Xiaowen Chen inlämningsuppgifter av Machines Learning

1. Teoretiska frågor

1.1 Kalle delar upp sin data i ”Träning”, ”Validering” och ”Test”, vad används respektive del för?

Svar: Träning: Träningsdata används för att träna modellen, som använder dessa data för att lära sig känna igen mönster och samband i daten. Detta är huvuddelen av datapartitionering som gör att modellen kan justera sina interna parametrar baserat på in- och utdata som tillhandahålls för förutsägelse- eller klassificeringsändamål.

Validering: Valideringsdata används för att justera modellparametrar och förhindra överanpassning. Överanpassning är när en modell presterar bra på träningsdata men presterar dåligt på nya, osynliga data. Valideringsdata hjälper till att utvärdera en modells förmåga att prestera på okända data och säkerställer att modellen gör mer än att bara komma ihåg träningsdata. Under modellträning, använd valideringsdata för att testa modellens prestanda och justera modellkonfigurationen baserat på prestanda.

Test: När modellträning och validering är klar, använd testdata för att utvärdera modellens slutliga prestanda. Denna del av data är helt osynlig för modellen under träning och möjliggör därför en opartisk bedömning av modellens prestanda i verkliga tillämpningar. Testdata låter oss att se hur exakt och generell din modell är när du hanterar nya data. Testresultat utvärderar vanligtvis modellens prestanda genom en serie poängindikatorer.

1.2 Julia delar upp sin data i träning och test. På träningsdatan så tränar hon tre modeller; ”Linjär Regression”, ”Lasso regression” och en ”Random Forest modell”. Hur skall hon välja vilken av de tre modellerna hon skall fortsätta använda när hon inte skapat ett explicit ”valideringsdataset”?

Svar: Om Julia delar upp sin data bara i träning och test, utan ett särskilt valideringsdataset, kan hon välja mellan sina tre modeller (Linjär Regression, Lasso Regression och Random Forest) genom att använda korsvalidering. Detta innebär att hon delar upp sin träningdata i flera delar, tränar modellen på vissa av dessa delar och testar på en del som inte används för träning. Genom att göra detta flera gånger och beräkna genomsnittet av modellens prestanda över dessa tester får hon en bra uppfattning om hur varje modell presterar. Den modell som presterar bäst på dessa valideringstester är den hon bör använda. Viktigt är att inte använda testdatan för att välja modell, utan bara för att testa den valda modellens slutliga prestanda.

1.3 Vad är ”regressionsproblem? Kan du ge några exempel på modeller som används och potentiella tillämpningsområden?

Svar: Ett regressionsproblem inom maskininlärning handlar om att förutsäga ett kontinuerligt värde baserat på en eller flera oberoende variabler. Det skiljer sig från klassificeringsproblem som syftar till att förutsäga en kategori eller klass tillhörighet.

Exempel på regressionsmodeller inkluderar:

Linjär regression: Den enklaste formen av regression som försöker passa en linje genom datapunkterna så att summan av kvadraten på avstånden från datapunkterna till linjen minimeras.

Polynomisk regression: En form av linjär regression där relationen mellan oberoende variabel och beroende variabel modelleras som ett polynom.

Lasso regression (Least Absolute Shrinkage and Selection Operator): En typ av linjär regression som använder regularisering. Lasso tenderar att göra vissa koefficienter exakt noll, vilket gör den användbar för variabelurval och för att skapa enklare, mer tolkningsbara modeller.

Ridge regression: Liknar lasso regression men använder en annan typ av straff för regularisering, vilket leder till att koefficienterna minskar men sällan blir exakt noll.

Random Forest för regression: En ensemblemetod som använder flera trädmodeller för att göra en mer stabil och noggrann förutsägelse.

Potentiella tillämpningsområden för regressionsmodeller inkluderar:

Prissättning av fastigheter: Förutsäga priset på ett hus baserat på dess egenskaper som storlek, läge och antal rum.
Aktiemarknadsförutsägelser: Använda historiska data för att förutsäga framtida aktiekurser eller marknadstrender.
Energiförbrukning: Förutsäga en byggnads energiförbrukning baserat på olika faktorer såsom väderförhållanden och byggnadens storlek.
Försäljningsprognoser: Estimera framtida försäljning baserat på tidigare försäljningssiffror och andra relevanta faktorer som marknadsföringsutgifter.

1.4 Hur kan du tolka RMSE och vad används det till:

$$RMSE = \sqrt{\sum_i (y_i - \hat{y}_i)^2}$$

Svar: RMSE (Rotmedelkvadratfelet) är ett sätt att mäta fel i förutsägelser. Det visar hur stor skillnaden är mellan vad vi förutsäger och vad som faktiskt händer, i genomsnitt.

(ŷ_i) är förutsagt värde
(y_i) är riktiga värdet

Användning:
Den visar hur mycket våra förutsägelser i snitt avviker från de verkliga värdena.
Ett mindre RMSE-värde betyder att modellen förutsäger mer exakt.
Man kan använda RMSE för att jämföra hur bra olika modeller förutsäger.
Så, RMSE hjälper oss att förstå hur långt bort våra modellers förutsägelser brukar vara från det som faktiskt sker.

1.5 Vad är ”klassificieringsproblem? Kan du ge några exempel på modeller som används och potentiella tillämpningsområden? Vad är en ”Confusion Matrix”?

Svar: Klassificeringsproblem hänvisar till uppgiften i maskininlärning att tilldela data till fördefinierade kategorier eller klasser. Detta skiljer sig från regressionsproblem, som förutsäger kontinuerliga värden snarare än diskreta kategorier.

Exempel på klassificeringsmodeller inkluderar:

Logistisk regression: Även om den har "regression" i namnet, används logistisk regression för binära klassificeringsproblem.

Beslutsträd: Dela upp data i kategorier genom en serie regler.

Support Vector Machines (SVM): Hitta den bästa gränslinjen som kan separera olika kategorier längst bort.

Random Forest: Konstruera flera beslutsträd och kombinera deras resultat för att förbättra klassificeringsnoggrannheten.

Neurala nätverk: simulerar nätverksstrukturen hos mänskliga hjärnneuroner och kan hantera komplexa klassificeringsproblem.

Potentiella användningsområden för klassificeringsmodeller inkluderar:

Skräppostfiltrering för e-post: Bestäm om ett e-postmeddelande är skräppost.
Kundsegmentering: Dela in kunder i olika grupper utifrån shoppingbeteende eller preferenser.
Medicinsk diagnos: Att fastställa vilken sjukdom en patient kan ha baserat på hans eller hennes data.
Bildigenkänning: som att automatiskt identifiera ansikten på foton.
Confusion Matrix är en specifik tabelllayout som används för att visualisera prestandan för en klassificeringsmodell, speciellt när det finns fler än två kategorier. Den visar förhållandet mellan de faktiska kategorierna och kategorierna som förutsägs av maskininlärningsmodellen, vilket gör att vi explicit kan se antalet korrekta och felaktiga förutsägelser av modellen för varje kategori.

Förvirringsmatrisen innehåller:

Sanna positiva (TP): Antalet gånger som modellen korrekt förutsäger en positiv klass.
Falska positiva (FP): Antalet gånger som modellen felaktigt förutsäger en positiv klass.
Sanna negativa (TN): Antalet gånger som modellen korrekt förutsäger den negativa klassen.
Falska negativa (FN): Antalet gånger som modellen felaktigt förutsäger en negativ klass.
Genom förvirringsmatrisen kan vi se inte bara modellens noggrannhet, utan också modellens specifika prestanda när det gäller att förutsäga varje kategori.

1.6 Vad är K-means modellen för något? Ge ett exempel på vad det kan tillämpas på.

Svar: K-means är en typ av oövervakad maskininlärningsmodell som används för klusteranalys. Modellen försöker dela in datan i K olika kluster där varje datapunkt tillhör det kluster med närmaste medelpunkt.

Ett exempel på tillämpning är kundsegmentering i marknadsföring. Genom att använda K-means kan ett företag gruppera kunder baserat på köpbeteende eller andra egenskaper för att skräddarsy sin marknadsföringsstrategi för varje segment.

1.7 Förklara (gärna med ett exempel): Ordinal encoding, one-hot encoding, dummy variable encoding. Se mappen ”l8” på GitHub om du behöver repetition.

Svar:

Ordinal encoding är när du har saker i en bestämd ordning och du ger dem nummer efter den ordningen. Till exempel, om du har storlekar som "Liten", "Mellan" och "Stor", kan du ge dem nummer 1, 2 och 3.

One-hot encoding är när du gör en ny kolumn för varje kategori och sätter en etta (1) för kategorin som gäller och nolla (0) för de andra. Om du har färger som "Röd", "Grön" och "Blå" så får "Röd" kolumnen [1, 0, 0], "Grön" får [0, 1, 0], och "Blå" får [0, 0, 1].

Dummy variable encoding är liknande one-hot encoding, men du använder en kolumn mindre för att undvika problem. Med färger igen, om "Röd" är standard så behöver du bara två kolumner. "Röd" blir [0, 0], "Grön" blir [1, 0] och "Blå" blir [0, 1].

Det här hjälper datorn att förstå och räkna på kategorier som om de vore siffror.

1.8 Göran påstår att datan antingen är ”ordinal” eller ”nominal”. Julia säger att detta måste 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?

Svar: Både Göran och Julia har på sätt och vis rätt. Göran påpekar att data kan vara "ordinal" (med en ordning) eller "nominal" (utan ordning). Julia menar att hur man tolkar datan kan variera. Färger som {röd, grön, blå} är vanligtvis "nominala" eftersom de inte har en naturlig ordning. Men i Julias exempel, där en röd skjorta innebär att man är vackrast på festen, ges färgerna en ordning baserat på skönhet, vilket gör dem "ordinala". Så, det beror på sammanhanget huruvida datan är ordinal eller nominal.

1.9 Vad är Streamlit för något och vad kan det användas till?

Svar: Streamlit är ett bibliotek för Python som gör det enkelt att skapa interaktiva webbappar för data och maskininlärning. Du kan snabbt visa data, köra modeller och interagera med informationen.

Datautforskning och visualisering.

Bygga interaktiva dashboards för att visa upp dataanalys och maskininlärningsresultat.

Dela prototyper av maskininlärningsmodeller med icke-tekniska intressenter.

Utveckla och dela verktyg inom data science och maskininlärning för intern eller extern användning.