Uvod

Ovaj master rad se bavi problemom nedostajućih vrednosti u podacima koji predstavljaju ulaz za algoritme mašinskog učenja. Ukoliko podaci koji služe za trening nisu potpuni, može se desiti da istreniran model nije dovoljno pouzdan i da kao rezultat donosi pogrešne zaključke.

Skup podataka koji će se koristiti u radu je pogodan za regresionu analizu. Pokazaće se kako različiti procenti

# Експериментални подаци

## Опис скупа података

Иако ће сам експеримент бити заступљен у другом делу рада,

# Регресиона анализа

## Линеарна регресија

Регресиона анализа коцептуално представља једноставан метод проналажења функционалних зависности између променњљивих. [4] Та зависност је приказана у облику формуле у коме се са једне стране налази зависна променљива, а са друге стране скуп независних променљивих. Полази се од претпоставке да вредности независних променљивих утичу на вредности зависних променљивих.

Означимо ли зависну променљиву са *Y*, а остале променљиве *X1*, *X2*, … , *Xp* линеарну регресију можемо представити једначином.

|  |  |
| --- | --- |
|  | (1.1) |

Ознаком се означава грешка апроксимације. Уколико са означимо апроксимирану вредност (1.2), једначина (1.1) постаје:

|  |  |
| --- | --- |
|  | (1.2) |
|  | (1.3) |

Из једначине (1.3) је јасно да грешка представља разлику између очекиване и апроксимиране вредности, и пожељно је да та разлика буде што ближа нули.

Овај рад ће се фокусирати на линерану регресију, и због тога се (1.1) се представља:

|  |  |
| --- | --- |
|  | (1.4) |

Или векторски:

|  |  |
| --- | --- |
|  | (1.5) |
|  | (1.6) |

Проналажењем параметара вектора , таквих да је вредност минимална, одређује се зависност између и . Величина вектора и је где представља број независних промељивих. Разлог за додавање вредности у вектор је једноставан. Како параметри тог вектора одређују апрксимирајућу функцију, она би без вредности засигурно пролазила кроз координатни почетак. Да би вредност увек била присутна, додата је вредност у вектор , и као што је приказано у једначини (1.4), она увек има исту вредност.

## Примена

Сви екперименти у овом раду ће се извршавати на подацима о пацијентима који болују од дијабетеса. У табели (1) је визуелно представљен само подскуп података. Иницијални скуп података је садржао податке о 442 пацијента, што је за потребе овог рада подељено у два подскупа: први са 400 обсервација (тренинг подскуп) и други са 42 обсервације (тестни подскуп).

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Старост | Пол | ИТМ\* | КП\*\* | Резултати серумских мерења | | | | | | резултат |
| Пацијент |  |  |  |  |  |  |  |  |  |  |  |
| 1 | 59 | 2 | 32.1 | 101 | 157 | 93.2 | 38 | 4 | 4.9 | 87 | 151 |
| 2 | 48 | 1 | 21.6 | 87 | 183 | 103.2 | 70 | 3 | 3.9 | 69 | 75 |
| 3 | 72 | 2 | 30.5 | 93 | 156 | 93.6 | 41 | 4 | 4.7 | 85 | 141 |
| 4 | 24 | 1 | 25.3 | 84 | 198 | 131.4 | 40 | 5 | 4.9 | 89 | 206 |
|  |  |  |  |  |  |  |  |  |  |  |  |
| 399 | 52 | 1 | 27.8 | 85 | 219 | 136 | 49 | 4 | 5.1 | 75 | 242 |
| 400 | 65 | 2 | 28.5 | 109 | 201 | 123 | 46 | 4 | 5.1 | 96 | 232 |

Табела Тренинг подскуп. Поред демографских података (старост и пол), на пацијентима је вршено 8 мерења који заједно утичу на ниво дијабетеса годину дана касније

\* ИТМ, индекс телесне масе

\*\* КП, крвни притисак

Резултат представља ниво дијабетеса код људи годину дана након мерења. [5] Другим речима, линеарна регресија би требало да процени ниво дијабетеса годину дана након мерења. За ту потребу, пронађени су параметри коришћењем тренинг подскупа, и затим је цео модел проверен тестним подскупом. Резултати се налазе у табели 2.

|  |  |
| --- | --- |
|  | Вредност |
|  | -321.14 |
|  | 0.02 |
|  | -22.73 |
|  | 5.62 |
|  | 1.03 |
|  | -1.03 |
|  | 0.70 |
|  | 0.31 |
|  | 6.85 |
|  | 64.40 |
|  | 0.37 |
| Средња квадратна грешка (MSE) | 1668.71 |
| Корен средње квадратне грешке | 40.85 |

Табела Резултати линеарне регресије

Горњи део табеле 2 показује све вредности вектора . Очигледно је да је параметар додат са разлогом јер има вредност другачију од нуле, што значи да има утицаја на предвиђењу вредност. Доњи део табеле показује колико истренирани модел добро предвиђа.

|  |  |
| --- | --- |
|  | (1.5) |

Разлика између предвиђене вредности и тражене вредности се квадрира, а затим се израчуна просечна вредност такве разлике међу свим пацијентима. Добијена вредност се зове средња квадратна грешка (MSE). Додатно се рачуна и корен ове грешке јер он даје бољу слику колико модел греши. У овом случају, модел приликом процене нивоа дијабетеса за следећу годину греши за око 40 јединица. Важно је приметити да средња квадратна решка не даје никакве информације о знаку, тј. не зна се да ли модел предвиђа више вредности или ниже од очекиваних.

Овај рад се неће даље бавити линераном регресијом, али ће ови резултати служити као мерило (референтне вредности) у будућим експериментима.

# Недостајуће вредности

## Механизми недостајућих вредности

Технике машинског учења као сто су надгледано и ненадгледано учење се могу посматрати као системи у ком су улази представљени као подаци, а излази су истренирани модели (алгоритми). Из овог угла посматрања, квалитет података директно утиче на каснију тачност предвиђања истренираног алгоритма. Међутим, улазни подаци често нису комплетни и као такви често онемогућавају тренинг алгоритама.

Некомплетан скуп података се означава и као скуп података са недостајућим вредностима. У таквим скуповима вредности недостају по једном од три могућа механизма, а механизам представља математички однос између података и недостастка. [1]. Механизми могу бити:

**MCAR** – Missing Completely At Random – Nедостајуће вредности су присутне без икакве законитости. Уколико скуп података има 2 променљиве (колоне), непостојаност податка у првој колони нема никакву повезаност са вредностима из обе колоне. Овај случај је веома чест, обзиром да углавном настаје људском ненамерном грешком. На пример, испитаник је случајно превидео одређено питање и оно је остало неодговорено. [3]

**MAR** – Missing At Random – недостајуће вредности присутне унутар једне променљиве немају никакву законитост (повезаност) са том променљивом. Уколико посматрамо исти скуп података (2 колоне), непостојаност податка у првој колони не зависи од вредности те колоне, али зависи вредности из друге колоне(а). На пример, прва колона садржи податке о просечној оцени током студија, а друга колона резултате теста приликом запослења. Испитаници (редови у скупу података) са ниском просечном оценом неће бити ни узети у разматрање, па њихова оцена са теста је ирелевентана, и не садржи се у скупу података.

**MNAR** – Missing Not At Random – недостајуће вредности једне променљиве су директно зависне од посматране променљиве. У скупу са две колоне, недостајуће вредности прве колоне недостају због ње саме, и немају никакве повезаности са другом колоном. На примеру скупа података који садржи резултате теста као променљиву (назив колоне), подаци те колоне могу да недостају у свим редовима где је резултат теста мањи од одређене оцене.

За потребе експеримената у овом раду посматраће се искључиво MCAR механизам недостајућих вредности. Разлог за ту одлуку је могућност добијања недостајућих вредности синтетичким путем. Скуп података из 1.2 (Примена регресионе анализе) ће бити ''пробушен'' више пута насумично и том приликом ће проценат недостајућих вредности бити различит. На тај начин од првобитног комплетног скупа података добиће се више некомплетних, и као додатна погодност знаће се иницијални вредности (касније ће се те почетне вредности користити за евалуацију технике уметања података).

## Технике уметања података

У случају недостајућих података, понекад је најлакше одбацити обсервације које нису потпуне. На пример, уколико скуп садржи 100 обсервација (редова) и 10 атрибута (колона) , и од тога 10 различитих редова има тачно једну недостајућу вредност (тачно једну једну колону непопуњену), овом јендноставном техником остало би 90 редова (инстанци) као улаз за алгоритам машинског учења.

Описан скуп података се састоји од 100 редова и 10 колона. Матрица података потенцијално садржи 1000 вредности (потенцијално јер неке вредности нису присутне). Уколико 10 вредности недостаје, ова матрица ће садржати 990 ненедостајућих вредности, што представља 99%. Уколико применимо технику одбацивања обсервација, избацићемо 10 редова, односно укупно 100 вредности, и коначна матрица ће садржати само 90 редова (900 вредности, 90%).

Јасно се види да се оваквим приступом због 1% недостајућих података, може елиминистаи чак 10% укупних вредности. У сваком случају, избацивање података може касније довести до већих грешака предвиђања, јер се скуп података који служи за тренинг драстично смањује [6].

Стога, у овом делу ће бити описане само технике уметања података, где ће се поља која недостају у матрици података заменити (највероватнијом) вредношћу.

### Уметање средње вредности

Уметање средње вредности је вероватно најједноставнија метода. Она подразумева замену недостајуће вредносте за сваку променљиву (колону) са средњом вредности познатих обсервација у посматрној колони. Овај приступ може бити погодан у случају када мало података недостаје и то по MCAR механизму. У сваком случају смањује се варијанса међу подацима као и корелација између променљивих. [7] Логично је да се варијанса смањује јер до сада непознату вредност замењујемо ''очекиваном'' вредности, и самим тим смањујемо одступање од те ''очекиване'' вредности. Када се каже да ће корелација бити мања мисли се на корелацију између варијабли (колона). Како овим приступом покушавамо да пронађемо везу између свих обсервација (редова) унутарт једне колоне, логично је да смањујемо корелацију између колона. На пример, скуп података садржи две колоне (висина и тежина), и 5 обсервације. (Табела3)

|  |  |
| --- | --- |
| Висина | Тежина |
| 160 | 67 |
| 165 | 65 |
| 158 | 59 |
| ? | 98 |
| 191 | 98 |

Табела Уметање средње вредности

Очигледно је да постоји веза између прве и друге колоне, и да би недостајућа вредност требало бити замењена са 191 (или неком вредношћу блику кој). Међутим, овом техником се та веза (корелација) занемарује и недостајућа вредност ће постати 168.

Уметање средњих вредности се због описаних недостатака неће даље разматрати у раду. Екпериментални скуп података садржи вредности о особама, где неке променљиве (колоне) имају велику варијансу, па свакако није добра идеја сманњивати ту варијансу. Такоће, испитује се утицај разних фактора на ниво дијабетеса, па је корелација један од главних предуслова.

### Преношење задњег запажања

Преношење задњег запажања је још једна техника која захтева MCAR механизам недостајућих вредности. Веома је популарна у ситуацијама где се посматрају одређене појаве (субјекти) кроз време. Изузетно може бити знимљива у истраживањимакоје садрже номиналне типове атрибута (вредности колона). На пример, посматрајмо медицинско истраживање које прати пацијента (субјекат) кроз време и бележи да ли је узео терапију или није. Другим речима, постоји номинална променљива са могућим вредностима ДА/НЕ. Уколико је пацијент одређеног дана заборавио да унесе да ли је узео лек или није, та вредност ће се попунити са вредношћу из претходног дана.

Међутим, за променљиве нумеричког типа, ова техника није препоручљива јер повећава пристраснот модела, и такође измњује средњу вредност и варијансу (по променљивој) [8]. Ни ова теника неће бити коришћења у експериментима у овом раду, јер се скуп података за тренинг сасатоји углавном од нумеричких типова, и притом павијенти се не посматрају кроз време.

### Уметање података коришћењем регресије

Ова техника подразумева да постоји корелација између промен

Ста је уметанје података, мало историје о томе кад је поцело

Које су технике погодне ѕа мцар

О свакој техници по мало

Фокус рада на мцар, како це се мцар користити мало псеудо кода, датасет који имам

[1] Applied Missing Data Analysis, 3

[2] Little, R.J.A. and Rubin, D.B. (1987) Statistical Analysis with Missing Data. John Wiley & Sons, New York

[3] Prevention And Treatment Of Item Nonresponse, 155

[4] Regression Analysis By Example, 1

[5] Least Angle Regression

[6] Data Analysis Using Regression and Multilevel/Hierarchical Models, 531

[7] Principal component analysis with missing values, 659

[8] The Application of Last Observation Carried Forward in the Persistent Binary Case