Úvod do umělé inteligence (NAIL120)

10. cvičení

Jirka Fink

https://ktiml.mff.cuni.cz/~fink/

Katedra teoretické informatiky a matematické logiky
Matematicko-fyzikální fakulta
Univerzita Karlova v Praze

Letní semestr 2023/24

Poslední změna 29. dubna 2024

8. domácí úkol: Teorie her

Zadání (zkráceno)

Pravidla hry

- Máme dva hráče, kteří po jednom odebírají kameny očíslované $1, 2, \ldots, n$.
- Hráč může odebrat libovolný kámen, který je násobek nebo dělitel předchozího odebraného čísla.
- Pokud hráč nemůže odebrat kámen, tak prohrává.

Napište funkci, která dostane seznam zbývajících kamenů a poslední odebraný kámen a rozhodne, zda je daná situace vyhrávající a pokud ano, tak vrátí kámen, který má hráč odebrat v dalším tahu.

Dizkuze

- K čemu se používá strojové učení?
- Jaké nástroje strojového učení znáte?
- Proč nezačneme rovnou s neuronovými sítěmi?

Základní přístupy

- **Učení s učitelem** (supervised learning): Při trénování dostáváme příklady i s očekávanou odpovědí a při testování máme na další příklady uhodnout správnou odpověď.
- **Učení bez učitele** (unsupervised learning): Očekávané odpovědi nejsou k dispozici při trénování, ale můžeme například aspoň zkusit zadaná data rozdělit do skupin podle podobnosti.
- **Zpětnovazební učení** (reinforcement learning): Při trénování máme tipovat výsledek a dozvíme se, jestli (jak dobře) jsme se trefili, a tím zlepšujeme další odpovědi.

Popis

Cílem je data rozdělit do dvou nebo více skupin, například

- Na základě symptomů rozhodnout, zda pacient je zdravý
- Pojmenovat zvířata na obrázcích
- Zprávy rozdělit do kategorií

Měření kvality binární kvasifikace

	Zdravý	Nemocný
Test negativní	TN	FN
Test pozitivní	FP	TP

- Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$: Správnost odpovědí ze všech testů
- Precision = $\frac{TP}{TP+FP}$: Správnost odpovědí ze všech pozitivních testů
- Recall = $\frac{TP}{TP+FN}$: Správnost odpovědí ze všech nemocných jedinců

Měření kvality binární kvasifikace

	Zdravý	Nemocný
Test negativní	TN	FN
Test pozitivní	FP	TP

- Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$: Správnost odpovědí ze všech testů
- Precision = $\frac{TP}{TP+FP}$: Správnost odpovědí ze všech pozitivních testů
- Recall = $\frac{TP}{TP+FN}$: Správnost odpovědí ze všech nemocných jedinců

Proč nestačí jedno kvalitativní číslo?

Určete kvalitu testu, jestliže nemoc má 1 % obyvatel a

1. test je vždy negativní
2. test z nemocných pozná 1 % a jinak je negativní
3. test je vždy pozitivní

Simpsonův paradox, 1951

Který student je lepší?

Tabulka udává úspěšnost dvou studentů u zkoušek v průběhu dvou semestrů.

	1. semestr		2. semestr		součet	
Student A	2/8	25%	2/2	100%	4/10	40%
Student B	1/5	20%	4/5	80%	5/10	50%

Další příklady například na Wikipedii.

Mark Twain, 1907

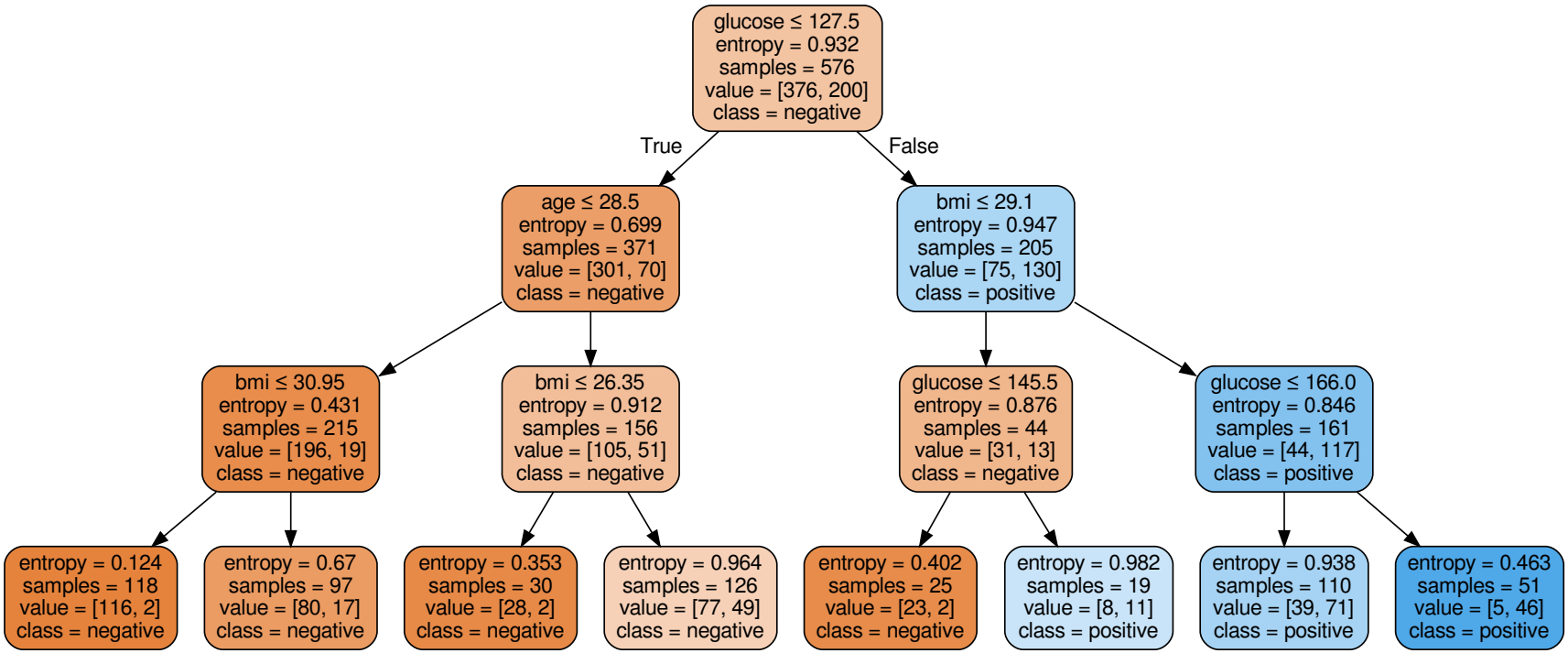
Existují tři druhy lží:

- lež
- nehorázná lež
- statistika

9. domácí úkol: Rozhovovací stromy

- Chtěli bychom rozpoznat, zda člověk s danými diagnostikami má cukrovku
- Máme k dispozici CSV soubor s různými diagnostikami včetně cukrovky k učení i testování
- Cílem je nastavit parametry vytváření rozhodovacího stromu tak, aby úspěšnost byla co největší
- Postupujte tak, aby výsledek byl statisticky relevantní
- Odevzdejte zprávu v PDF obsahující
  - Popište parametry, které měli vliv na výsledek
  - Napište nejlepší parametry, kterých jste dosáhli
  - Vytvořte graf závislosti přesnosti a úplnosti na velikosti testovací množiny
  - Nakreslete nejlepší rozhodovací strom

Rozhodovací stromy: Příklad



Rozhodovací stromy: Konstrukce (trénování)

Jak vytvořit rozhodovací strom

- Určit nejdůležitější atribut, podle kterého vrchol rozvětvíme
- Rozdělit vzorky podle tohoto kritéria
- Pokračujeme rekurzí dokud máme vzorky stejné kategorie

Jak určit nejdůležitější atribut?

Využijeme entropii

- Máme n vzorků rozdělených do c kategorií
- n_i je počet vzorků kategorie $i = 1, \ldots, c$
- Definujeme $p_i = n_i / n$
- Entropie je $H = -\sum_{i=1}^{c} p_i \log_2 p_i$

Najdeme atribut, který nejvíce sníží entropii

- Atribut rozdělí vzorky do dvou skupin velikosti k_1 a k_2 entropie H_1 a H_2
- Vážený součet entropií je $\frac{k_1}{k} H_1 + \frac{k_2}{k} H_2$

Rozhodovací stromy: Jak zlepšit výsledek

- Vynechat atributy, které nemají vliv na výsledek
- Vrcholy s málo vzorky nedělit
- CCP/CPA analýza vynechávající podstromy, které nezlepšují výsledek
- Podívat se do dokumentace, co knihovna nabízí

Rozhodovací stromy: příklady

Otázka

Jestliže při rozdělování nějakého vrcholu u podle libovolného atributu dostáváme vždy stejnou entropii jako je entropie ve vrcholu u, má smysl pokračovat ve vytváření podstromu?

XOR

- Vytvořte rozhodovací strom pro funkci XOR tří argumentů
- Jak vypadá nejmenší rozhodovací strom, který vždy správně rozhoduje?

Rozhodovací graf

- Rozhodovací graf se liší od rozhodovacího stromu tím, že jeden vrchol může mít více otců, takže do jednoho vrcholu se můžeme dostat více rozhodnutími
- Jak vypadá nejmenší rozhodovací graf funkce XOR, který vždy správně rozhoduje?
- Proč se rozhodovací grafy v praxi nepoužívají?