SLOVENSKÁ TECHNICKÁ UNIVERZITA
Fakulta informatiky a informačných technológií
Ilkovičova 2, 842 16 Bratislava 4

Umelá inteligencia

Zadanie 4a – Klasifikácia

Lucia Murzová
Ing. Ivan Kapustík - Streda 13:00
13.12.2021

Lucia Murzová
Streda 13:00

Obsah

Lucia Murzová
Streda 13:00

1. Klasifikácia

Máme 2D priestor, ktorý má rozmery X a Y, v intervaloch od -5000 do +5000. V tomto priestore sa môžu nachádzať body, pričom každý bod má určenú polohu pomocou súradníc X a Y. Každý bod má unikátne súradnice (t.j. nemalo by byť viacej bodov na presne tom istom mieste). Každý bod patrí do jednej zo 4 tried, pričom tieto triedy sú: red (R), green (G), blue (B) a purple (P). Na začiatku sa v priestore nachádza 5 bodov pre každú triedu (dokopy teda 20 bodov). Súradnice počiatočných bodov sú:

R: [-4500, -4400], [-4100, -3000], [-1800, -2400], [-2500, -3400] a [-2000, -1400]

G: [+4500, -4400], [+4100, -3000], [+1800, -2400], [+2500, -3400] a [+2000, -1400]

B: [-4500, +4400], [-4100, +3000], [-1800, +2400], [-2500, +3400] a [-2000, +1400]

P: [+4500, +4400], [+4100, +3000], [+1800, +2400], [+2500, +3400] a [+2000, +1400]

Vašou úlohou je naprogramovať klasifikátor pre nové body – v podobe funkcie classify(int X, int Y, int k), ktorá klasifikuje nový bod so súradnicami X a Y, pridá tento bod do nášho 2D priestoru a vráti triedu, ktorú pridelila pre tento bod. Na klasifikáciu použite k-NN algoritmus, pričom k môže byť 1, 3, 7 alebo 15.

Na demonštráciu Vášho klasifikátora vytvorte testovacie prostredie, v rámci ktorého budete postupne generovať nové body a klasifikovať ich (volaním funkcie classify). Celkovo vygenerujte 20000 nových bodov (5000 z každej triedy). Súradnice nových bodov generujte náhodne, pričom nový bod by mal mať zakaždým inú triedu (dva body vygenerované po sebe by nemali byť rovnakej triedy):

- R body by mali byť generované s 99% pravdepodobnosťou s X < +500 a Y < +500
- G body by mali byť generované s 99% pravdepodobnosťou s X > -500 a Y < +500
- B body by mali byť generované s 99% pravdepodobnosťou s X < +500 a Y > -500
- P body by mali byť generované s 99% pravdepodobnosťou s X > -500 a Y > -500

(Zvyšné jedno percento bodov je generované v celom priestore.)

Návratovú hodnotu funkcie classify porovnávajte s triedou vygenerovaného bodu. **Na základe týchto porovnaní vyhodnoťte úspešnosť** Vášho klasifikátora pre daný experiment.

Experiment vykonajte 4-krát, pričom zakaždým Váš klasifikátor použije iný parameter k (pre k = 1, 3, 7 alebo 15) a vygenerované body budú pre každý experiment rovnaké.

Vizualizácia: pre každý z týchto experimentov vykreslite výslednú 2D plochu tak, že vyfarbíte túto plochu celú. Prázdne miesta v 2D ploche vyfarbite podľa Vášho klasifikátora.

Dokumentácia musí obsahovať opis konkrétne použitého algoritmu a reprezentácie údajov. V závere zhodnoťte dosiahnuté výsledky ich porovnaním.

Poznámka 1: Je vhodné využiť nejaké optimalizácie na zredukovanie zložitosti, napríklad pre hľadanie k najbližších bodov si rozdelíme plochu na viaceré menšie štvorce, do ktorých umiestňujeme body s príslušnými súradnicami, aby sme nemuseli vždy porovnávať všetky body, ale len body vo štvorci, kde sa nachádza aktuálny bod a susedných štvorcoch.

Je tiež možné využiť algoritmus na hľadanie k najmenších hodnôt.

Poznámka 2: Úlohu je možné riešiť aj pomocou neurónovej siete, pričom je vhodné použiť nejaký framework (napríklad PyTorch).

Lucia Murzová
Streda 13:00

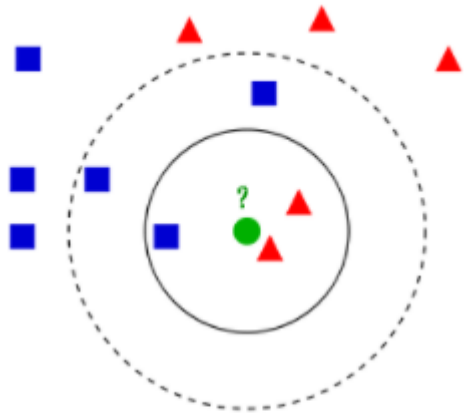
2. KNN algoritmu

Algoritmus K-nearest neighbor alebo aj K najbližších susedov je z hľadiska princípu jeden z najjednoduchších algoritmov strojového učenia, ktorý sa využíva na klasifikáciu.

Algoritmus v tréningovej fáze najprv umiestni prvky tréningovej množiny do priestoru tak, aby prvky z rovnakých tried vytvárali zhluky. Nasleduje fáza klasifikácie kde je prvok, ktorý je potrebné klasifikovať umiestnený do priestoru podľa jeho vlastností. Potom sa v priestore nájde k najbližších prvkov okolo klasifikovaného prvku. Následne je prvok klasifikovaný do triedy, do ktorej patrí väčšina k najbližších prvkov (obr. 1).

Číslo K musí byť prirodzené číslo s pri klasifikačnom probléme s dvomi triedami musí byť k nepárne číslo. Inak by mohlo dôjsť pri klasifikácii k situácii, že by sa počet prvkov oboch tried vo vyhradenom priestore rovnal a algoritmus by nedokázal určiť, ku ktorej triede klasifikovaný prvok pravdepodobnejšie bude patriť. Pre tento istý dôvod k nesmie byť ani násobok počtu tried.

ZDROJ: https://smnd.sk/mcibula/alg/KNN.html



Obr. 1: Príklad klasifikácie vykonávanej algoritmom *k-nearest neighbors*. Ak $k = 3$, zelený kruh bude klasifikovaný ako červený trojuholník. Ak $k = 5$, zelený kruh bude klasifikovaný ako modrý štvorec.

3. Opis riešenia

Globálne premenné

- Vsetky_x [] – Pole s uloženými x polohami všetkých vygenerovaných bodov
- Vsetky_y [] – Pole s uloženými y polohami všetkých vygenerovaných bodov
- Vsetky_f [] - Pole s uloženými farbami všetkých vygenerovaných bodov
- Uspesne_klasifikovane – počet bodov, ktoré boli úspešne klasifikované
- Pocet_bodov – počet bodov, ktoré majú byť vygenerované z jednej farby

Výstup

Výstupom programu sú súbory s uloženými polohami a farbami jednotlivých bodov pre k = 15, k = 7, k = 3 a k = 1. Následne je potrebné pustiť program vykreslenie.py, ktorý načíta tieto uložené výstupy a vykreslí ich grafy.

Vykreslenie

Program je rozdelený do dvoch súborov – main.py a vykreslenie.py z dôvodu, že pre rýchlejšie výpočty som používala just in time compiler PyPy 3.10, ktorý však momentálne nepodporuje knižnicu matplotlib, potrebnú pre vykreslenie. Program z tohto dôvodu uloží do jednotlivých súborov pre všetky k x, y a farby finálneho rozloženia bodov.

Algoritmus

Program na začiatku inicializuje tréningovú skupinu bodov, ktoré pridá do poľa pre všetky vygenerované body. Následne začína s k = 15 a postupne generuje body z farieb červená, zelená, modrá a fialová, pričom kontroluje predošlé body aby boli súradnice vždy unikátne.

Po vygenerovaní prechádza všetky predošlé body a počíta ich vzdialenosti. Pole vzdialeností následne usporiada a vyberie z neho K najbližších. Podľa najväčšieho počtu jednotlivých farieb najbližších k susedov určí farbu daného bodu ktorú táto funkcia vráti. Pridelená farba je porovnávaná s triedou v ktorej bol bod generovaný a ak sa tieto farby zhodujú, navýši sa počítadlo úspešne klasifikovaných.

Tieto výsledky sa uložia do súborov pre výpis a program pokračuje s k = 7, k = 3 a k =1, pričom pri týchto už negeneruje nové hodnoty, ale pracuje s tými, ktoré sa vygenerovali v prvom cykle. Každé k má svoj súbor pre uloženie x, y a farieb jednotlivých bodov, s ktorými následne pracuje program pre vykreslenie.

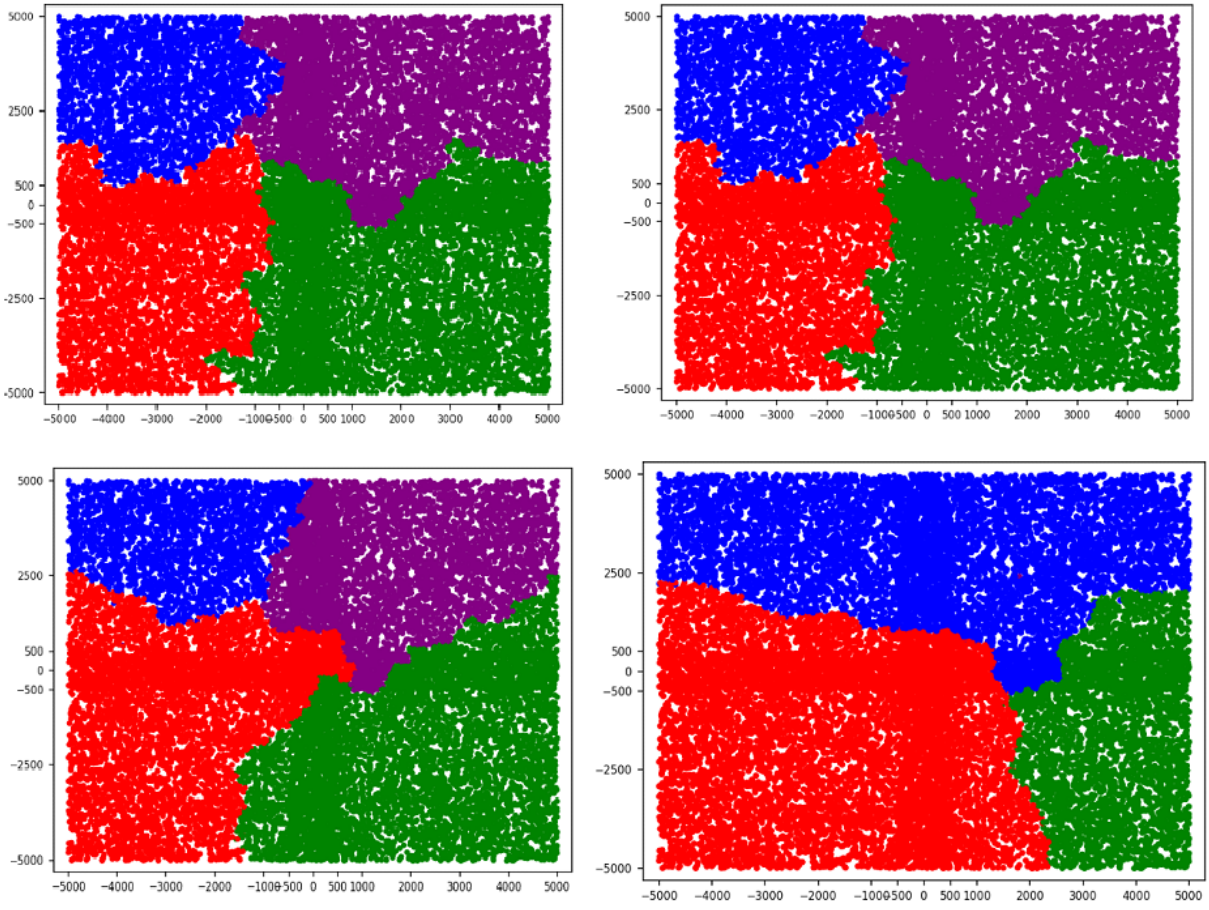
Lucia Murzová
Streda 13:00

4. Testovanie

Všetky testy boli spravené s rovnakými sadami bodov pre všetky k – 1, 5, 7 a 13. Časy pre k = 15 sú vrátane generovania unikátnych bodov, pri ďalších k je uvedený celkový čas programu, nie len čas behu danej klasifikácie. Nižšie uvedené obrázky jednotlivých testov sú v poradí k = 1, k = 5, k = 7 a k = 13.

Test. Č. 1

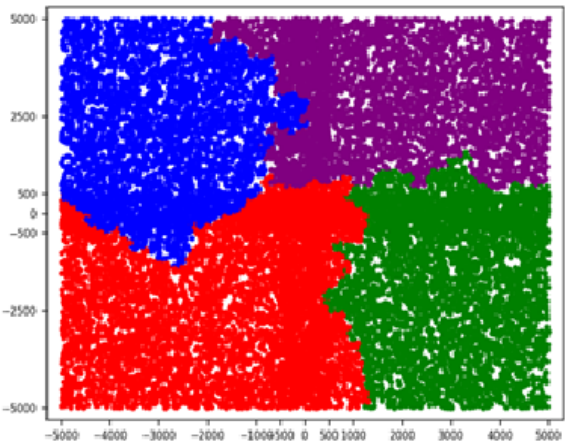
```
Uspesnost k=15 0.5397
--- 158.0356628894806 seconds ---
Uspesnost k=7 0.71295
--- 315.6047217845917 seconds ---
Uspesnost k=3 0.73555
--- 503.66103076934814 seconds ---
Uspesnost k=1 0.73545
--- 663.3404679298401 seconds ---
```

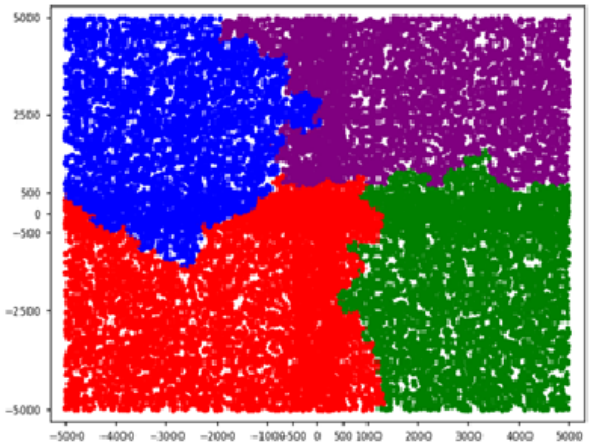


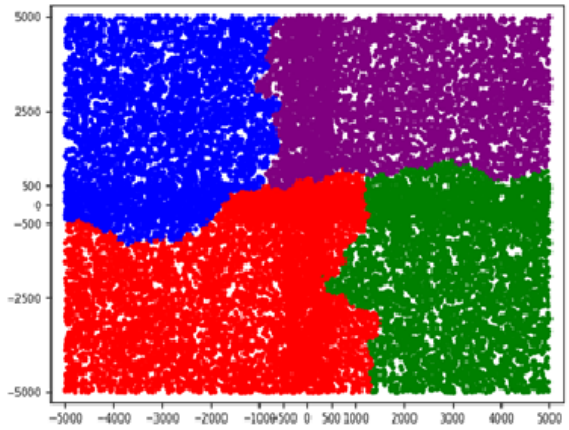
Lucia Murzová
Streda 13:00

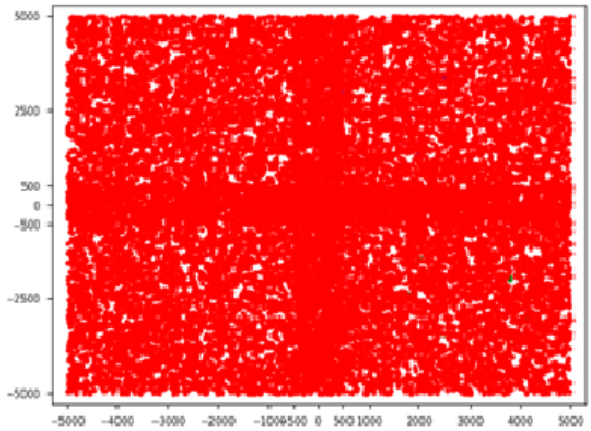
Test č. 2

```
Uspesnost k=15 0.25015
--- 215.65861797332764 seconds ---
Uspesnost k=7 0.76055
--- 397.57795095443726 seconds ---
Uspesnost k=3 0.75965
--- 579.8604960441589 seconds ---
Uspesnost k=1 0.7601
--- 752.8167960643768 seconds ---
```





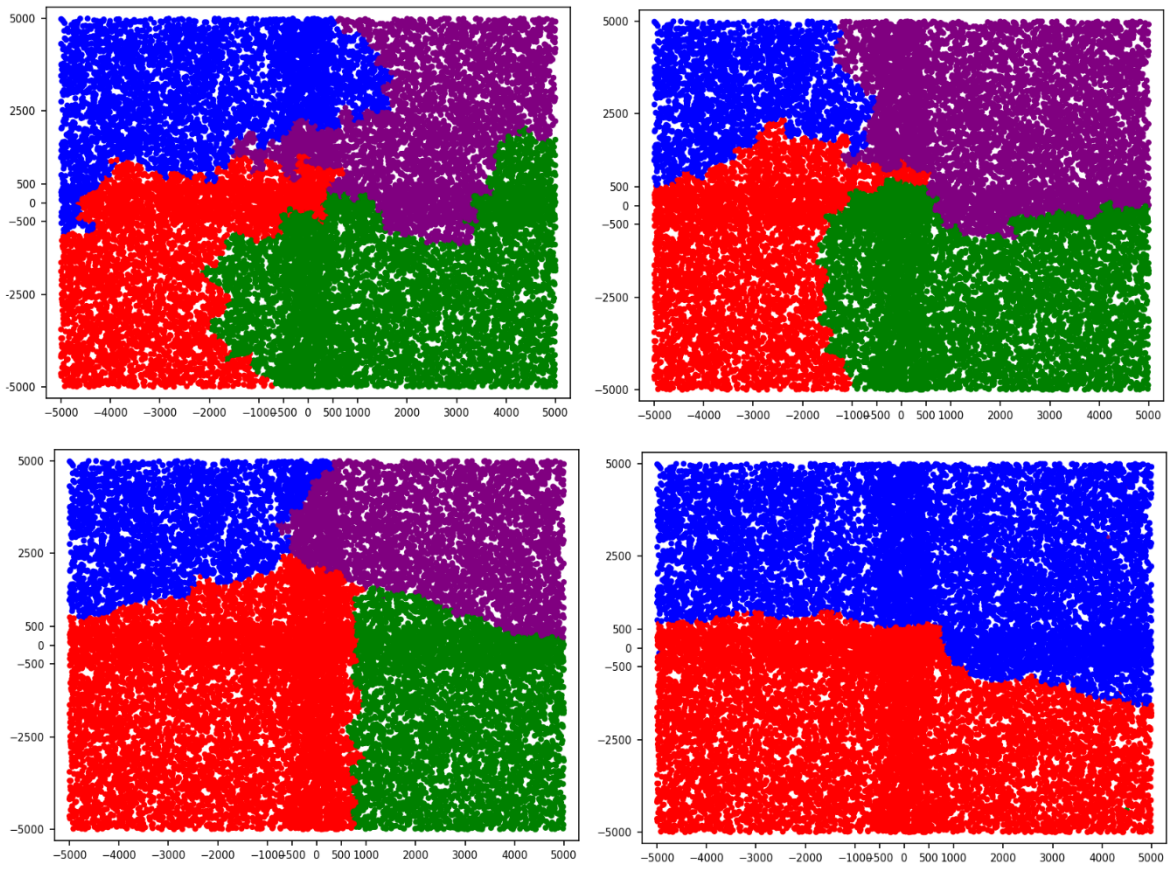




Lucia Murzová
Streda 13:00

Test č. 3

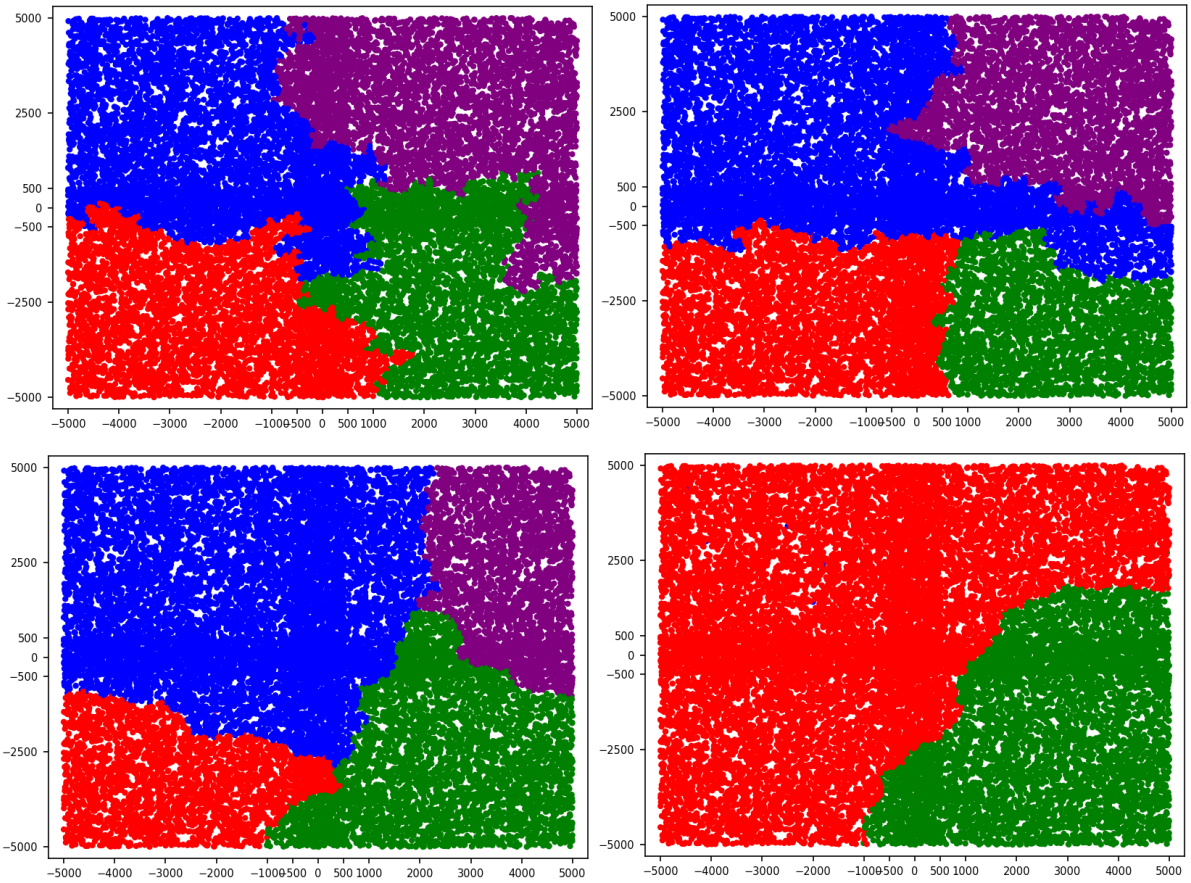
```
Uspesnost k=15 0.44125
--- 163.58044695854187 seconds ---
Uspesnost k=7 0.7416
--- 327.3952741622925 seconds ---
Uspesnost k=7 0.72985
--- 491.46408200263977 seconds ---
Uspesnost k=7 0.7217
--- 650.6664900779724 seconds ---
```



Lucia Murzová
Streda 13:00

Test č. 4

```
Uspesnost k=15 0.4391
--- 228.8017818927765 seconds ---
Uspesnost k=7 0.68135
--- 403.86546897888184 seconds ---
Uspesnost k=7 0.74105
--- 561.3688569068909 seconds ---
Uspesnost k=7 0.75765
--- 718.7751989364624 seconds ---
```



Lucia Murzová
Streda 13:00

5. Zhodnotenie testovania a riešenia

Priemerná úspešnosť:

- K = 1: 74,29 %
- K = 5: 74,14 %
- K = 7: 72,4 %
- K = 15: 41,75 %

Z testovania je jasné, že najnižšia úspešnosť je pri vyššom počte susedov – v tomto prípade pri k = 15. V tomto prípade preto dochádza často k zanedbaniu niektorých tried – v teste č. 1 fialovej. V teste č. 3 a 4 boli zanedbané dve farby a v teste č. 2 dokonca až k troch farieb.

Pri k = 1 a 3 sú výsledky veľmi podobné, čo vo väčšine prípadov boli aj konkrétne výsledky.