

1 Besvara nedanstående teoretiska frågor koncist.
2
3 1. Lotta delar upp sin data i "Träning", "Validering" och "Test", vad används respektive del för?
4
5 Tex om 3 olika modeller används så används träning för att köra "fit metoden" på alla modeller.
6 Sen predikteras y värden fram med hjälp av valideringsdatats X. där de predikterade y värdena tillsammans med valideringsdatans y värden används för att se hur bra modellen är.
7 Modellerna jämförs mellan varandra och en modell väljs som är bäst och antas ge bäst prediktioner.
8 Den valda modellen kan tränas om på träning och validerings datat.
9 Sen så körs en prediktering på den valda modellen med test datats X värde.
10 Det predikterade y värdet tillsammans med test datats y värde används nu för att se hur bra modellen är igen och om den håller lika bra som innan på valideringsdatat och inte är overfitted eller andra problem.
11 Varför den predikteras med både validering och test är för att säkerställa att inte den slumpmässigt valda datan för validering endast är bra för träningsdatan.
12 Omträning med träning och validering är för att modellen ska få bättre resultat och kan generalisera bättre på osedd data.
13
14
15
16 2. Förklara (gärna med ett exempel): Ordinal encoding, one-hot encoding, dummy variable encoding.
17
18 Ordinal encoding
19 Är när värden har inbördes mening mellan varandra. Tex om värdena låg, mellan och hög ska användas så skulle de kunna användas så här
20
21 0 = låg
22 1 = mellan
23 2 = hög
24 Det skulle tex kunna betyda ju högre värde desto "högre" upp.
25
26 one-hot encoding
27 i det här fallet skulle 3 kolumner i datat skapas. En för varje värde.
28 så om de ligger i den här ordningen i datat låg, mellan, hög så skulle varje värde representeras av att det finns en etta i dess kolumner
29
30 låg mellan hög
31 låg = 1 0 0
32 mellan = 0 1 0
33 hög = 0 0 1
34
35 dummy variable encoding
36 Här tas en column bort och tex låg skulle kunna representeras av att alla kolumner har värdet noll.
37 Det skulle då se ut så här
38
39 mellan hög
40 låg = 0 0
41 mellan = 1 0
42 hög = 0 1
43
44 detta används för att undvika något som heter multicollinearity.
45
46
47 3. Göran påstår att datan antingen är "ordinal" eller "nominal". Julia säger att detta måste
48 tolkas. Hon ger ett exempel med att färger såsom {röd, grön, blå} generellt sett inte har någon inbördes ordning (nominal) men om du har en röd skjorta så är du vackrast på festen (ordinal) - vem har rätt?
49
50
51
52 Båda har lite rätt. Datat kan vara ordinalt eller nominalt men det är också någonting som måste tolkas.
53 Det är vi som tar fram modeller som måste avgöra om det finns samband mellan värdena i datat.

54 För den som inte tycker om rött så är kanske personen med röd skjorta inte vakrast på
festen.
55 Lite "context is king" på det här.
56
57
58 4. Läs följande länk:
<https://stackoverflow.com/questions/56107259/how-to-save-a-trained-model-by-scikit-learn>
59 (speciellt svaret från användaren som heter "sentence") som beskriver "joblib" och
"pickle".
60 Det är alltså ett sätt att spara modeller och innebär att man kan träna en modell och
sedan
61 återanvända den för att göra prediktioner utan att behöva träna om modellen. Detta
62 kommer ni ha nytta av om ni satsar på VG delen.
63 Svara på frågan: Vad används joblib och pickle till?
64
65 Båda används till att spara och ladda data men de har lite olika användningsområden
beroende på den datan som ska sparas/laddas.
66 pickle är mer generell och kan spara de flesta datastrukturerna men fungerar sämre för
större mängder data och kan ha lite säkerhetsbrister pga att den tillåter sån variation
av data att kunna laddas in.
67 joblib är rekommenderad för ML data (scikit-learn) och kan hantera större mängder data
bättre med kompression och optimering för numerisk data.