

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 2021/22

Poslední změna 3. května 2022

Licence: Creative Commons BY-NC-SA 4.0

Zadání (zkráceno)

Pravidla hry

- Máme dva hráče, kteří po jednom odebírají kameny očíslované $1, 2, \dots, n$.
- 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?

Dizkuze

- K čemu se používá strojové učení?
- Jaké nástroje strojového učení znáte?

Dizkuze

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

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.

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.

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

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 klasifikace

	Zdravý	Nemocný
Test negativní	TN	FN
Test pozitivní	FP	TP

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 klasifikace

	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 klasifikace

	Zdravý	Nemocný
Test negativní	TN	FN
Test pozitivní	FP	TP

Měření kvality binární klasifikace

	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 klasifikace

	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

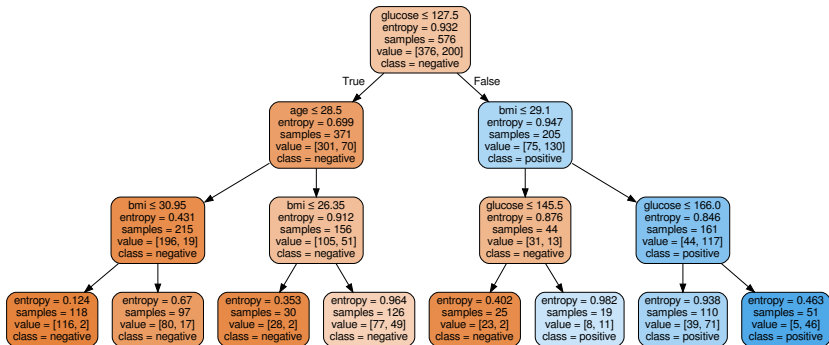
Proč nestačí jedno kvalitativní číslo?

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

- 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 Python program a zprávu v PDF obsahující
 - Popište parametry, které měli vliv na výsledek
 - Vytvořte graf závislosti přesnosti a úplnosti na velikosti testovací množiny
 - Nakreslete jeden výsledný rozhodovací strom

Rozhodovací stromy: Příklad



Jak vytvořit rozhodovací strom

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

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 $i = 1, \dots, c$
- Definujeme $p_i = n_i/n$
- Entropie je $H = - \sum_{i=1}^c p_i \log_2 p_i$

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 $i = 1, \dots, c$
- 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 $\frac{k_1}{k} H_1 + \frac{k_2}{k} H_2$

- 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

Otázka

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

Otázka

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

XOR

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

Otázka

Jestliže při rozdělování podle libovolného atributu dostáváme vždy stejnou entropii jako je původní entropie, 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?