Examensarbete 30 hp i matematik/data science

Fusion of multi-level categorical variables using lasso regression

Region Västerbotten arbetar för att god hälsa och hållbar utveckling ska stärka varandra. Vi tar ansvar för en jämlik välfärd och för att forskning och innovation ger resultat. För att klara morgondagens utmaningar vill vi i Region Västerbotten möjliggöra mer tid till vård genom smarta servicelösningar och digitalisering.
Centrum för informationsteknik och medicinsk teknik har som syfte att arbeta långsiktigt och modernt med rätt förmågor för att leverera och realisera regionens behov av utveckling, systemförvaltning, verksamhetsnära teknik och IT-infrastruktur. Vi är ca 300 anställda och finns representerade på länets tre sjukhus - i Lycksele, Skellefteå och Umeå. Vi bidrar med all form av teknik för att stötta hela regionen och i samverkan med dem bidrar vi till digitaliseringen av vården. Vi jobbar medvetet för att skapa ett högt värde för individen och en stolthet över det vi gör tillsammans.

Bakgrund

AMHOS är ett paraplyprojekt på Region Västerbotten med målsättning att integrera **A**I och **M**askininlärning i **H**älso- **O**ch **S**jukvård. Delprojekten inom AMHOS innefattar att konstruera maskininlärningsmodeller med tillhörande pipelines för att prediktera vårdtid och -tyngd på postoperativ avdelning (PostOP) samt intensivvårdsavdelning (IVA). Prediktionerna görs med kliniska patientdata, vilka måste genomgå lämplig förbehandling för att reducera dimensionaliteten av datan och öka den *effektiva* komplexiteten.

En utmaning med detta innefattar glesheten som uppstår givet variabler med ett stort antal nivåer/målvärden. Två viktiga sådana variabler utgörs av diagnos- och operationskoder, av vilka det finns tusentals målvärden [1,2]. Detta resulterar i ett stort antal variabler vid applikation av metoder såsom one-hot-encoding, vilket leder till att datan blir gles samt att var unik kod har liten representation i träningsdatan. Av denna anledning är det av intresse att undersöka metoder för gruppering av koder med likvärdig korrelation till resultatet.

Målsättning

Målsättningen med arbetet är att, givet en uppsättning av grupperingsmetoder, identifiera på vilket sätt diagnos- och operationskoderna bör grupperas för att effektivt användas vid prediktion av vårdtid på PostOP.

Arbetsbeskrivning

Målsättningen uppnås genom att implementera och utvärdera resultaten av olika grupperingsmetoder på en klassifikations- och regressionsmodell. För att erhålla en baseline skall koderna i rådatan grupperas enligt deras sina 2-4 första karaktärer och sedan undergå one-hot-encoding. Den kodlängd som ger bäst resultat skall utgöra baseline.

Baseline skall jämföras med de resultat som erhålles enligt metoden beskriven av [3]. Metoden innefattar en fusion av kategoriska variabler genom en lasso-regression vars straffterm baseras på skillnad mellan olika variablers relation till vårdtid på PostOP. Olika styrkor av straffermen skall undersökas.

Slutligen skall minst en till metod undersökas, *i mån av tid*. Denna extra metod avser att utgöra ytterligare grund för jämförelse. Metoden väljs av studenten i samråd med handledare och kan vara en av nedanstående eller någon annan metod såsom föreslagen av studenten.

- Weight-of-evidence encoder [4].
- Gruppering enligt kliniskt förväntad tid på PostOP.
- Icke-linjär diskriminantanalys baserad på kernelfunktioner såsom beskriven i [5].

Arbetet skall genomföras i Python. Detta arbete karaktäriseras av en överföring av matematisk teori till implementation. Studenten bör vara systematisk och ha god vana att läsa och söka vetenskapliga texter. Vidare

behöver studenten ha en grundläggande förståelse för maskininlärning samt tidigare erfarenhet av programmering med matematiska applikationer.

Referenser

[1] https://www.socialstyrelsen.se/statistik-och-data/klassifikationer-och-koder/icd-10/
[2] https://www.socialstyrelsen.se/statistik-och-data/klassifikationer-och-koder/kva/
[3] https://projecteuclid.org/journals/annals-of-applied-statistics/volume-4/issue-4/Sparse-modeling-of-categorial-explanatory-variables/10.1214/10-AOAS355.full
[4] https://datasciencestunt.com/weight-of-evidence-woe-and-information-value-iv/
[5] https://proceedings.neurips.cc/paper/1999/file/c0d0e461de8d0024aebcb0a7c68836df-Paper.pdf

Handledare vid IT (preliminärt)
Amanda Bertgren
CIMT, IT: Forskningsutveckling
Region Västerbotten
E-post: amanda.bertgren@regionvasterbotten.se