Databas från CSV-fil*

Matthew H. Motallebipour

April 30, 2024

1 Teoretiska frågor

1.1 Träning, validering, test

Träning används för att skapa en estimator, validering används frekvent för att testa resultatet av funktionsuppskattningen som delmål. Test-data används däremot efter att en slutlig estimator är färdigställt och redo att ”deploy”as.

1.2 Jämförelse av modeller

Om hon inte kan använda funktionen för uppdelning av data i träning, validering och test kan hon kanske använda funktioner som finns för korsvalidering av data, om inte även det är förbjudet att använda. Dessa funktioner delar upp data i flera del mängder under körning, där varje delmängd används som valideringsdata i varje runda och de andra delar används till träning av modellen. Efter att alla delmängder är använda beräknar prestandan för den producerade modellen som möjliggör jämförelse mellan de nämnda funktionerna.

1.3 Regression

Regressionsproblem handlar om att hitta en funktion/ hyperplan som kan approximera majoriteten av punkter i en datamängd i ett N-dimensionellt rum. Värden som fås från funktionen är ofta reella. Exempel på modeller som genererar sådana funktioner:

- Linjär regression: ofta används det inom ekonomi och medicin, där sambandet mellan två fenomen betraktas.
- Polynomial regression: också används inom ekonomi, där man baserad på flera faktorer bestämmer om företag kommer att vara lönsamt året därpå; se kursboken sidan 36.
- Beslutsträd: används många gånger där en beroende faktor är bestäms av flera andra faktorer. Trots att den genererade hyperplanen är ”fyrkantig” används det ändå ofta för det ändamålet.
- Random Forest: använder sig av flera beslutsträd och klassificeras därför som en ensemble-metod.

1.4 RMSE

RMSE är enkelt sagt ett medelvärde för avståndet mellan riktiga värden och deras uppskattningar som är framställd mha en datormodell. Nogrannare är detta värde beräknat genom roten (R) ur medelvärdet (M) av summan (S) av alla data punkters avvikelser från den uppskattade hyperplanen (E) i kvadrat.

*https://github.com/Zomnipotential/Power_BI_Quiz_I

1.5 Klassificering

Klassificeringsproblem är där man önskar konstruera modeller som delar datapunkter i olika grupper eller kategorier baserat på deras olika egenskaper. Klassificering, tvärtemot regressionsproblem, framställer oftast diskreta resultat, även om de genererade talen kan vara reellvärda. Skillnaden mot regression är att i regression är själva den predikterade hyperplanen en uppskattning av datapunkterna medan i klassificering är hyperplanen en uppskattning av gränsen mellan datapunkterna i de olika kategorierna.

Dessa använda t ex för för olika igenkänningsproblem, såsom igenkänning av handskrivna bokstäver, klassificering av växter och djur osv. Exempel på modeller som löser sådana problem är

- Logistisk regression: i sin enklaste, typiska form används den för binär klassificering
- Olika typer av neurala nätverk, såsom faltningsneuronnät (CNN),
- Random Forest Classifier
- Support Vector Classifier

Confusion matrix bygger på relationen mellan antalet riktiga datapunkter och deras uppskattningar. Raderna fördelas mellan negativa och positiva riktiga datapunkter och kolumnerna fördelas mellan negativa och positiva uppskattningar av samma datapunkter. Resultatet ser ut som följer:

		Prediction outcome		
		p	**n**	**total**
actual value	**p′**	True Positive	False Negative	P′
	n′	False Positive	True Negative	N′
	total	P	N	

1.6 K-means

K-means är en modell för klassificering av data utan tillgång till labels, en så kallad oövervakad inlärningsalgoritm. Till exempel gruppering av kunder baserad på deras intressen eller inhandlade varor.

1.7 Data Encoding

När det finns features/ kolumner som innehåller kategorisk data kan de inte användas i ML-modeller. Därför måste de omformas, dvs ersättas av siffror. Detta kallar vi kodning. Det finns tre typer av kodning.

1.7.1 Ordinal kodning

Görs när man inser att data har en inbördes rangordning. Denna rangordning representerar vi med siffror. T.ex. kan barn, tonårig, ung, mellanålder och gammal ersättas av siffrorna 1, 2, 3, 4 och 5. Detta kräver endast en kolumn att ersätta den ordinarie kolumnen.

ung	3
gammal	5
barn	1
barn	1
barn	1
mellanålder	4
barn	1
mellanålder	4
gammal	5
gammal	5
barn	1
ung	3
barn	1
gammal	5

1.7.2 I one-hot encoding

får vi lägga till en extra kolumn i tabellen för varje distinkt värde i den ordinarie kolumnen. I dessa nya kolumner markerar 1 att respektive kategori finns på den raden i den ordinarie kolumnen. Så vårt tidigare exempel får ett nytt utseende

	barn	tonåring	ungdom	mellanålder	gammal
ung	0	0	1	0	0
gammal	0	0	0	0	1
barn	1	0	0	0	0
tonåring	1	1	0	0	0
barn	1	0	0	0	0
mellanålder	0	0	0	1	0
barn	1	0	0	0	0
mellanålder	0	0	0	1	0
gammal	0	0	0	0	1
gammal	0	0	0	0	1
barn	1	0	0	0	0
ung	0	0	1	0	0
barn	1	0	0	0	0
gammal	0	0	0	0	1

1.7.3 Dummy kodning

Utförs med dummy variabler och sparar en kolumn genom att ta bort det självklara alternativet att där alla andra kolumner innehåller nollor ska det antas att den ordinarie kolumnen antar det alternativ för vilket det inte finns någon kolumn. I den tabell som följer tar vi bort barn-kolumnen och märker att vi ändå kan se var värdet i den ordinarie kolumnen ska vara 1 (barn).

	tonåring	ungdom	mellanålder	gammal
ung	0	1	0	0
gammal	0	0	0	1
barn	0	0	0	0
tonåring	1	0	0	0
barn	0	0	0	0
mellanålder	0	0	1	0
barn	0	0	0	0
mellanålder	0	0	1	0
gammal	0	0	0	1
gammal	0	0	0	1
barn	0	0	0	0
ung	0	1	0	0
barn	0	0	0	0
gammal	0	0	0	1

1.8 Ordinal eller nominal

Julia har rätt. Att vara ordinal eller nominal beror helt på vår underliggande mening med det data vi har samlat i tabellen. Även siffror kan användas där det inte finns någon egentlig inbördes rangordning. Ett bra exempel är siffrorna på tröjorna som basketspelare brukar ha på sig.

1.9 Streamlit

Är ett ramverk för att bygga Python-applikationer. Detta kan användas till att göra applikationen tillgänglig genom en server.

I denna rapport kommer tabeller att annoteras med **fet stil** kolumner med *kursiv text*.

1.10 Databasen

Databasen, som presenterats i form av csv-filer är extraherade från en huvuddatabas under namnet AdventureWorks2022 och innehåller 6 tabeller som är hopkopplade i form av 3 grupper

1. **DimProduct**, **FactInternetSale**, och **DimSalesTerritory**
    - **FactInternetSale** som är kopplad till **DimSalesTerritory** genom *SalesTerritoryKey*
    - **FactInternetSale** som är kopplad till **DimProduct** genom *ProductKey*
2. **DimProductSubcategory** och **ProductCategory** genom *ProductCategoryKey*
3. **DimDate** är en ensamstående tabell

För att koppla ihop samtliga tabeller i en enda grupp, en så kallad data modell, söker vi och hittar *ProductSubcategoryKey* som gemensamt nyckelord mellan **DimProduct** och **DimProductSubcategory**, samt *DateKey* i **FactInternetSale** och **DimDate**. Resultatet ser ut som följer i bilden överst på nästa sida och påminner om den så kallade datamodellen snöflinga.

1.11 Rapporten

Rapporten är på begäran bestående av tre sidor, där den första sidan innehåller bolagets logotyp samt data i stora drag, den andra sidan innehåller intressanta trender som vi hittade i vår data, och den tredje sidan visar värdet av den totala försäljningen delad över olika regioner i världen.

1.11.1 Första sidan – Introduction

Visar företagets logotyp och hur bolaget har presterat under hela sin historia

- Den totala levererade beställningar Count of *SalesAmount*,
- Deras totala värde Sum of *SalesAmount*,
- Den mest sålda detaljprodukten Top Selling Category, som är en Measure i **FactInternetSale**. I detta fall är measure beräknad som den största utav de aggregerade värdena för alla enskilda, unika detaljprodukter i *EnglishProductSubCategoryName*.
- Den minst sålda detaljprodukten Least Selling Category, som också är en Measure i **FactInternetSale** och beräknad på samma sätt som ovan, där det minsta värdet är använt.
- Därefter följer *SalesAmount* för *ShipDate*, *DueDate*, och *OrderDate*.

1.11.2 Andra sidan – Sales Trend

Vi har försökt titta på den cykliska trenden i data, nämligen hur de olika månaderna påverkar försäljningen i stort.

- Överst på sidan presenteras den totala försäljningens värde under alla året bolaget har opererat.
- Under denna, och på höger sidan ser vi en uppdelning av samma historik, en uppdelning av de tre huvudprodukterna, accessories, bikes och clothing.
- Mittemot dessa kan man konstatera den cykliska försäljningen månadsvis. Detta visar att logiskt nog ökar försäljningen från januari fram till och med juni och därefter faller försäljningen tills den återigen ökar avsevärt under december månad, speciellt med tanke på julhandeln, då människor förbereder sig inför den stundande varma perioden.

1.11.3 Tredje sidan – Sales Amounts

Här visas den totala försäljningen per land, region, och kontinent i en interaktiv karta. För att överskådliggöra försäljningen för har även värden för de tre kontinenterna markerats som tre kort på topologiskt relevanta platser runtom kartan.

1.12 Övrigt

En mörkare färg har valts för hela rapporten för att undvika alltför mycket utstrålning av ljus som kan kännas besvärligt för ögat.