**Linearna regresija**

Prvi korak predstavlja izbor atributa. Za atribute sam izabrao kvadraturu, sprat na kom se nalazi stan i udaljenost od centra grada u kilometrima. Ove parametre sam izabrao jer imaju najveći uticaj prilikom određivanja cene stana. Ciljnu promenljivu predstavlja cena stana. Linearna regresija tada ima sledeći oblik:

– kvadratura

– sprat

– udaljenost od centra grada

Potrebno je odrediti parametre linearne regresije , , na osnovu dostupnog skupa podataka.

Implementacija algoritma

Na početku algoritma potrebno je inicijalizovati vektor nulama, gde je

Kako se vrednost parametara traži pomoću stohastičkog gradijentnog spusta, odlučio sam se da koristim mini-batch stohastički gradijent. Jedan mini-batch sadrži 200 primera. Vrednost stope učenja postavio sam na 0.01, kako se pokazala kao najbolja (za ovu vrednost sigurno dostiže globalni minimum kriterijumske funkcije). Kriterijumska funkcija je srednja vrednost kvadrata razlike , gde je broj primera u mini–batch-u.

Vektor koji predstavlja jedan primer proširio sam na početku sa 1, kako bi dimenzije vektora parametra i vektora atributa bile iste i tada se predikcija cene računa kao .

Kada se odredi predikcija na osnovu trenutnih vrednosti parametara, ažuriraju se vrrednosti parametara na sledeći način:

Uslov za konvergenciju algoritma je kada kriterijumska funkcija dostigne minimum.

Vrednosti koje se dobiju za parametre su sledeće:

Vrednosti parametara koje vraća ugrađena funkcija su sledeći:

Iz priloženog može se zaključiti da algoritam dobro određuje vrednosti parametara. Parametar ofgovara srednjoj vrednosti cena, što i jeste slučaj. Primećuje se da najveći uticaj na cenu stana ima kvadratura, zatim udaljenost što je i očekivano. Udaljenost ima negativan učinak. Sa povećanjem udaljenosti cena će opadati.

Iz datog supa prosečna kvadratura iznosi 83, sprat 3 i udaljenost 3km. Prosečna cena stanova je 269496. Kada se unesu ovi podaci cenu koju daje naš model iznosi 267473, što je očekivan rezultat.

|  |  |  |  |
| --- | --- | --- | --- |
| Kvadratura | Sprat | Udaljenost | Prediktovana cena |
| 40 | 2 | 2.5 | 138219 |
| 90 | 3 | 5 | 230341 |
| 120 | 1 | 4 | 329057 |

**K - najbližih suseda**

Atributi koji su korišćeni u ovom zadatku su: kvadratura, udaljenost od centra grada, uknjiženost i broj soba. Pored ovih atributa isprobao sam i atribute kao što su: da li ima parking ili ne, da li ima lift ili ne, broj kupatila, medjutim oni su davali lošije rezultate. Opseg izlaznih vrednosti podelio sam na 6 klasa, ne sledeći način: do 50k, 50k-100k, 100k-150k, 150k-200k, 200k-500k, od 500k. Faktor k samo odredio kao . Funckije rastojanja koje sam koristio su: euklidsko rastojanje i menhetn distancu.

Euklidska distanca:

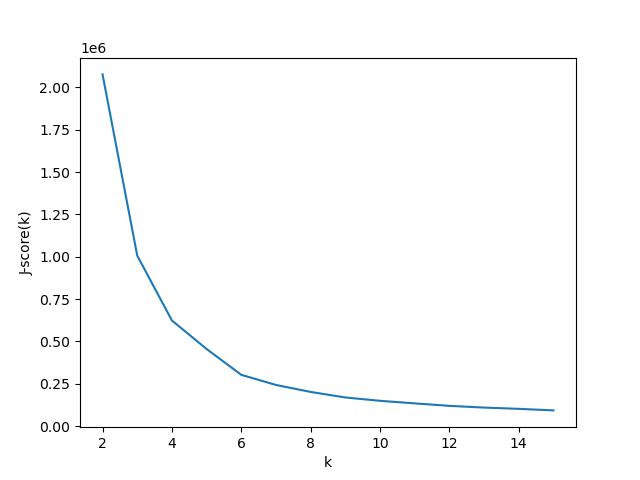
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Kvadratura | Udaljenost | Uknjiženo | Sprat | Broj soba | Klasa |
| 50 | 4 | ne | 6 | 2 | 100k-150k |
| 80 | 1 | da | 10 | 3 | 200k-500k |
| 80 | 7 | da | 3 | 3 | 150k-200k |
| 30 | 6 | da | 0 | 1 | 50k-100k |
| 75 | 8 | ne | 2 | 3 | 100k-150k |

Menhetn distanca:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Kvadratura | Udaljenost | Uknjiženo | Sprat | Broj soba | Klasa |
| 50 | 4 | ne | 6 | 2 | 100k-150k |
| 80 | 1 | da | 10 | 3 | 200k-500k |
| 80 | 7 | da | 3 | 3 | 150k-200k |
| 30 | 6 | da | 0 | 1 | 50k-100k |
| 75 | 8 | ne | 2 | 3 | 200k-500k |

Rezultati su praktično isti za obe distance.

**K-means**

Atributi na osnovu kojih se vrši klasterizacija: kvadratura, sprat i udaljenost od centra grada. Metod za oređivanje broja klastera koji sam koristio je lakat metod. Vise puta je pokrenuta k-means klasterizacija, gde je k uzimalo vrednosti od 2 do 15. Za svako k pribeležena je vrednost J scora i prikazana grafički. 

Za optimalnu vrednost k treba uzeti onu vrednost gde je funkcija ’najoštrija’. Vrednost koju sam uzeo za k je 6.

Centri klastera su:

1. [[ 83.46659708 3.22442589 2.5989408 ]
2. [203.1119403 4.21641791 2.23306252]
3. [154.32945736 4.5503876 2.26904117]
4. [ 37.37954939 1.76083189 3.75627396]
5. [113.33562823 4.10499139 2.30335271]
6. [ 59.84615385 2.73992674 3.60579184]]

Слика на којој се налази текст, снимак екрана, мапа, Графика

Опис је аутоматски генерисан

Klasteri su pretežno određeni prema kvadraturi, a vidi se i blagi uticaj druga dva atributa kada se grafik posmatra iz druge perspektive.