lineární funkce

$$\hat{y} = \hat{\beta}_0 + \sum_{j=1}^{n} x_j \hat{\beta}_j$$

- Pokud do x přidáme 1, tj. vytvoříme vektor $\langle 1, x \rangle$, $\hat{\beta}_0$ schováme do $\hat{\beta}$ a píšeme:

$$\hat{y} = \sum_{j=0}^{n} x_j \hat{\beta}_j = x^T \beta$$

- Sumu $\sum_{j=0}^{n} x_j \hat{\beta}_j$ můžeme zapsat vektorově jakožto skalární součin $x^T \beta$.

Lineární regrese

- Pokud navíc necháme index i procházet jednotlivé trénovací příklady, můžeme y chápat jako vektor odpovědí na jednotlivé příklady, X jako matici $N \times n$ jednotlivých příkladů a psát:

$$\hat{y} = X\beta$$

- Hledáme takové hodnoty parametrů $\hat{\beta}$, aby chyba aproximace byla co nejmenší. Za míru chyby se téměř vždy bere součet čtverců reziduí (RSS – residual sum squares), tj.

$$RSS(\beta) = \sum_{i=1}^{N} (y_i - x_i^T \beta)^2 = (y - X\beta)^T (y - X\beta)$$

Lineární regrese

- Derivací podle β dostaneme normální rovnici

$$X^T(y - X\beta) = 0$$

- Není–li $X^T X$ singulární, dostaneme jednoznačné řešení

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

- a odhad $\hat{y}$ pro dané x_i je $\hat{y}(x_i) = x_i^T \hat{\beta}$
- Není–li $X^T X$ invertibilní, uvereme závislé sloupce (tj. atributy) nebo se pokusíme překódovat či filtrovat data tak, aby matice invertibilní byla.

Lineární regrese pro klasifikaci

Dvě třídy

- jednu třídu kódujeme 0, druhou 1, najdeme lineární model této kódované funkce.
- Pokud model predikuje $y \leq 0,5$, predikujeme první třídu, jinak predikujeme druhou třídu.
- Hranice $\{x : x^T\beta = 0,5\}$ se nazývá **rozhodovací hranice (decision boundary)**.

Lineární regrese pro klasifikaci

K tříd

- Každý příklad v datech patří do (právě jedné) z k tříd $G_1, \ldots, G_K$. Pak zavedeme indikátory, tj. proměnné y_k nabývající 1 právě když příklad patří do třídy G_k, jinak $y_k = 0$.
- Spočteme naráz modely pro všechny indikátory, tj. Y bude matice $K \times N$ a

$$\hat{B} = (X^T X)^{-1} X^T Y$$

- Pro klasifikaci nového příkladu x pak nejdříve spočteme vektor predikcí indikátorů

$$\hat{f}(x) = [\langle x, 1 \rangle \hat{B}]^T$$

- a pak najdeme takovou třídu, jejíž indikátor nabývá největší

hodnoty, tj.

$$\hat{G}(x) = argmax_{k=1,\ldots,K}\hat{f}_k(x)$$

- Při použití lineární regrese pro klasifikaci může dojít k maskování tříd, např. pro tři třídy v přímce klasifikuji vždy do jedné z krajních, střední třída nikdy nenabyde maximální hodnoty indikátoru.