Political Data Science

Lektion 10

Superviseret læring I

Opgave 1

Indlæs datasættet `valg2015.rdata`, som er tilgængeligt i mappen `data` på fagets GitHub. Datasættet renset udgave af Valgundersøgelsen 2015, hvor et subset af 27 variable indgår. Brug eksempelvis pakken `rio` til at importere data med funktionen `import()`.

Kodebogen `valg2015.txt` er tilgængelig samme sted. Dan dig et hurtigt overblik over data.

Opgave 2

Del datasættet op i et træningssæt og et testsæt, hvor træningssættet består af 80 % af data og testsættet består de af resterende 20 %. Brug evt. koden herunder:

```
set.seed(42)
index <- sample(nrow(valg2015), nrow(valg2015)*0.8)

valg2015_train <- valg2015[index, ]
valg2015_test <- valg2015[-index, ]
```

Bruger du koden ovenfor, forklar da kort logikken i koden.

Opgave 3

Træn en lineær OLS-model og en logit-model på træningssættet [hint: `lm()` og `glm()`], hvor variablen `partivalg` er outcome og du bruger alle de øvrige variable som prædiktorer. Husk at sætte argumentet `family = "binomial"` for logit-modellen.

Print dine outputs med `summary()`. Kommenter kort på de to outputs.

Opgave 4

Prædikter outcome i testsættet med hver af dine to modeller, hhv. OLS og logit [hint: `predict()`]. Husk at specificere argumentet `type = "response"` for logit-modellen. Gem de prædikterede outcomes i hver deres kolonne i testsættet.

Lav to plots, hvor du har det prædikterede partivalg på x-aksen og det faktiske partivalg på y-aksen for hver af de to modeller. Ser det ud til, at modellerne har høj eller lav prædiktions-performance baseret på de to plots?

Opgave 5

Lav to nye kolonner i testsættet, hvor du transformerer de prædikterede partivalg for hver af dine to modeller fra en kontinuert variabel til en kategorisk variabel, der tager værdierne 0 og 1, afhængigt af om de prædikterede partivalg er under/over gennemsnittet af det prædikterede outcome.

Lav en confusion-matrix for hver af de to modeller, hvor du en tabel med de faktiske outcomes og prædikterede outcomes [hint: `table()`].

Beregn accuracy, dvs. andelen af korrekte klassifikationer, for hver af de to modeller.

Hvilken model har den højeste accuracy? Begrund, om du vil vurdere modellernes accuracy til at være høj, lav eller midt imellem givet hvad du ved om outcome-variablen `partivalg`. Er accuracy et godt mål i sammenhængen her - hvorfor, hvorfor ikke?

Opgave 6

Installer og load pakken `pROC`, som du kan bruge til at plotte ROC-kurver og beregne AUC.

Gem et ROC-objekt med datapunkter for en ROC-kurve for hver af dine to modeller ved at bruge funktionen `roc()`. Funktionen tager to inputs, hhv. et faktisk outcome og et prædikteret outcome. Det kan se ud som herunder:

```
ROC_ols <- roc(df$y, df$y_hat)
```

Plot ROC-kurver for hver af de to ROC-objekter, fx ved brug af funktionen `plot()`, der kan tage et ROC-objekt som input.

Brug funktionen `auc()` til at beregne AUC for hver af de to modeller. Funktionen kan også tage et ROC-objekt som input. Hvilken af dine modeller leverer de bedste prædiktioner? Er der stor forskel?

Bonus-opgave

Opstil en logit-model og prædiktér outcome-variablen `partivalg` på baggrund af alle de øvrige variable i hele det oprindelige datasæt `valg2015`. Brug 5-fold cross-validation til at estimere modellens out-of-sample performance. Brug eksempelvis funktionen `kWayCrossValidation()` fra pakken `vtreat` og skriv et loop, hvor logit-modellen på skift trænes på `k - 1` folder og prædikterer om det udeladte fold, indtil du har prædikteret om alle folder. Er modellens AUC højere eller lavere end da du brugte validation set tilgangen ovenfor? Er resultatet som forventet?