LINEARNA REGRESIJA

1. Navedite model višestruke linearne regresije i objasnite sve oznake.
   - $y = b_0 + b_1x_1 + b_2x_2 + \ldots b_nx_n$
   - y – predviđana vrednost
   - b_i - parametri
   - x_i – nezavisne promenljive
2. Objasnite čemu služi koeficijent determinacije (R^2).
   - Procena koliko se dobro regresioni model uklapa u podatke.
   - Kreće se između 0-1.
   - 0 – nikakvo uklapanje.
   - 1 – savršeno uklapanje.
   - $R^2 = 0.72$ - 72% varijanse u y možemo objasniti pomoću x, 28% varijacije je šum koji nije obuhvaćen modelom.
3. Objasnite šta se optimizuje pomoću Metode Najmanjih Kvadrata.
   - Minimizujemo grešku, odnosno razliku predikcija i stvarnih vrednosti sa ciljem da dobijemo parametre modela.
4. Definišite zbir kvadrata grešaka regresionog modela koristeći koje god oznake želite.
   - $\sum_{i=0}^{N}(yi - \tilde{y}i)^2$
5. Objasnite zašto je regresija pomoću polinoma trećeg stepena takođe lin reg.
   - Linearna je po parametrima, ne po nezavisnoj promenljivoj.
6. Objasnite čemu služi t-test.
   - Služi za testiranje statističke hipoteze. Postavimo hipotezu lošeg slučaja – da je parametar = 0, ako je p-vrednost < 0.05 odbacujemo hipotezu, znači da je taj parametar različit od 0. To znači da postoji linearna veza između y i x uz koji stoji taj parametar.
7. Na koji način se interpretira p-vrednost sa 95% pouzdanosti.
   - $p < 0.05$ odbacujemo hipotezu da ne postoji linearna veza između y i x, odnosno postoji linearna veza.
8. Objasnite na koji način se interpretira koeficijent regresionog modela kod višestruke lin reg.
   - Ako povećam neko x za 1 vrednost, koliko će se povećati srednja vrednost y ako su ostali x fiksirani.
9. Ako je dat regresioni model za predikciju cena kuća koji uključuje površinu kuće i broj kupatila, kod koga koeficijenta ispred površine kuća ima negativnu vrednost b_p na koji način se interpretira ta negativna vrednost.
   - Cena kuće opada sa povećanjem površine, ako je fiksan broj kupatila – npr. ogromna kuća sa 1 kupatilom.
10. Objasnite zašto nam je kod višestruke regresije potreban prilagođeni koeficijent determinacije.
    - R^2 – kako dodajemo promenljive biće isti ili rasti, ne znamo da li smo overfitovali i da li su dodate promenljive besmislene, dok prilagođeni uzima u obzir broj nezavisnih promenljivih.
11. Objasnite pojam preprilagođavanja overfittinga.
    - Model je loše uslovljen, za male promene x dobijam velike promene y. Kreće da se prilagođava ne samo signalu nego i šumu.

12. Objasnite na koji način možete da utvrdite da li je neka nezavisna promenljiva korisna u modelu višestruke lin reg.
    - Kod promenljivih kod kojih je p-vrednost < 0.05 te su korisne jer odbacuje hipotezu da nije 0, znači da postoji linearna veza.
13. Objasnite u kojoj situaciji nije dobro ukloniti nezavisnu promenljivu koja ne doprinosi modelu višestruke regresije.
    - Cilj predikcija – možemo da uklonimo iz modela ako nam nije korisna.
    - Cilj istraživanje – hoćemo da vidimo baš vezu između te promenljive i y.
14. Navedite pretpostavke lin reg.
    - Greške prate normlanu rasporedu najverovatnije će podatak biti na samom regresionom modelu, sa malo manjom verovatnoćom udaljen.
    - Na početnom grafiku greške manje pa veće kasnije, a na grafiku reziudala – greške ravnomerno osciluju.
    - Linearnost podataka.
    - Nezavisnost grešaka - y_i ne zavisi od y_{i-1}, ako zavisi onda je model vremenskih serija.
15. Objasnite pretpostavku o linearnosti.
    - Postoji linearna veza između x i y. Ako je narušena radimo transformaciju npr. x^2
16. Objasnite pretpostavku o nezavisnosti grešaka.
    - Nezavisnost grešaka y_i ne zavisi od y_{i-1}. Ako je narušena onda se koristi model vremenskih serija.
17. Objasnite pretpostavku o multikolinearnosti.
    - npr. $x_1 = 2x_2 + x_3$ -> problem ako rešavamo sa metodom najmanjih kvadrata sistem jer bismo dobili sistem sa d jednačina i d nepoznatih, a ako su neki u međusobnoj vezi dobili bismo da sistem nema rešenja.
    - Ako imamo približnu multikoliearnost onda je model nestabilan, skloniji smo overfittingu.
18. Objasnite pretpostavku o konstantnoj varijansi grešaka.
    - Na početnom grafiku greške manje pa veće kasnije, a na grafiku reziudala – greške ravnomerno osciluju.
    - Onda možemo da probamo logaritamsku ili korensku transformaciju y.
19. Objasnite šta su reziduali modela.
    - y_i (stvarno) – y_i (predviđeno) je rezidual.
20. Koja pretpostavka je narušena na datom grafiku (slika).
    - Nezavisnost grešaka. Radili bismo modelom vremenskih serija.
21. Koja pretpostavka je narušena na datom grafiku reziduala (slika).
    - Independence of squares. Na početnom grafiku greške manje pa veće kasnije, a na grafiku reziduala – greške ravnomerno osciluju.

INTERPOLACIJA

1. Opišite problem koji rešavamo interpolacijom i objasnite kako ga rešavamo (navedite osnovnu ideju, nije neophodno da navodite konkretan postupak ili formulu).
   - Imamo skup tačaka i hoćemo da vidimo šta se dešava između tih tačaka. Interpolacija - pretpostavka da će to biti interpolacioni polinom koji precizno prolazi kroz svaku od tačaka.
2. U kom slučaju biste primenili interpolaciju, a u kom regresiju?
   - Regresiju kada tražimo trend u podacima – cena nekretnine za istu kvadraturu ima prodate kuće za različitu cenu. Interpolacija - merenja su precizna – npr. zvuk, slika i izmereni podaci ostaju takvi kakvi jesu pa tražimo samo šta je između podataka.
3. Nepoznata funkcija f je zadata u N tačaka. Interpolacijom određujemo jedinstven polinom koji prolazi kroz svaku od zadatih tačaka: (1) Napišite oblik traženog polinoma, koje parametre određujemo interpolacijom i koliko ih ima. (2) Ukoliko je stepen polinoma prevelik, koju alternativu možemo da primenimo?
   - $y = a_0 + a_1x + a_2x^2 + ... + a_{n-1}x^{n-1}$, određujemo n parametra od a_0 do a_n.
   - splajn – na svakom intervalu imamo poseban splajn, npr. linearni, kvadratni.
4. Kako se polinom predstavljaju u Pythonu? Napišite predstavu polinoma 5.5*x3 + x - 3.
   - Kao vektori, [5.5 0 1 -3].
5. U interpolaciji, kada određujemo jedinstven polinom stepena N-1 koji prolazi kroz N zadatih tačaka, oblik polinoma možemo predstaviti na dva načina: (1) standardni zapis g(x) = a1 + a2x + a3x2 + ... + aNxN-1, (2) Lagranžov zapis i (3) Njutnov zapis. Zašto određivanju interpolacionog polinoma tipično ne koristimo standardni zapis polinoma, već druge alternative?
   - Kod standardnog zapisa – n jednačina sa n nepoznatih, problem je loš uslovljen sistem jer za male promene podivlja rešenje. (crtež skoro 2 paralelne prave)
6. U interpolaciji, kada određujemo jedinstven polinom stepena N-1 koji prolazi kroz N zadatih tačaka, oblik polinoma možemo predstaviti na dva načina: (1) standardni zapis g(x) = a1 + a2x + a3x2 + ... + aNxN-1, (2) Lagranžov zapis i (3) Njutnov zapis. Navedite Njutnov zapis polinoma drugog stepena.
   - $g_N(x) = b_1 + b_2(x-x_1) + b_3(x-x_1)(x-x_2)$
7. U interpolaciji, kada određujemo jedinstven polinom stepena N-1 koji prolazi kroz N zadatih tačaka, oblik polinoma možemo predstaviti na dva načina: (1) standardni zapis g(x) = a1 + a2x + a3x2 + ... + aNxN-1, (2) Lagranžov zapis i (3) Njutnov zapis. Navedite Lagranžov zapis polinoma drugog stepena.

   gL(x) = $\frac{(x-x2)(x-x3)}{(x1-x2)(x1-x3)}f(x1) + \frac{(x-x1)(x-x3)}{(x2-x1)(x2-x3)}f(x2) + \frac{(x-x1)(x-x2)}{(x3-x1)(x3-x2)}f(x3)$
8. U interpolaciji koja dva faktora utiču na tačnost procene dobijene interpolacionim polinomom?
   - Razdaljina tačaka – što je manja biće preciznije.
   - Broj tačaka – stepen polinoma kako se povećava dobijamo preciznije rezultate.
9. U interpolaciji, kada određujemo jedinstven polinom stepena N-1 koji prolazi kroz N zadatih tačaka, oblik polinoma možemo predstaviti na dva načina: (1) standardni zapis g(x) = a1 + a2x + a3x2 + ... + aNxN-1, (2) Lagranžov zapis i (3) Njutnov zapis. Koja je prednost Njutnovog zapisa nad njegovim alternativama?
   - Ako dodamo još jednu tačku, odnosno uvećamo stepen interpolacionog polinoma, oni niži koeficijenti ostaju nepromenjeni pa bismo računali samo još jedan dodatni koeficijent.

10. Šta je ekstrapolacija?
    - Korisitmo isti polinom kao kod interpolacije koja nam govori šta se dešava između tačaka, a kod ekstrapolacije šta se dešava daleko od tih tačaka i trudimo se da ekstrapolaciju izbegnemo.
11. Opišite osnovnu ideju interpolacije splajnom.
    - Podelimo interval N-1 ako je N broj tačaka. Na svakom intervalu imamo poseban splajn i na kraju dobijemo krivu koja prolazi kroz sve tačke.
12. Skicirajte interpolaciju splajnom. Šta je problem kod linearnog splajna?
    - Problem - koristi linearni polinom pa kriva može da ne bude dovoljno glatka i da onda ne prati baš precizno oblik podataka.

13. Skicirajte interpolaciju kvadratnim splajnom za 4 tačke – naznačite oblik polinoma za svaki od podintervala i koliko ukupno nepoznatih koeficijenata moramo da odredimo (slika).