KAUNO TECHNOLOGIJOS UNIVERSITETAS

MATEMATIKOS IR GAMTOS MOKSLŲ FAKULTETAS

Vytautas Janilionis

Studijų modulio
P160M129 DAUGIAMATĖS STATISTINĖS ANALIZĖS MODELIAI

Projekto užduotis

KAUNAS, 2023

DARBO TIKSLAS

Panaudojus statistikos metodus bei pasirinktas programines priemones (R, SAS, Python) atlikti dėstytojo pateiktų daugiamačių duomenų statistinę analizę: sudaryti duomenų analizės modelius, sukurti atitinkamas programas, atlikti pateiktų duomenų analizę, parinkti duotiems duomenims geriausius ir interpretuojamus modelius, parengti modelių sudarymo proceso ataskaitą, pateikti gautų modelių kokybės vertinimą, interpretaciją ir išvadas.

1 DALIS. TIESINĖS REGRESIJOS MODELIS

DUOMENYS

Moodle patalpintas duomenų failas **Reg21.csv**. Jame yra nepriklausomi kiekybiniai kintamieji X1, X2, ..., X110, kokybiniai kintamieji C1 ir C2, priklausomi kintamieji Y1, Y2, ..., Y45.

UŽDUOTIS

Sudarykite geriausią interpretuojamą tiesinės regresijos modelį Y_N prognozavimui ir nepriklausomų kintamųjų (regresorių) įtakos Y_N vertinimui, čia N yra varianto numeris.

Projektą rengia vienas studentas arba komanda (max. du 2 nariai). Jeigu projektą rengia du studentai, tuomet N = mažiausiam projekto komandos narių varianto numeriui (kiekvieno studento varianto numeris N pateiktas Moodle kurso medžiagos pradžioje). Apie sprendimą sudaryti komandą studentai informuoja užsiregistruodami PROJEKTO KOMANDŲ SĄRAŠE, kuris yra Moodle skyrelyje PROJEKTAS iki gegužės 14 d. Jeigu studentas rengia projektą vienas, registruotis nereikia.

Nurodymai modelio sudarymui ir ataskaitos rengimui:

1. Modelio kūrimui galima laisvai pasirinkti tinkamas R, SAS, Python procedūras bei funkcijas.
2. Modelio sudarymui naudokite šiuos nepriklausomus kintamuosius (regresorius): X1, X2, ... , X110 (arba jų transformacijas, jeigu reikia tiesinti); kokybinius kintamuosius C1, C2; kokybinio kintamojo C1 sąveikas (sandaugas) su kiekybiniais kintamaisiais X_{N+40},.X_{N+41},…,X_{N+45}; kiekybinių kintamųjų X_{N+30}, X_{N+31},…, X_{N+35} visas tarpusavio porines sąveikas (sandaugas po du), čia N - varianto numeris.
3. Modelio sudarymui panaudokite bent tris kintamųjų (regresorių) atrankos metodus ir parinkite tinkamiausius atrankos metodų parametrus. Bent vienas atrankos metodas turi būti su reguliarizacija (Lasso (L_1), Ridge (L_2), Elastic net (L_1, L_2) ir t.t., parinkite tinkamiausias reguliarizacijos parametrų reikšmes). Palyginkite gautus modelius tarpusavyje ir išrinkite geriausią. Pateikite modelių atrankai naudotus modelių kokybės

rodiklius {ASE (paklaidų kvadratų vidurkis), MSE (vidutinė kvadratinė paklaida), RMSE (vidutinis kvadratinis nuokrypis), RSQ (apibrėžtumo koeficientas R^2), ADJRSQ (pataisytasis apibrėžtumo koeficientas R^2adj), AIC, AICC, BIC, SBC (Akaikės, Akaikės pataisytasis, Bajeso, Švarco ir Bajeso informaciniai kriterijai), CP (Mallows' Cp) ir kitus}. Pastaba: ataskaitoje be jūsų naudotų kriterijų reikės pateikti AIC ir RSQ. Atrenkant modelius rekomenduojama panaudoti ir kryžminį patvirtinimą. Atrenkant geriausią modelį stenkitės parinkti interpretuojamą ir mažiau parametrų turintį modelį, neužmirškite patikrinti ar modelis tenkina prielaidas. Pastaba: geriausias modelis, tenkinantis prielaidas (arba tik su labai „švelniais" pažeidimais), turėtų gautis su 5-12 regresorių.

4. Ataskaitą parenkite pagal Moodle pateiktą šabloną. Ataskaitoje būtina pateikti programos kodą, modelio sudarymo etapų pagrindinius rezultatus (paaiškinti su kokiomis problemomis susidūrėte ir kaip jas sprendėte, pateikti sudarytų modelių palyginimo rezultatus (būtina pateikti AIC, RSQ ir kitų naudotų kriterijų vertes), geriausio atrinkto modelio regresijos lygtį, pateikti išvadas apie regresijos modelio prielaidų tenkinimą, apibūdinti atrinkto modelio privalumus ir trūkumus, pateikti išvadas apie imties ir populiacijos regresijos lygties koeficientus prie visų regresorių įrašytų į regresijos lygtį, išvadas apie populiacijos Y_N individualios reikšmės ir Y_N vidurkio prognozę laisvai pasirinktame taške.

2 DALIS . LOGISTINĖS REGRESIJOS MODELIS

DUOMENYS

Moodle patalpintas duomenų failas **LogReg.csv**. Jame yra nepriklausomi kiekybiniai kintamieji X1, X2, ..., X100, kokybiniai kintamieji C1 ir C2, priklausomi kokybiniai kintamieji Y1, Y2, ..., Y45.

UŽDUOTIS

Sudarykite geriausią interpretuojamą logistinės regresijos (logit) modelį $\ln(P(Y_N=1)/P(Y_N=0))$ prognozavimui ir nepriklausomų kintamųjų (regresorių) įtakos vertinimui, čia N yra varianto numeris.

Nurodymai modelio sudarymui ir ataskaitos rengimui:

1. Modelio kūrimui galima laisvai pasirinkti tinkamas R, SAS, Python procedūras bei funkcijas.

2. Modelio sudarymui naudokite šiuos nepriklausomus kintamuosius (regresorius): X1, X2, ...,X100; kokybinį kintamąjį C2; kokybinio kintamojo C1 sąveikas (sandaugas) su kiekybiniais kintamaisiais X_{40},.X_{41},…,X_{80}.
3. Modelio sudarymui panaudokite bent tris kintamųjų (regresorių) atrankos metodus ir parinkite tinkamiausius atrankos metodų parametrus. Bent vienas atrankos metodas turi būti su reguliarizacija (Lasso (L_1), Ridge (L_2) ir t.t., parinkite tinkamiausias reguliarizacijos parametrų reikšmes). Palyginkite gautus modelius tarpusavyje ir išrinkite geriausią. Pateikite modelių atrankai naudotus modelių kokybės rodiklius {AIC, AICC, SBI (Akaikės, Akaikės pataisytasis, Švarco ir Bajeso informaciniai kriterijai) ir kitus}. Pastaba: ataskaitoje be Jūsų naudotų kriterijų reikės pateikti SBI kriterijaus vertes. Atrinktiems modeliams parinkite klasifikavimo slenkstį ir pateikite bendrą modelio klasifikavimo tikslumą (Accuracy), specifiškumą (Specificity), jautrumą (Sensitivity) ir t.t. Neužmirškite įvertinti modelio tinkamumą duomenims. Atrenkant modelius rekomenduojama panaudoti ir kryžminį patvirtinimą. Atrenkant geriausią modelį stenkitės parinkti interpretuojamą ir mažiau parametrų turintį modelį.
4. Ataskaitą parenkite pagal Moodle pateiktą šabloną. Joje būtina pateikti programos kodą, modelio sudarymo etapų pagrindinius rezultatus (paaiškinti su kokiomis problemomis susidūrėte ir kaip jas sprendėte, pateikti sudarytų modelių palyginimo rezultatus (būtina pateikti ir komentuoti SBI ir kitus kriterijus, bendrą modelio klasifikavimo tikslumą, specifiškumą , jautrumą, ir t.t.), parašyti atrinkto modelio logistinės regresijos lygtį, pateikti išvadas apie logistinės regresijos modelio tinkamumą duomenims, pateikti išvadas apie imties ir populiacijos logistinės regresijos lygties koeficientus ir jų įtaką tikimybei P(Y=1) bei galimybių santykiams (odds ratio).

Papildymas. Tie kurie pasirinks projekto vykdymui SAS, rekomenduojama naudoti:

- regresinės analizės procedūras: GLMSELECT , HPREG, GLM, REG;
- logistinės regresinės analizės procedūras: LOGISTIC, HPLOGISTIC (abiejų procedūrų operatorius **model** yN (Event='1') = regresorių sarašas / CTABLE LACKFIT LINK=LOGIT kiti parametrai;);
- nuoroda į kurso biblioteką, kurioje saugomi SAS formato duomenų failai reg21 ir logreg: libname duomenys "/home/pakeisti_į_student_SASID/my_shared_file_links/vytautasjanilio0/DSAM_PR";