Applicerad AI Inlämning 2

Teknikhögskolan

January 2022

1 Introduction

Inlämningsuppgiften innefattar klassifikation, regression och klustring. Det rekommenderas starkt att göra en ordentlig exploratory data analysis för att leta efter korrelationer, imbalanser, onödiga features och eventuella NA-värden i dataseten. Det fungerar alltid att ”tuta och köra”, men för bättre resultat bör ordentlig pre-processing göras. Med det sagt finns det dock inga krav i form utav ”uppnå x i score y” som examinerande moment.

2 Klassifikation

Ladda in det givna datasetet och gå loss!

För G:

2.1 Dataanalys

Börja med att utreda datasetet och ta reda på antal klasser och om det finns någon form av klassimbalans eller NA-värden.

2.2 Modell

Använd valfri klassificeringsalgoritm eller modell för att få ett så högt F_1-score som möjligt (var för sig) för ”weighted”-parametern ifrån den här länken..

För VG:

2.3 Dataanalys

Utred om det finns någon form utav klassimbalans och åtgärda isåfall den. Avgör om denna modifikation förbättrar resultaten.

Använd PCA för att avgöra hur många features som är överflödiga. Gör en plot på den förklarade variansen samt en heatmap över korrelationerna i datan.

Avgör om dimensionsreducering med PCA förbättrar dina resultat.

Avgör om feature selection behövs och om det förbättrar dina resultat. Länk:

2.4 Modell

Maximera F_β-score för $\beta = 0.5, 1, 2$ för minoritetsklassen. Ni ska alltså göra separata approacher för alla dessa fall.

Använd två olika, valfria klassificeringsalgoritmer och jämför resultaten från dessa.

3 Regression

För G:

Ta bort NA-värden från datasetet.

Gör en korrelationsplot i t.ex. seaborn för att avgöra vilka par av features som har högst korrelation. Ta bort en feature från detta par.

Gör en fullständig one-hot encoding för datasetet

Använd en valfri algoritm/modell för regression.

För VG: Ta bort NA-värden, encoda stringvariabler och

Använd två olika, valfria algoritmer/modeller för regression.

Plotta en korrelationsmatris

Använd feature importance/feature selection för att plocka bort de 5 features som har lägst inverkan på slutresultatet. Om du använder LASSO-regression för feature selection - hitta först det bästa α-värdet och ta bort alla features som är 0.

Plotta distributionen av era targetvariabler i t.ex. ett histogram för att få en överblick. Pre-processa därefter dina targetvariabler genom att göra en log(1+x)-transform med

$$val_i \rightarrow ln(1 + val_i) \tag{1}$$

och se på distributionen igen. Gör en fit av modellen på den transformerade datan och använd för predictions metoden

$$pred_i -> exp(pred_i) - 1 \tag{2}$$

Dessa operationer finns i t.ex. numpy med np.expm1, np.log1p. Det här steget kan med ibland användas för pre-processing av regressionsdata som har en skev distribution som avviker mycket från en normaldistribution och/eller har många outliers. Utred om det här steget ger en förbättring av era modeller.

4 Klustring

Det dataset som ges har inga labels - det är upp till er att med hjälp av klustringsalgoritmer och tolkning av labels försöka ansätta korrekt antal egentliga klasser. Ansätt klasstillhörighet till datasetet med hjälp av klustringen (dvs - klustringen ger en klass till varje observation som kan användas vid visualiseringar).

För G:

Använd K-means clustering för att bestämma antalet kluster. Plotta dina resultat med t.ex. pairplot från seaborn.

För VG: För alla klustringsalgoritmer ska ni använda en majoritetsomröstning från 3 olika klustringsmetrics (t.ex. Calinski Harabasz, Davies Bouldin, Silhouette Coefficient, Bayesian Information Criterion, Elbow Rule) för att bestämma antal kluster. Fundera på om du behöver skala variabler för varje enskild klustring.

Dimensionsreducera till 2 dimensioner och plotta datan med hjälp av t-SNE, UMAP och PCA var för sig. Ser det ut att vara distinkta grupper?

Använd Gaussian Mixture Models som klustringsalgoritm och avgör vilken kombination av antal kluster och kovarianstyp som ger bäst klustring. **OBS!** Bara BIC-score behöver tas hänsyn till i denna klustring.

Använd DBSCAN som klustringsalgoritm och säkerställ att minst 85% av datan är klassificerad som någonting annat än brus (noise). (Detta kan exempelvis göras med en grid search av parameterkombinationer av ϵ och antal core points).