МАТЕМАТИЧКА ГИМНАЗИЈА

МАТУРСКИ РАД

из рачунарства и информатике

Филтрирање података (математички модели, машинско учење, text, data и duo mining) и примери примене на класификацији података прикупљених софтвером за развијање способности визуелне перцепције особа са сметњама у развоју

Ученик
Милош Арсић, IVb

Ментор
Јелена Хаџи-Пурић

Београд, јун 2013

Садржај

Резиме

Циљ овог рада је да објасни, поред основних појмова, и примену машинског учења, data и text mining-а. Рад садржи апликацију за развијање способности визуелне перцепције којом су прикупљени подаци који су служили као задатак за класификовање SVM класификатору. При изради и анализирању апликација у раду водили смо се званичним методама пронађеним у литератури. На једноставан начин, кроз мноштво примера, у раду је стављен акценат на примене машинског учења, data и text mining-а са којима се најчешће сусрећемо. Највећи део рада чине примери из праксе где се види како се ова дисциплина и процеси успешно примењују при филтрирању и обради података не само у информатици већ и у економији, биологији, криминологији, медицини, трговини, маркетингу и у многим другим областима.

***Кључне речи*:** SVM класификатор, data mining, text mining, примена.

1. Увод

У протеклој години, према истраживању Републичког завода за статистику, установљено је да 55.2% домаћинстава поседује рачунар, што је за 3.1% више од 2011. године. Истим истраживањем установљено је да је број особа које поседују мобилни телефон све већи и да чак 76% испитаника свих старосних доби сматра да се развитком одговарајућих апликација и информационих система побољшава живот и рад савременог човека. Савремени начин живота и пословања карактеришу брзе промене у окружењу човека и начину његовог пословања, што захтева обраду великих количина података. Захтеви савременог доба постављају задатке за развијање бржих, квалитетнијих и адекватнијих програма за обраду података. Тако су се, у току двадесетог и двадесет првог века, развијали *data* и *text mining* (истраживање текстуалних података) као процеси прикупљања, обраде и анализе података. Разлика између data и text mining-a је да data mining тражи обрасце унутар структуираних података-база података, а text mining унутар неструктуираних података што су белешке и документа. Код data mining-a статистичар бира одговарајући алгоритам(е) за пословни проблем, припрема податке за анализу, а затим фино подешава модел заснован на резултатима. Text mining често при обради користи семантичке анализе и таксономију и тиме усклађује статистичке податке и вештачку интелигенцију. У раду ће такође бити објашњен појам duo mining-a који представља комбиновање data и text mining-a и најновија је техника која се користи у банкама.

Слика 1. Text и data mining

Као активиста и полазник обука на Harvard универзитету, био сам део тима који се залагао за израду сајта за оцењивање лекара. То ме је упознало, поред предходних, са још једном граном вештачке интелигенције-*машинским учењем*.

Машинско учење представља скуп теоријских резултата и примена из великог броја области као што су вероватноћа и статистика, теорија израчунљивости, информациона теорија, психологија и многе друге. Циљ је развити програме који су спремни да „уче“, да усвајају знање, и који су способни да се прилагоде на нове ситуације на основу претходног искуства, чиме се омогућава проверавање исправности неких модела и боље закључивање код великих количина података.

Слика 2. Машинско учење

Током свог волонтерског рада упознао сам се са начином рада и живота особа са сметњама у развоју као и са њиховим потребама.

Тим поводом у свом матурском раду направио сам апликацију која помаже у развијању способности визуелне перцепције особа са сметњама у развоју. Апликација је рађена у Scratch-у. Према резултатима истраживања, апликација је високо оцењена у домену јасног начина коришћења и резултата након њеног коришћења. Дакле, циљ овог матурског рада је да покаже, преко резултата успешности коришћења креиране апликације и машинског учења како програм може разликовати неке унете тачке и повући функцију (хипер-раван) која их најбоље одваја, што је увод за каснију обраду текста и података, као и за моделовање природних језика што ће бити показано у раду.

2. Софтвер у служби особе са сметњама у развоју

2.1 О перцепцији

Када кажемо перцепција, било да је у питању аудитивна, визуелна или нека друга, прво што помислимо је начин на који неко опажа свет. Да би перцепција била адекватна, неопходно је да све мождане структуре, а поготово мождана кора буду у потпуности зреле. Нажалост, код особа са интелектуалним тешкоћама, ове мождане структуре готово никада не достигну своју пуну функционалност.

Наиме, информација о ономе што смо видели или чули постаје део наше свести тек када прође мождану обраду и то кроз три поља.

Прво поље је прости центар-примарно поље вида, слуха и мириса. На нивоу овог поља присутно је просто опажање-има ли или нема светла, тишина је или се нешто чује. Друго поље су секундарна мождана поља где се одвија и прва обрада података. Примећујемо да је нешто црвено или плаво, звук који чујемо је у ствари људски глас. На нивоу терцијарних поља информација која је прошла прва два поља постаје значајна за нас. Перцепција се одвија на нивоу секундарних поља мождане коре.

2.2 Број особа са сметњама у развоју

Према подацима WHO (World Health Organization) број особа са интелектуалним тешкоћама у укупној популацији је око 2.5 %. Особе са менталним поремећајем често не могу остварити своја људска права и изложене су различитим облицима дискриминације, која је најприсутнија у области образовања, запошљавања, културног живота и приступа јавним местима и службама.

У Србији не постоји свеобухватна званична евиденција и регистар особа са сметњама у развоју. Према подацима Савеза друштава за помоћ ментално недовољно развијеним особама које чине 58 општинских и међуопштинских друштава, евидентирано је 89943 особа са сметњама у менталном развоју и са вишеструким сметњама у развоју.

2.3 Развој визуелне перцепције

Значајан део мождане коре обрађује визуелне податке. Сматра се да 90% свих информација о свету око нас усвајамо преко чула вида. И неке друге вештине, као што је писање, захтевају визуелну контролу иако су то у суштини моторни процеси.

Чуло вида нам омогућава да примимо информацију на даљину односно без непосредног додира, мириса или испитивања укуса.

На рођењу, новорођенчад нису у могућности да процене дубину и имају веома ограничену могућност распознавања боја. До шестог месеца деца почињу да се интересују за боје, на кратко фокусирају мање предмете, могу да прате кретање објеката и да их визуелно истражују, могу да лако пребацују поглед са ближих предмета ка даљим и обрнуто. Дакле, за првих шест месеци живота, код здравог детета се развија визуо-моторна контрола. Тада деца воле да гледају како предмети падају и да гледају куда је предмет нестао. Између шестог и деветог месеца дете може бити заинтересовано за све геометријске дезене. Од деветог месеца до годину дана дете примећује објекте величине пар милиметара у свом окружењу. Јако се интересује за мимику, распознаје лица, покушава да имитира одрасле и разликује непознате од познатих људи. До друге године дете може да распознаје одређене облике и да их слаже по неком критеријуму, може да именује животиње и предмете у дечијим књигама. До треће године дете може да решава једноставне слагалице, да нацрта круг и ставља мање предмете у веће. Од пете до седме године дете ће развити готово све особине и функције које поседује одрастао човек, али неће бити у могућности да их стави у социјални контекст (недовољна развијеност која ће почети у пубертету).

Код деце са сметњама у развоју, поготово са комплексним сметњама, ово сазревање се често не одигра до краја. Поједине функције могу бити више или мање развијене. Зато је и опажање предмета и људи око њих отежано. Уколико, уз то, дете има и тешкоће са моториком (церебрална парализа) изостаће и визуо-моторна контрола, што додатно отежава развој визуелне перцепције. Ипак, мозак детета поседује особину пластичности односно могућност да други делови мождане коре преузму функције делова који су оштећени. Ово нам даје шансу да, користећи адекватне софтвере, помогнемо у вежбању одређене функције. Кроз понављање визуелних стимулуса, излагањем и упамћивањем овај процес се доста убрзава. Уколико помогнемо особи која има проблем са визуелном перцепцијом да усаврши своје опажање, даћемо јој шансу да боље разуме свет око себе и да се више и квалитетније укључи у друштво.

2.4 Како направити софтвер за развијање способности визуелне перцепције особе са сметњама у развоју

Важно је пратити физиолошки пут развијања визуелне перцепције и правовремено задавати стимулусе. Никада не треба ићи испред онога што је очекивано за дати узраст особе. Са друге стране, када развој перцепције касни, потребно је пронаћи ниво који је дете усвојило па наставити даље.

Код креирања софтвера, битно је да је он једноставан за коришћење. Потребно је да буде компатибилан и лак за састављање и рад са помоћним средствима (специјалним мишевима и тастатурама) односно да ради на клик или неку другу команду.

Слика 3. Мишеви и тастатуре за особе са сметњама у развоју

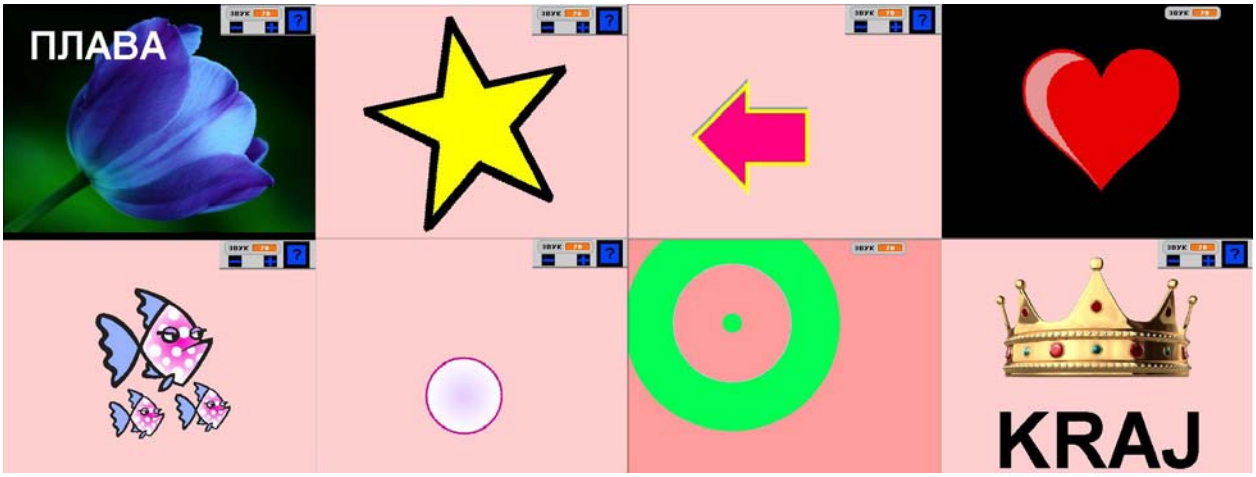
Пошто желимо да потпомогнемо памћење, увек је корисно користити глас и звукове који прате одређени стимулус. Потребно је да вежба траје кратко. Било би од велике користи да боје буду контрасне, да се могу користити монитори велике резолуције, а да су стимулуси који се детету представљају очигледни, интересантни и њему блиски. На почетку би требало да број детаља буде минималан. Касније се стимулуси могу усложњавати и њихов број може бити већи.

2.5 Опис садржаја апликације

Основна намена апликације је развијање способности визуелне перцепције. Апликација је рађена у Scratch-у и садржи неколико различитих визуелних и аудио стимулуса:

контрасне боје, предмети који нестају и појављују се на клик тастера, адекватан звук (глас и музика), просторне одреднице.

Апликација се једноставно користи-једини догађај је клик на тастер чиме се покреће следећа анимација која траје у просеку око петнаест секунди.



Слика 4. Неке од сцена креиране апликације

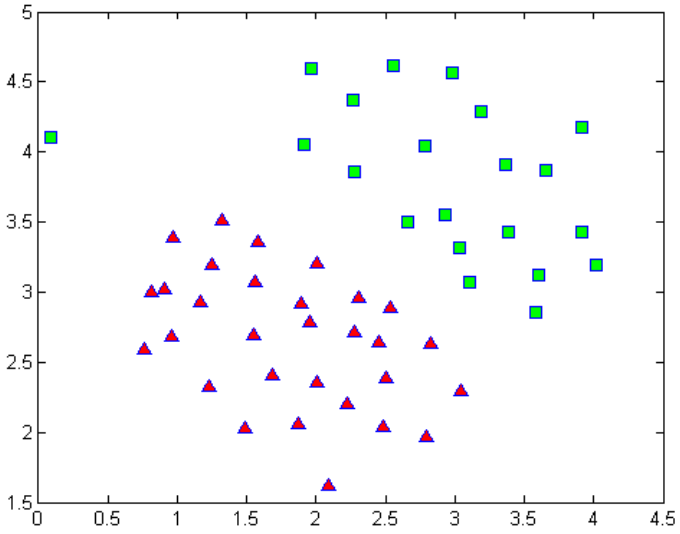
2.6 Истраживање и резултати истраживања

Након пробне групе која је тестирала апликацију извршили смо истраживање. У истраживању су учествовале две групе особа А и В. Група А је била група ометених особа које или имају мање од 13 година или су са тежим сметњама у развоју, а група В је била група ометених особа које или имају више од 13 година или су са лакшим сметњама у развоју. У истраживању су учествовала деца школског узраста и млађе особе са сметњама у развоју (до тридесет година) из неколико дневних боравака са територије Београда.

Обе групе су показале изузетну заинтересованост за коришћење апликације. Истраживање је трајало седам дана и током њега утврђено је да је 95% особа са сметњама у развоју успело да користи апликацију, да су особама групе А најзанимљивије анимације са контрасним бојама и необичним звуком, а особама групе В први ниво у коме се уче боје и ниво у коме је приказано срце.

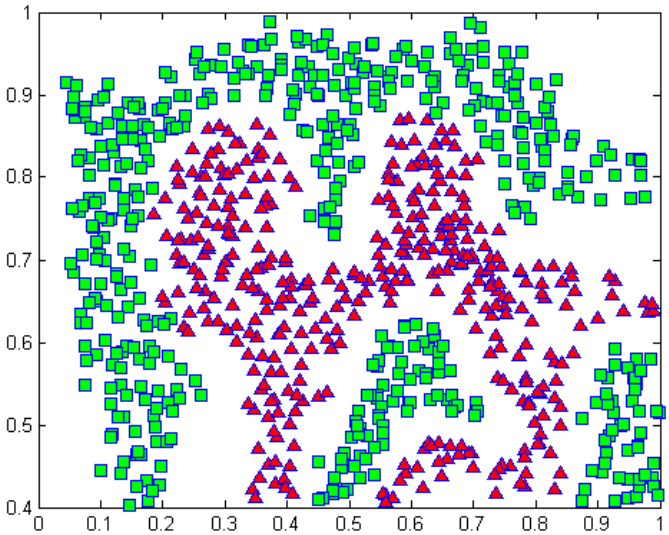
Истраживање је изведено, заједно са дефектолозима, на три различита нивоа, а резултат тог истраживања су следећа три графика (особе групе А означене су троуглом, а групе В квадратом):

1. график зависности просечне оцене реакције, у односу на просечну оцену психичког стања особе пре сваког коришћења апликације (оцене су дали дефектолози, а просечна вредност је резултат пет неузастопних коришћења апликације);



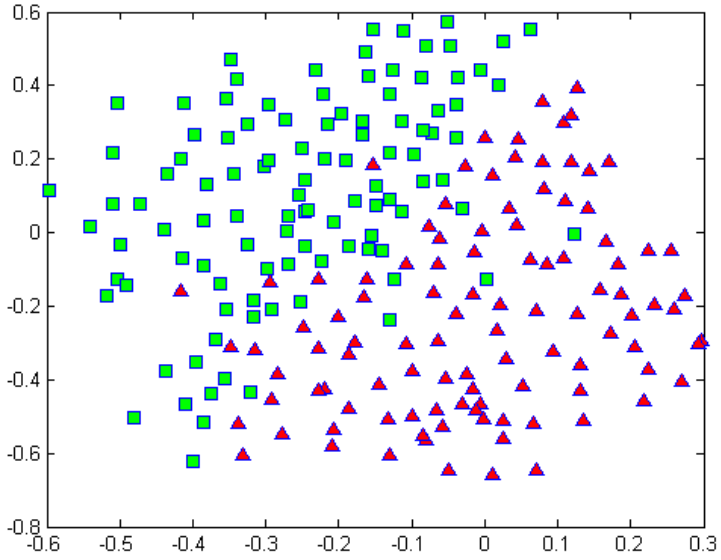
Слика 5. График 1

2. график зависности успешности коришћења апликације (резултат пет неузастопних коришћења апликације изражен је у процентима) у односу на проценат оштећења мозга;



Слика 6. График 2

3. график зависности просечне јачине реакције особе (може бити позитивна или негативна и резултат је пет неузастопних коришћења апликације изражен у процентима) у односу на просечну очекивану јачину реакције (резултат је пет различитих процена дефектолога пре сваког од пет неузастопних коришћења апликације и изражен је у процентима).



Слика 7. График 3

Покушајмо сада на графику резултата поставити границу између особа А и В. Уколико је рачунар тај који треба да одлучи о томе где се налази ова граница (хипер-раван) ситуација се знатно усложњава. Очигледно је да је на првом графику (Слика 5) могуће конструисати линеарну функцију која ће раздвојити резултате особа групе А од резултата особа групе В, али шта радити у ситуацијама на слици 6 и 7 и како направити програм који за унете тачке може генерисати „границу одлучивања“-хипер-раван која може на најбољи начин да раздвоји унете тачке? О томе зашто нам је ово неопходно и значајно и којим методом се може направити програм који ово може извести биће објашњено у следећем поглављу овог рада.

3. Машинско учење

3.1 Увод у машинско учење

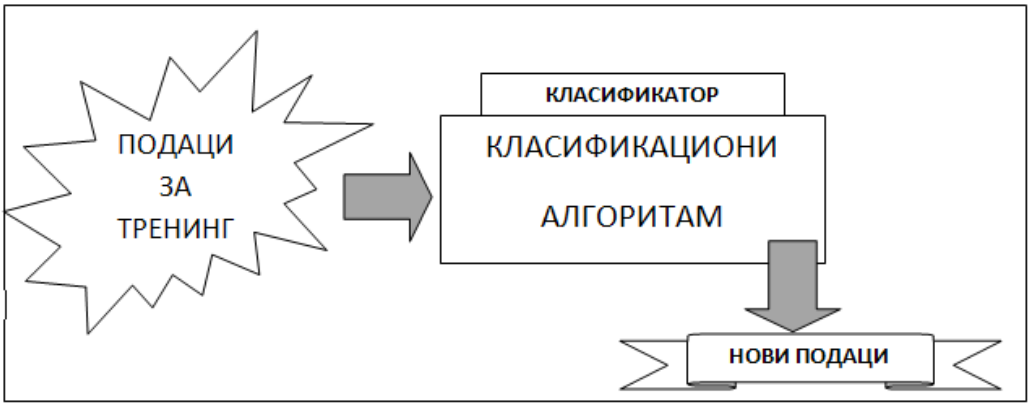
Машинско учење је једна од веома значајних грана вештачке интелигенције. Представља дисциплину која се бави пројектовањем прилагодљивих рачунарских система који су способни да побољшају своје перформансе стицањем искуства-учењем. У последњих двадесет година машинско учење представља метод који даје велике могућности рада са рачунарима. Скоро свакодневно се срећемо са примерима употребе ове области. На многим сајтовима сакупљају се информације о томе које су све књиге продате и ком члану и који су све филмови погледани. Ови подаци касније дају могућност, након обраде одговарајућим програмима, да се на клијентову email адресу пошаљу адекватне понуде и промоције у складу са оним што он најчешће гледа и купује на поменутом сајту. Посматрајмо програм који може да преводи рукопис или говор у текст. Такав програм морао би да са слике и из изговорене реченице „препознаје“ знакове и гласове које би после „преводио“ у текст, али како људи не пишу и не говоре на исти начин било би јако тешко направити такав програм. Посматрајмо сада велику количину лекарских извештаја. Њиховом обрадом били бисмо у могућности да добијемо закључке о трендовима болести у медицини и најуспешнијем лечењу одређених болести.

На ове и на бројне друге проблеме и питања одговоре даје машинско учење, као дисциплина која се бави проучавањем генерализације и конструкцијом алгоритама којим се може генерализовати. Машинско учење развило је велики број алгоритама за успешну обраду података. У овом раду, при обради података прикупљених апликацијом за развијање способности визуелне перцепције особа са сметњама у развоју, користили смо SVM алгоритам који ће детаљније бити објашњен у следећем делу овог поглавља.

3.2 Класификација и SVM алгоритам

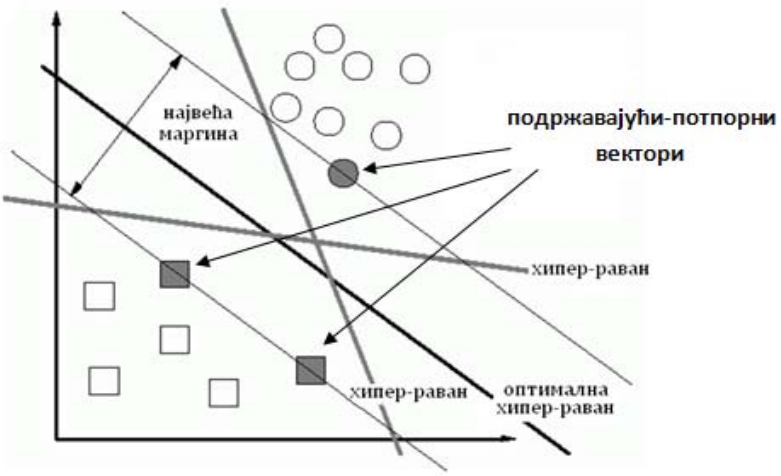
Класификација је процес којим се подаци пресликавају у предефинисане скупове података (класе). Према типу података и начину класификације постоји велики број класификација система као што су: класификација система према интеракцији са околином, класификација система према промењивости стања система (статички и динамички) као и многе друге. Свака класификација састоји се из две фазе.

У првој је потребно добро дефинисати модел (класификациони алгоритам) на основу унетих-тренирајућих података, а у другој је потребно применити дефинисани модел на нове податке.



Слика 8. Процес класификације

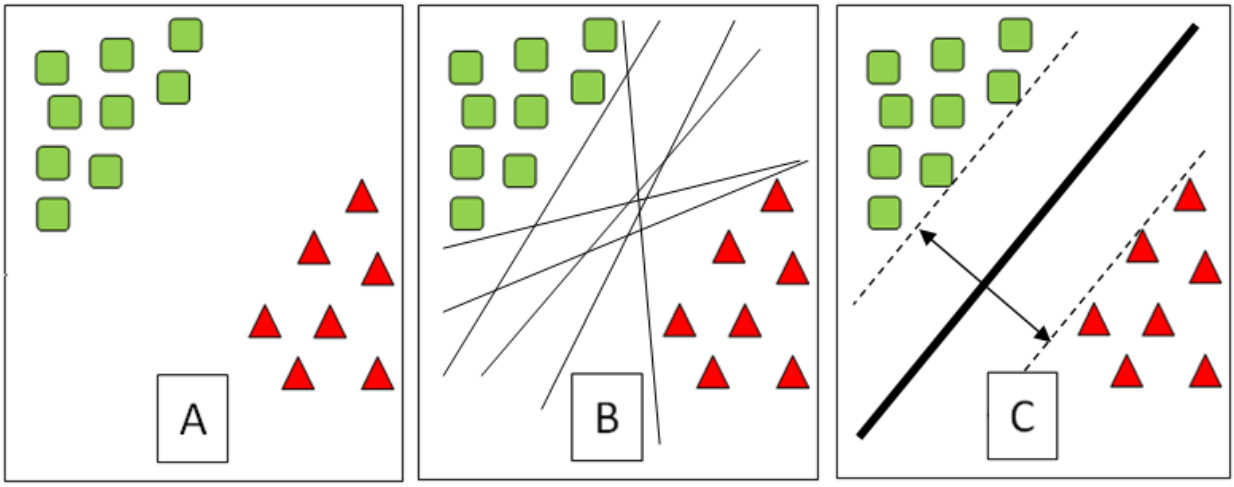
Један од великог броја метода за класификацију је алгоритам SVM. SVM (Support Vector Machine) је метод подржавајућих-потпорних вектора и један је од најпопуларних алгоритама за класификацију података. Користи се када је број димензија података велики. Основни циљ употребе овог алгоритма је да се раздвоје подаци различитих класа одговарајућом хипер-равни тако да су сви (или највећих број) података из исте класе са исте стране хипер-равни. Постоји велики број једноставних алгоритама којима се може одредити хипер-раван која раздваја класе међутим једино се алгоритам SVM користи уколико је потребно пронаћи оптималну раван. Хипер-равни потпуно су одређене специфичним подскупом података за тренинг-подржавајућим (потпорним) векторима.



Слика 9. Приказ одређивања оптималне хипер-равни

3.2.1 Линеарни SVM класификатор

Нека су дати линеарно раздвојиви подаци за тренинг система распоређени као на слици 10 у случају A. Јасно је да су на слици означене две класе података (квадрати и троуглови). Покушајмо сада одредити хипер-раван која најбоље раздваја ове две класе. Интуитивно је јасно да постоји велики број хипер-равни које могу раздвојити ове податке. Неке од њих приказане су на слици 10 у случају B. Међутим, како одредити која је хипер-раван оптимална и која ће хипер-раван бити резултат класификације SVM алгоритмом? Потребно је дефинисати маргину као растојање хипер-равни од података за тренинг. Маргина је означена на слици 10 у случају C. Тако се за сваку потенцијално оптималну хипер-раван може одредити њена маргина. Резултат класификације SVM алгоритмом биће она хипер-раван чија је маргина највећа и та хипер-раван зове се оптимална хипер-раван.



Слика 10. Одређивање оптималне хипер-равни

Употреба линеарних класификатора је веома битна. Уколико се подаци сместе у координатни систем тада се хипер-равни може одредити једначина њеног простирања ($ax + by - c = 0$). Посматрајмо слику 10 у случају C.

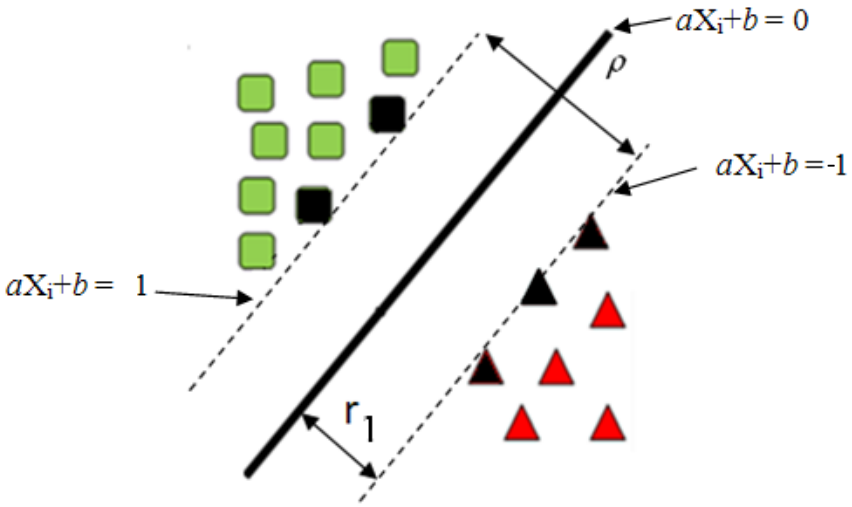
Нека је једначина хипер-равни $ax + by - c = 0$. Хипер-раван се још назива и граница одлучивања. Тада би важило да се у простору (координатном систему) за $ax + by \geq c$ налазе квадрати, а за $ax + by \leq c$ само троуглови. Ово ће бити случај са резултатима нашег истраживања са слике 5. О анализи ових резултата биће речи касније.

3.2.2 Линеарна алгебра и SVM класификатор

Како гласи формула по којој се може израчунати једначина хипер-равни? Прво, морамо увести претпоставке и ознаке како би могли дефинисати исправну једначину. Димензију простора који посматрамо означићемо са d, а сваки податак i биће приказан у облику вектора $X_i = (X^1, X^2, ..., X^d)$ где сваком податку X_i одговара тачно једна вредност за врсту класе $Y_i \in \{0, 1\}$. Из линеарне алгебре следи да је хипер-раван R приказана у облику:

$aX+b = 0$ где су a и b вектори записани у канонским облицима.

Познатом формулом може се одредити да је растојање било код податка од хипер-равни R чија је једначина облика $aX+b = 0$ једнако: $r_0(X, R) = \frac{aX + b}{\|a\|}$. Уколико ширину маргине означимо са ρ следи да је $\rho = 2 r_0$. Вектори a и b се канонски одређују тако да је растојање поменутих потпорних вектора од хипер-равни једнако 1. Одатле следи да су једначине хипер-равни које би садржале потпорне векторе $aX_i+b = 1$ односно $aX_i+b = -1$ у зависности да ли се потпорни вектор налази изнад или испод хипер-равни (слика 11).



Слика 11. Једначине потпорних вектора

Дакле, растојање потпорних вектора од хипер-равни је: $r_1(X, R) = \frac{1}{\|a\|}$, одакле следи да је маргина $\rho = \frac{2}{\|a\|}$. Уочимо да за сваки елемент класе квадрат важи да се налази изнад функције $aX_i+b = 1$. Ако се сваком елементу класе квадрат X_i додели вредност за тип класе $Y_i = 1$, а класе троугао $Y_i = 0$, следи да се закључак може записати као:

За сваки податак (X_i, Y_i) важи: $aX_i+b \geq 1$ уколико је $Y_i = 1$, а

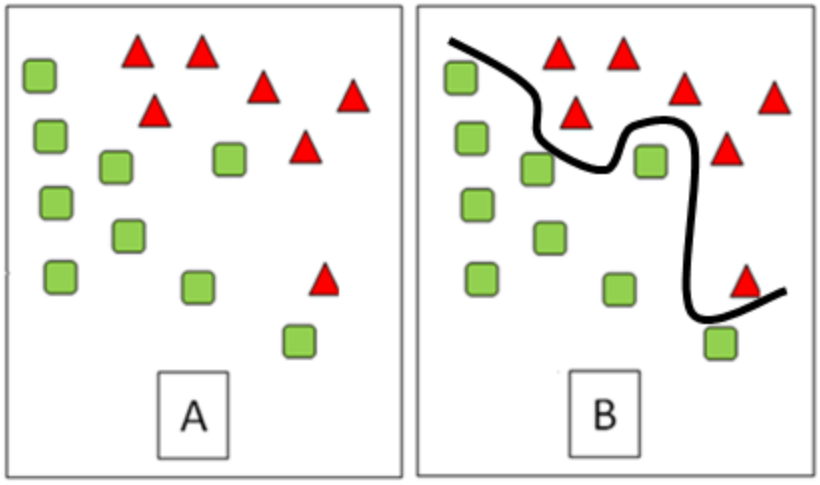
$aX_i+b \leq -1$ уколико је $Y_i = 0$.

Проблем дефинисања највеће маргине своди се дакле на проблем одређивања вектора *a* и *b* тако да $\rho = \frac{2}{\|a\|}$ буде максимално, уз услов : $aX_i+b \geq 1$ уколико је $Y_i = 1$ и $aX_i+b \leq -1$ уколико је $Y_i = 0$. Постоји велики број решења овог проблема, али технике његовог решавања превазилазе оквире овог рада. Детаљније о овој техници и њеним применама читалац може да потражи у [3].

Навешћемо само формулу која се добија користећи се једном од техника (Лагранжеовим мултипликатором). Након коришћења ове методе проблем се своди на одређивање вектора *a* и *b* тако да $L(a) = \frac{\|\vec{a}\|^2}{2} + C\left(\sum_{i=1}^{N}\xi_i\right)$ буде минимално, где су ξ_i променљиве које толеришу грешку (у току учења), а C параметар који контролише тачност процеса (што је C већи број то је тачност већа).

3.2.3 Нелинеарни SVM класификатор

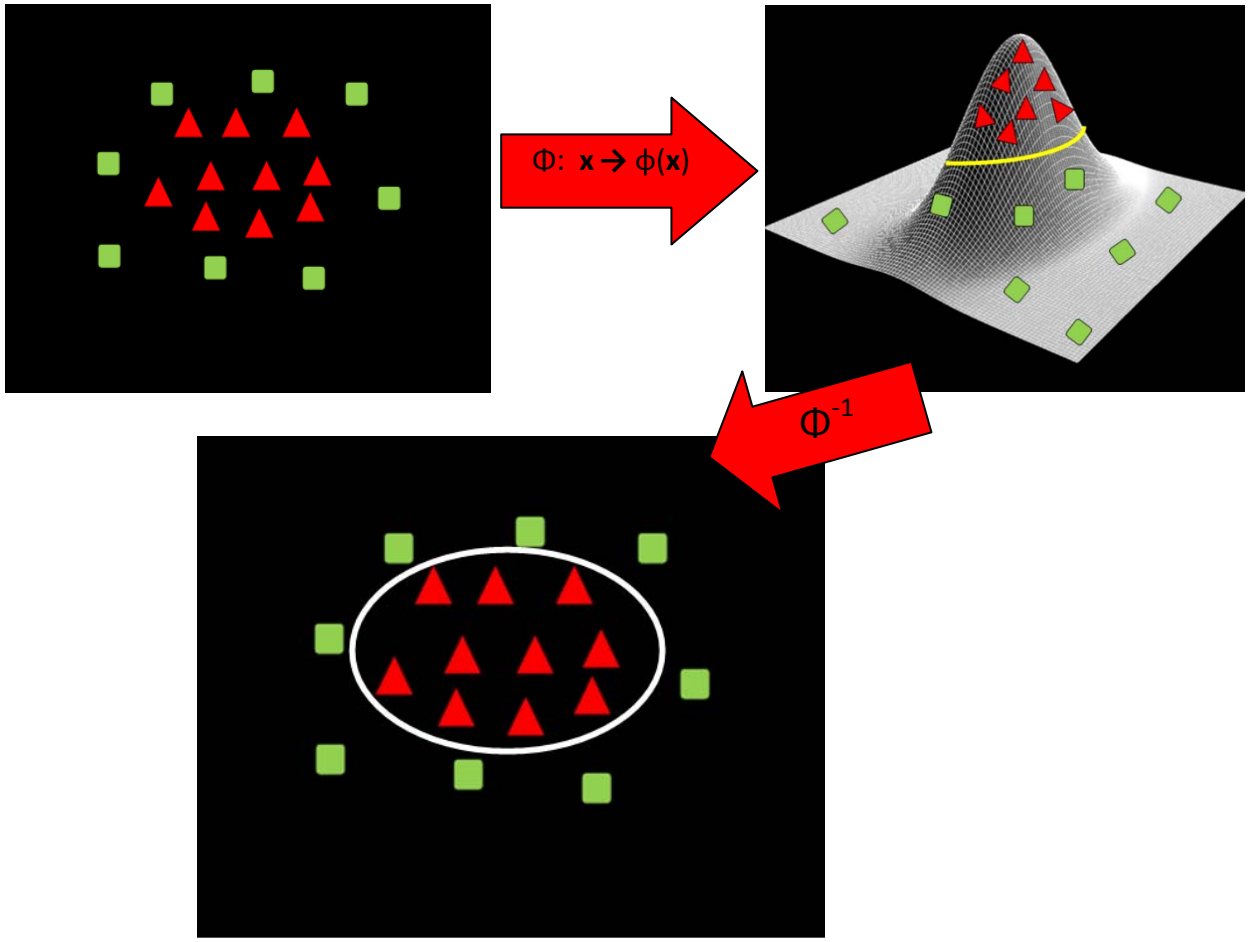
Знатно чешћи случај приликом раздвајања класа података јесте случај када су подаци за тренинг система линеарно нераздвојиви. У том случају, због сложености распореда података, врло је теже раздвојити податке за тренинг. Погледајмо слику 12 у случају А.



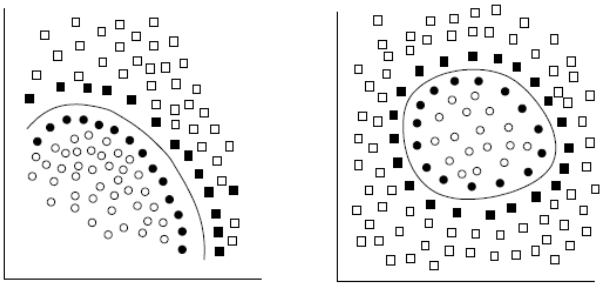
Слика 12. Пример линеарно нераздвојивих података за тренинг

Јасно је да је немогуће раздвојити ове податке линеарном хипер-равни те се стога прибегава другим методама.

Циљ је пресликати улазни векторски простор у неки вишедимензиони простор у коме је могуће раздвојити податке за тренинг. Метод који смо користили у раду и апликацији зове се Гаусов кернел. Погледајмо следећи пример на слици 13:



Слика 13. Пример употребе Гаусовог кернела за одређивање хипер-равни



Кернел је функција која одговара скаларном производу у неком проширеном простору. Користи се да би се подаци из линеарно нераздвојивих векторских простора пресликали у неке друге просторе где се могу раздвојити са хипер-равни.

Слика 14. а) полиномијални кернел б) RBF кернел

Постоји математичка теорија према којој се одређује да ли је нека функција кернел. Та теорија зове се Мерцерова теорија. Мерцерова теорија дефинише услове које дата функција треба да задовољава да би у неком векторском простору представљала скаларни производ (кернел). Неке од тих особина су: симетричност, позитивна дефинисаност, ...

3.2.4 Примена SVM класификатора

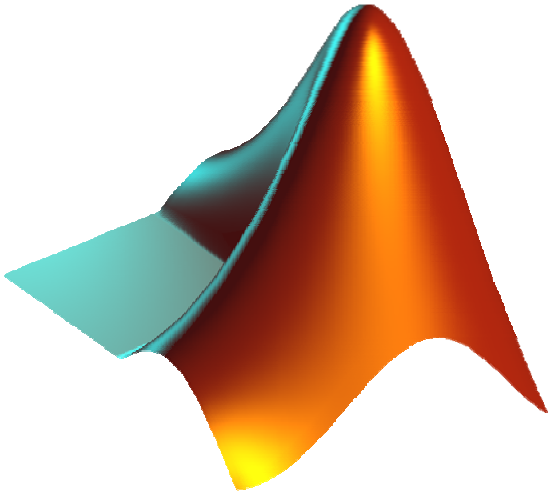
SVM је једна од највише коришћених и најуспешнијих техника са веома разноврсном применом. Неке од примена SVM-а су:

- у биоинформатици (Сваки ген/протеин приказује се као вектор дужине n, где је n или експериментално одређена вредност или вредност која одређује активност гена. Овде се SVM користи за предвиђање понашања нових гена, откривање хомолога, врши различите класификације гена/протеина према њиховом пореклу,...);
- код програма за препознавање гласа и рукописа;
- класификација докумената (У великим фирмама користећи се data minig-ом и SVM класификацијом, одвајају се различити типови електронских докумената (рачуни, уговори,...));
- у медицини (Обрадом великог броја извештаја лекара, добијају се информације потребне за проучавање трендова болести и успешног лечења);
- за препознавање објеката и лица људи (Највише се користи код програмирања робота. Специфичне црте лица неке особе преводе се у вектор, а потом се класификују SVM алгоритмом);
- у економији (За одређивање најбоље цене неких произода на тржишту);
- код електронске поште (Све примљене поруке разврставају се или на не-спам или на спам поруке);
- у банкама (Процена кредитне способности особе);
- у трговини и на интернету (За профилисање купца/корисника услуга).

У овом раду приказаћемо читаоцу употребу SVM-а у апликацији код класификације података прикупљених и приказаних у поглављу 2.6 овог рада. Апликација је рађена у MATLAB-у и OCTAVE-у о чему ће бити речи у следећим поглављима овог рада.

3.3 MATLAB и OCTAVE класификатор података

3.3.1 О MATLAB-у



MATLAB је математичко и симулационо окружење за нумеричке прорачуне, симулирање и анализирање процеса, развијање алгоритама, обраду података, приказ (визуелизацију), а све то кроз интерактиван и програмски рад.

Развија и продаје га фирма MathWorks, а скраћеница MATLAB је потекла од **MAT**rix **LAB**oratory-Лабораторија за матрице. MATLAB омогућава лако манипулисање подацима и матрицама, приказивање функција и лако повезивање са програмима писаним у другим програмским језицима. MATLAB је изумео мексички професор и шеф катедре за информатику Cleve Moler (Клив Молер).

Интерактивни рад у MATLAB-у постиже се задавањем наредби из командног прозора (Command Window). На тај начин се, након што унесете ваше наредбе и притиснете тастер ENTER, команде одмах извршавају. Поред овог прозора у оквиру MATLAB-а постоји и прозор где се могу видети претходне наредбе (Command History), прозор за писање и измену сопствених програма и функција (Program Editor), прозор који приказује тренутни фолдер (Current Folder), прозор који приказује радну површину (Workspace) и многи други. Такође, у оквиру MATLAB-а постоји велики број функција које се на основу намене групишу у класе (Toolbox-ове).

MATLAB нуди могућност коришћења помоћи (help-а) који, поред основних дефиниција наредби, садржи и примере њиховог коришћења чиме је знатно олакшано учење овог програмског језика.

MATLAB поред рада са подацима нуди могућност рада са векторима и приказ података на графицима због чега сам се одлучио да SVM класификатор направим баш у њему. Оригинална верзија MATLAB-а се плаћа, док се на сајту произвођача може преузети пробна верзија која траје само тридесет дана.

3.3.2 О OCTAVE-у

GNU OCTAVE је језик високог нивоа намењен нумеричким израчунавањима. Дизајнирали су га James B. Rawlings и John G. Ekerdt. Првобитно замишљен 1988. године као додатак уз уџбеник из хемије (програм за приказивање хемијских реакција). Након што су видели мане оваквог приступа проблему приказивања реакција, ова два професора сложила су се да покушају да направе много флексибилнији алат. Назив OCTAVE нема везе са музиком, већ је то име професора једног од креатора OCTAVE-а који је био познат по својим брзим и лаким прорачунима и амбициозном раду.

Многи људи саветовали су професоре да уместо да користе OCTAVE једноставно искористе Fortran који је био језик рачунарског инжењерства тог времена. Ипак, професори су схватили да ученици изгубе доста времена на откривању грешака у Fortran коду па су желели да развију ново, интерактивно и боље окружење.

Врло брзо, већ 1992. године, OCTAVE је доживео пуну фазу развоја и од 1993. године користи се како као помагало за решавање многих реалних (стварних) проблема, тако и у комерцијалне сврхе. Овај језик данас користи велики број људи.

OCTAVE пружа угодан интерфејс командне линије за решавање линеарних и нелинеарних проблема. За обављање других нумеричких експеримената OCTAVE користи језик који је веома компатибилан са MATLAB-ом чиме је омогућено лако превођење из једног у други програмски језик. Постоје и додаци који поред командног прозора омогућавају и графичко приказивање резултата и цртање функција.

За разлику од MATLAB-а OCTAVE се може бесплатно, под условима GNU лиценце, преузети са интернета са званичног сајта http://www.gnu.org/.

3.3.3 Опис садржаја апликације

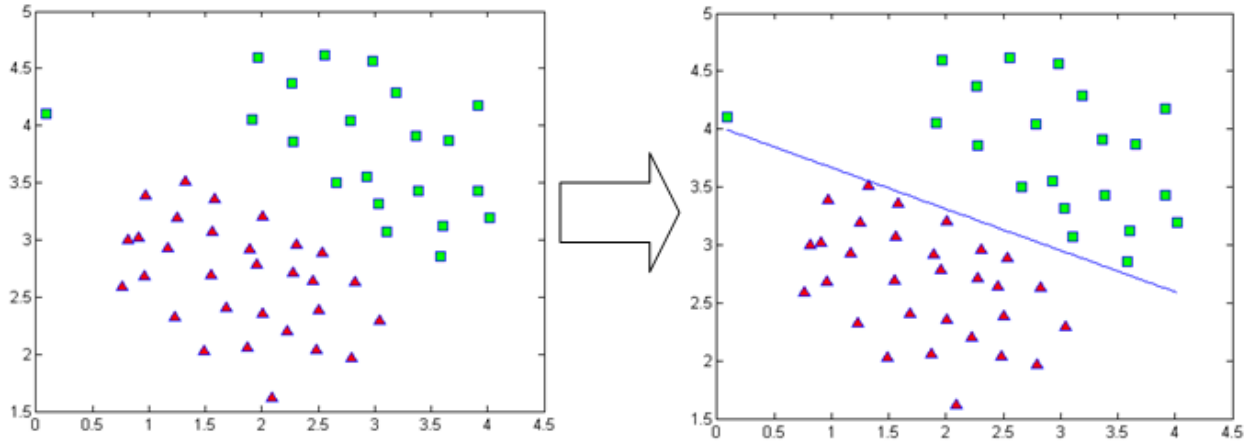
Апликацију је могуће покренути и у OCTAVE-у и у MATLAB-у. Користи се једноставно. Покретањем апликације и пратећи команде притисцима на тастер ENTER читалац може видети како рачунар раздваја, уз помоћ хипер-равни, две класе података прикупљених истраживањем из поглавља 2.6 овог рада. Следе исечци MATLAB кода апликације и објашњења њиховог функционисања. *Напомена*: Уз апликацију направљена су 3 .MAT фајла са подацима за унос тачака за сваки од графика. У њима су смештене координате тачака које програм треба да визуелизује на графику, а касније и раздвоји (класификује). При изради апликације коришћена су два пакета svmTrain.m и svmPredict.m из јавно доступне библиотеке функција за нумеричко израчунавање.

3.3.3.1 Цртање графика и одређивање линеарне хипер-равни

```
%% Masinsko ucenje
%  SVM klasifikator
clear ; close all; clc
%% ========Ucitavanje i prikaz podataka iz prve tabele========
fprintf('Ucitavanje i prikaz podataka ...\n')
load('podaciizstrazivanja1.mat');
crtajPodatke(X, y);
fprintf('Program zaustavljen. Pritisnite enter da nastavite. \n');
pause;
%% ====================Treniranje linearnog SVM-a====================
fprintf('Pocinje treniranje linearnog SVM na skupu podataka 1 i nakon
toga, na grafiku, bice iscrtana naucena granica odluke- hiper-ravan
\n');
load('podaciizstrazivanja1.mat');
fprintf('Treniranje linearnog SVM-a ...\n')
%unesite razlicite vrednosti broja C, sto vece C to veca tacnost
C = 1000;
model = svmTrain(X, y, C, @linearniKernel, 1e-3, 20);
prikazilinearnuhiperravan(X, y, model);
fprintf('Program zaustavljen. Pritisnite enter da nastavite. \n');
pause;
```

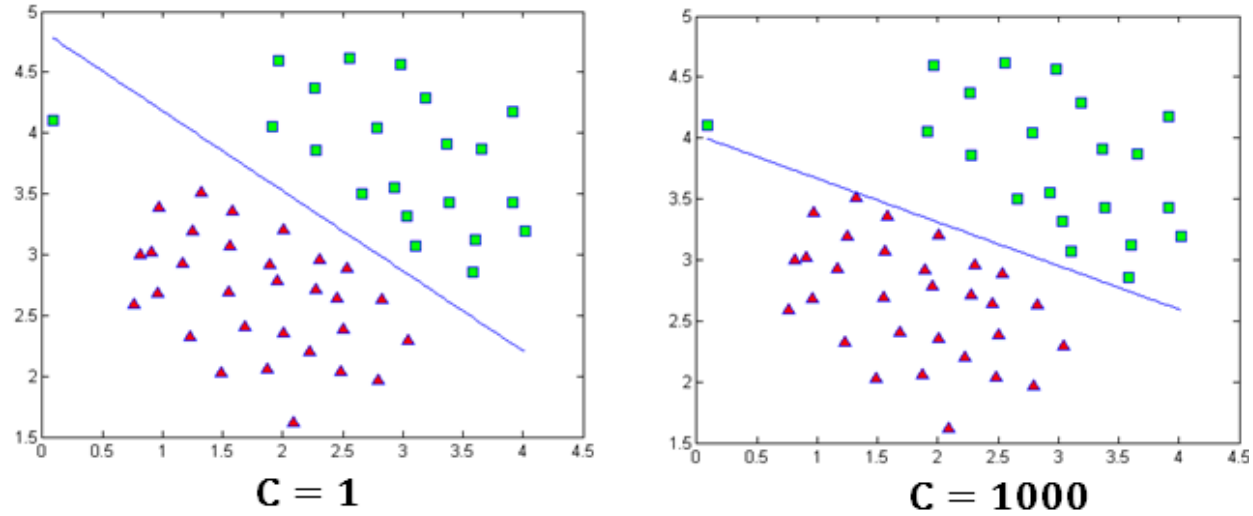
Код који је наведен изнад, учитава податке из првог дела истраживања, приказује их на графику и црта оптималну хипер-раван. Функција fprintf('text\n') исписује text на екрану, load('podaciizstrazivanja1.mat') учитава податке из podaciizstrazivanja1.mat фајла који се креира користећи wizard, crtajPodatke(X,y) црта график где је X апсциса, а y ордината, pause зауставља апликацију до притиска на тастер ENTER, svmTrain креира модел од датих података за тренинг и константе C (поглавље 3.2.2), а на крају prikazilinearnuhiperravan(X, y, model) приказује хипер-раван уз помоћ креираног модела.

Погледајмо како изгледа нацртана хипер-раван.



Слика 15. Одређивање хипер-равни код резултата првог дела истраживања

Уколико покушамо да променимо вредност параметра C промениће се и тачност (нагиб) хипер-равни. Погледајмо следећу слику која приказује хипер-раван у два случаја: за C = 1 и за C = 1000.

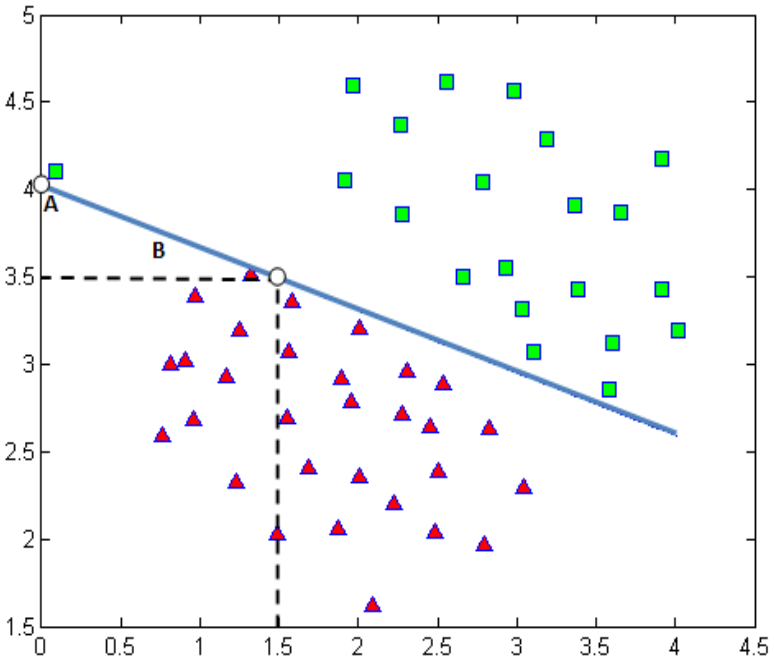


Слика 16. Положај хипер-равни у зависности од вредности параметра C

Приметимо да за вредност C = 1 нису сви елементи класе квадрат смештени са исте стране хипер-равни док за вредност C = 1000 јесу. Као што је напоменуто у поглављу 3.2.2 што је параметар C већи, то је тачност цртања хипер-равни већа. Након притиска тастера ENTER покреће се следећи део апликације приказан у поглављу 3.3.3.3 овог рада.

3.3.3.2 Примена линеарне хипер-равни

Посматрајмо нацртану хипер-раван и покушајмо одредити њену једначину.



Слика 17. Одређивање једначине линеарне хипер-равни

Претпоставимо да је једначина праве која раздваја класе са графика облика Ax + By - C = 0. Како та хипер-раван пролази кроз тачке A(0,4) и B(1.5,3.5) на познат начин може се одредити да је једначина праве са графика: x + 3y - 12 = 0. Из поглавља 3.2.1 и 3.2.2 овог рада следи да се добијена једначина праве може користити као неједнакост за предвиђање и класификацију нових података који би се унели у обраду касније.

Јасно је да важи да се свака особа групе A (троугао) налази испод хипер-равни на графику односно да за њу важи неједнакост $x + 3y - 12 \leq 0$, а да се свака особа групе B (квадрат) налази изнад хипер-равни на графику односно да за њу важи неједнакост $x + 3y - 12 \geq 0$, где је x оцена реакције особе на апликацију, а y оцена психичког стања особе пре коришћења апликације.

Следе неки од примера предвиђања користећи добијене неједнакости и једначину линеарне хипер-равни:

- За све особе које имају теже облике сметњи у развоју (Група A) следи да важи да њихова највиша оцена реакције на апликацију може бити оцена 4 што је највећа вредност функције која приказује хипер-раван на графику на интервалу [0,5].
- Уколико знамо да је оцена психичког стања особе била x, а оцена реакције y користећи се одређеним неједнакостима $x + 3y - 12 \leq 0$ и $x + 3y - 12 \geq 0$ може се закључити којој групи особа A или B неко припада у зависности да ли је $x + 3y - 12 \leq 0$ или $x + 3y - 12 \geq 0$. Ово је јако битно предвиђање и користи се највише у фармацији и медицини.

Дакле, одређују се симптоми и њихова јачина и према томе се може предвидети шанса за излечење особе и резултат деловања лека.

❖ Уколико лекар или дефектолог жели да унапред одреди да ли је нека апликација или лек добар за пацијента он може на основу одређивања оцене психичког стања особе пре употребе апликације односно лека, одредити уз помоћ ове технике, која ће бити оцена реакције особе на апликацију или организма на лек. Потребно је само пре тога извршити истраживање на тему коју желимо да испитамо. Погледајмо следећи пример.

Пример 1. Нека је дефектолог пре коришћења креиране апликације одредио да је оцена психичког стања пацијента 4. Уколико се зна да ће дефектолог понудити родитељима детета ову апликацију као терапијско помагало ако и само ако очекује да ће оцена реакције детета бити већа од 3, испитати да ли би требало да лекар понуди родитељима детета ову апликацију као терапијско помагало.

Решење: Како знамо једначину функције линеарне хипер-равни следи да важи:

$x + 3y - 12 = 0$. Како је према услову примера $x = 4$ следи да је $y = \frac{8}{3} < 3$. Дакле не би требало да лекар понуди ову апликацију као терапијско помагало иако особа има високу оцену психичког стања.

❖ Овим методом може се:

- ❖ на основу неких података о коришћењу апликације (дејству лека) предвидети групације којима особа припада;
- ❖ предвидети каква ће бити реакција особе на крају апликације и знати да ли јој треба дозволити да користи неку апликацију;
- ❖ може се предвидети очекивани опсег реакције детета или организма на неки лек;

Метод класификације SVM класификатором веома је битан метод којим се, као што је показано у раду, може побољшати начин рада здравства и фармацеутске индустрије, трговине, пословања, економије, биоинформатике и многих других области у чему се и огледа значај овог метода.

3.3.3.3 Цртање графика и одређивање нелинеарне хипер-равни

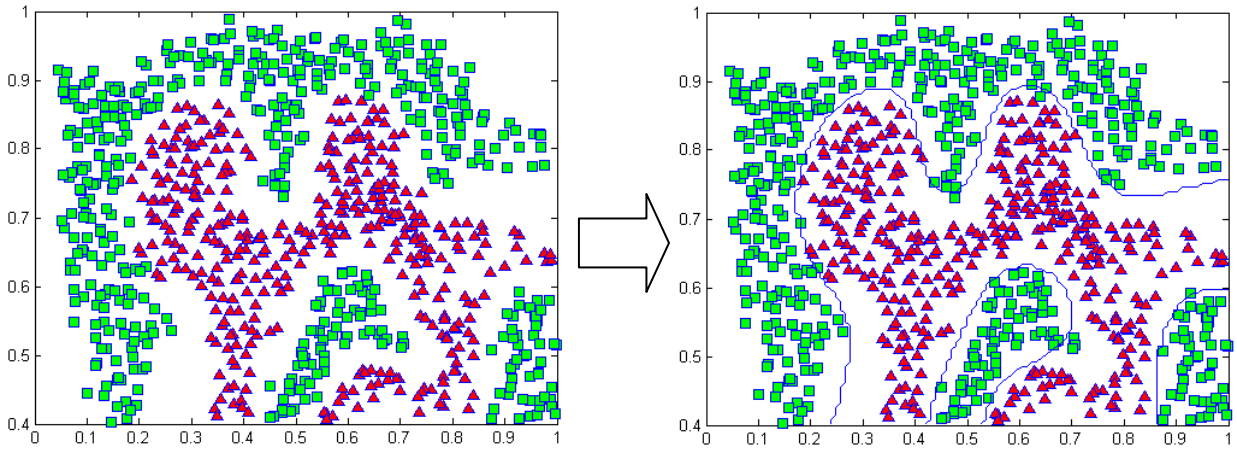
У другом и трећем делу апликације приказана је класификација линеарно нераздвојивих података Гаусовим кернелом. Гаусов кернел је функција облика:

$$K_{\textbf{gausov}}(x^{(i)}, x^{(j)}) = \exp\left(-\frac{\|x^{(i)} - x^{(j)}\|^2}{2\sigma^2}\right) = \exp\left(-\frac{\sum_{k=1}^{n}(x_k^{(i)} - x_k^{(j)})^2}{2\sigma^2}\right)$$

Погледајмо исечак кода којим се раздвајају подаци другог графика.

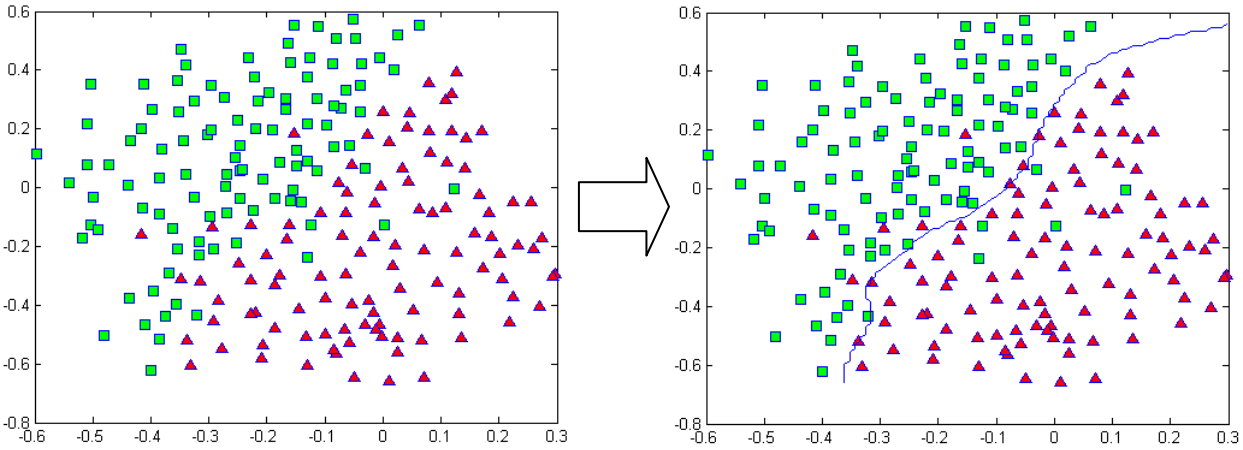
```
%% ====Ucitavanje i prikaz podataka iz druge tabele====
fprintf('Ucitavanje i prikaz sledeceg grafika \n')
load('podaciizstrazivanja2.mat');
crtajPodatke(X, y);
fprintf('Program zaustavljen. Pritisnite enter da nastavite. \n');
pause;
%% ==========Treniranje SVM-a sa Gausovim Kernelom==========
fprintf('Treniranje SVM-a sa Gausovim Kernelom (ovo moze trajati 1 do
2 minuta) ...\n');
load('podaciizstrazivanja2.mat');
% Unesite parametre
C = 1; sigma = 0.1;
model= svmTrain(X, y, C, @(x1, x2) gausovKernel(x1, x2, sigma));
prikazihiperravan(X, y, model);
fprintf('Program zaustavljen. Pritisnite enter da nastavite. \n');
pause;
```

Неке од функција објашњене су у поглављу 3.3.3.1. Модел који се креира преко функције svmTrain јесте средство које нам после помаже при цртању хипер-равни. Видимо да модел зависи од параметара C и сигма који регулишу тачност (прецизност) постављања хипер-равни. Функција prikazihiperravan(X, y, model) црта хипер-раван на графику.
На следећој слици приказан је график 2 и оптимална хипер-раван која раздваја податке на њему.



Слика 18. Одређивање хипер-равни код резултата другог дела истраживања

На следећој слици приказан је график 3 и хипер раван која га раздваја.



Слика 19. Одређивање хипер-равни код резултата трећег дела истраживања

Поред машинског учења и SVM класификатора постоје многи други методи за класификацију и обраду података. У наставку рада представићемо вам основе data, text и duo mining-а као и моделовање природних језика који представљају једне од основних метода обраде и филтрирања података данас.

4. Data mining

4.1 O data mining-y

Живимо у времену брзих промена услова у пословној средини што многим фирмама поставља задатак да непрестано прате шта то ради њихова конкуренција, како би успели да се одрже поред конкуренције. Како је број фирми, а самим тим и података које је потребно обрадити велики, многе фирме немају довољно адекватно обученог особља које би обрађивало те податке. Зато се овакви послови препуштају специјализованим програмима који обрађују податке на веома брз и квалитетан начин. Сви програми овог типа могу се сврстати у Business Intelligence програме којима се омогућава боља, бржа и квалитетнија обрада и филтрирање података, а самим тим се побољшава начин пословања многих фирми. Data mining јесте један од Business Intelligence производа. Једна је од нових технологија које су се крајем педесетих година прошлог века развиле са развојем научних техника и рачунарских програма и метода. Не постоји прецизна дефиниција овог појма, али неке од њих су:

- Data mining представља процес проналажења скривених законитости и веза међу подацима. То је поступак издвајања потенцијално корисних, интересантних и нових информација и законитости из структуираних података, а све у циљу бољег пословања.
- Други назив за data mining је Knowledge Discovery in Databases-откривање знања у базама података. Data mining даје могућност сагледавања информација на начин који раније није био могућ.

Приликом обраде података из великих база података data mining програми могу:

1. утврђивати особине које се јављају заједно код више података (које производе фирма набавља заједно у једној набавци) (**асоцијација-association**);

2. откривати скривене везе и елементе (функције) за њихово груписање у неку од класа (**класификација-classification**);

3. одређивати групе података који су међусобно веома слични, али различити од осталих група података (**кластеровање-clustering**);

4. открити и пратити понашање посматраног предмета током времена, а самим тим и предвиђати и одређивати, користећи се правилностима из унетих примера, различите очекиване нумеричке вредности (**предвиђање-scoring, predicting**).

Сваки data mining пројекат састоји се од следећих фаза:

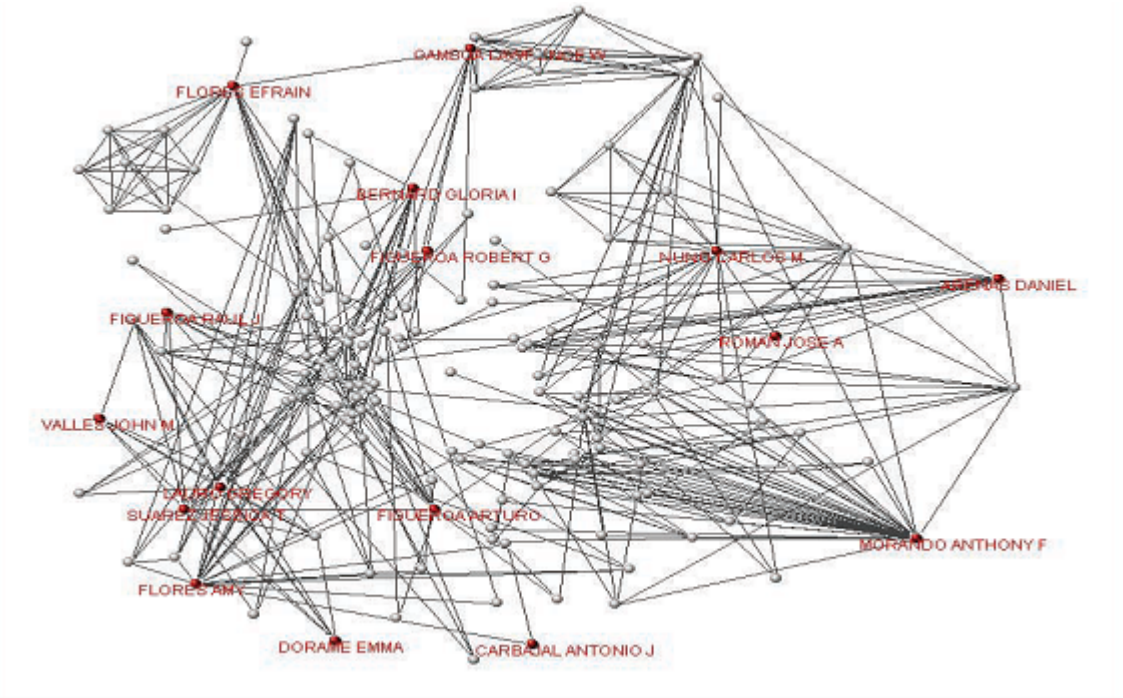
1. Прикупљање података-подаци морају бити ускладиштени у базе података тако да програм може да их анализира;
2. Филтрирање података и њихова трансформација-одстрањују се сувишне информације, непотпуни и дуплирани подаци, а након сређивања података долази до њихове трансформације у типове података са којима програм може да ради;
3. Креирање и избор модела-креира се након што су подаци прошли кроз прве две фазе. За креирање модела потребно је познавање циља data mining пројекта и поседовање одговарајућих алгоритама којима се модели могу направити;
4. Процена квалитета модела-након креирања модела мора се извршити његова евалуација (оцењивање);
5. Креирање извештаја-када је модел креиран и процењен евалуацијом као добар може се креирати извештај који се даје фирми на увид. Постоје два основна типа извештаја: извештај о пронађеним обрасцима и извештај о предвиђеним вредностима модела;
6. Оцењивање модела-припрема за употребу пронађеног модела и оцена предвиђања на основу тренираног модела и скупа нових података;
7. Интеграција data mining модела у апликацију-примена пословне интелигенције на пословни систем односно укључивање модела у апликације које се користе у фирмама и његово коришћење;
8. Управљање моделом-одржавање статуса модела је велики изазов зато што сваки data mining модел има свој животни циклус (због промена у средини). У неким областима примене, модели су стабилни, али у неким, попут економије и трговине, се стално мењају, те је потребно непрестано креирати нове и боље верзије модела.

4.2 Примена data mining-а

Уколико изузмемо велике компаније, данас постоји велики број фирми које не користе data mining приликом обраде података. Разлози који се најчешће помињу јесу велики трошкови, недостатак квалификованог особља и неразумевање рада data mining програма. Следе неке од основних примена data mining програма.

- осигурање-за предвиђање нивоа одштетних захтева и спречавање превара;

- банкарство-за утврђивање ризика код кредитних картица и предвиђање зараде од нових клијената;
- трговина-за спречавање крађа и превара и утврђивање плана набавке код малопродаје, као и за одређивање оптималних залиха у магацину;
- полиција- за праћење учесника злочина и њихово повезивање и лоцирање. На следећој слици налази се полицијски извештај након испитивања једног криминалног случаја. Data mining-ом успешно су повезани сви чланови различитих кланова, а на врховима графа налазе се њихови шефови;



Слика 20. Извештај полиције направљен data mining процесом

- маркетинг-утврђивање трендова, предвиђање понашања потрошача, утврђивање метода за спровођење директног маркетинга.

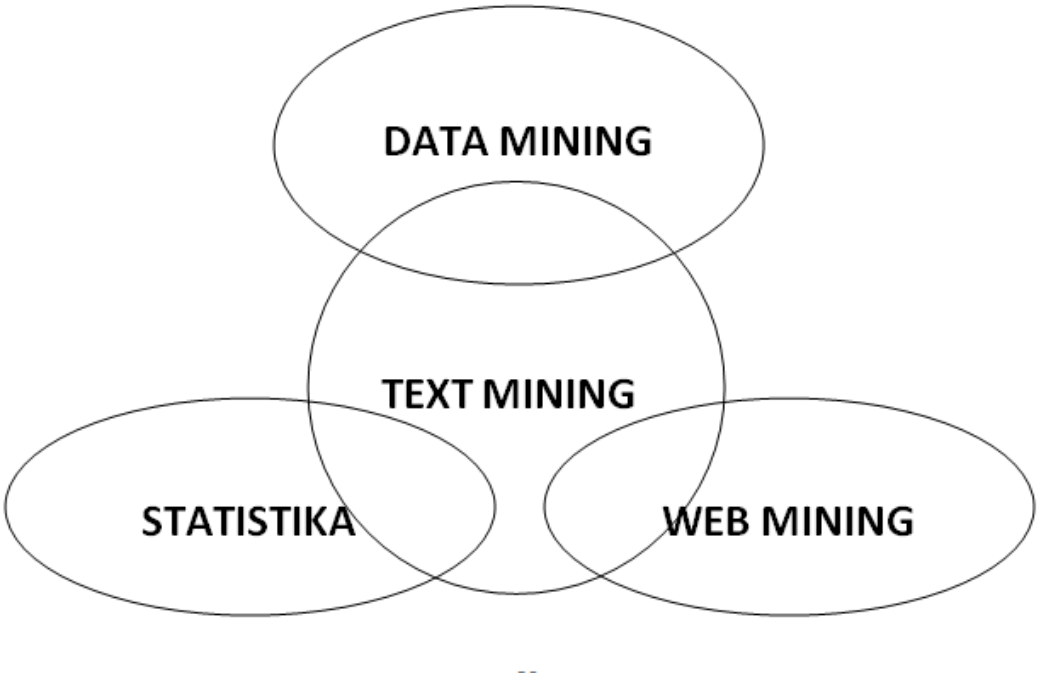
У следећем поглављу упознаћемо се са text и duo mining-ом.

5. Text и duo mining

5.1 О text mining-y

Text mining представља процес проналажења образаца и филтрирања података унутар неструктуираних скупова података (књиге, белешке и документа). Још једна од дефиниција text mining-a је: Text mining је процес који рачунарима даје могућност извођења закључака и информација из великих количина неструктуираних скупова података. Text mining често при обради користи семантичке анализе и таксономију и тиме усклађује статистичке податке и вештачку интелигенцију. Као и код data mining-a и text mining алати понекада могу захтевати од корисника да ручно именују документе и да им одреде тип. Ипак са развијањем овог процеса појавили су се софтвери које је могуће, преко података за тренинг, обучити како да препознају тип унетог документа. Text mining може послужити као средство за одређивање кључних образаца унутар великих скупова докумената, а самим тим може помоћи пословању неке фирме.

У поређењу са data и web mining-ом text mining представља бржи процес зато што су подаци код база података и интернет страница структуирани, а text mining користи податке из природних језика.



Слика 21. Text mining и његов однос са другим областима

Сваки text mining пројекат састоји се од следећих фаза:

1. Прикупљање података-докумената и списа који ће бити обрађени;
2. Препроцесирање података;
3. Трансформација текста-генерисање атрибута и грађење структуре;
4. Data mining-откривање образаца унутар текста и пролазак кроз све фазе data mining-а;
5. Евалуација интерпретираних резултата.

5.2 Примена text mining-а

Text mining се данас примењује у великом броју државних, али и пословних истраживања. Апликације засноване на text mining-у могу бити сортиране у више категорија по врсти анализе или пословне функције. Користећи овај приступ за класификацију решења, неке од примена text mining-а су:

- код предузећа-за предвиђање будућих корака у пословању и праћење рада конкуренције;
- за управљање записима података;
- безбедност-за управљање подацима обавештајних служби;
- за испитивање тржишта;
- код праћења трендова друштвених мрежа;
- код обраде медицинских извештаја;
- за извлачење кључних речи неког текста-највише се користи за испитивање збирке научних радова када се проучавају поглавља *резиме* и *закључак и дискусија* и тако се праве трендови-кључне речи збирке научних радова.

5.3 O duo mining-y

Као засебни процеси data и text mining постојали су годинама. Међутим, аналитичари су дошли до фантастичних резултата након што су спојили ова два процеса. Недавно су произвођачи као што су IR (Intelligent Results), SAS (Statistical Analysis System) и SPSS (Statistical Product and Service Solutions) почели да саветују својим клијентима да комбинују data и text mining. Због чега је ово добро?

Предузећа која су клијенти ових произвођача почела су нагло да проширују своје видике и да повећавају количине информација које користе и добијају комбинујући data и text mining.

Колекције и одељења за опоравак банака и компанија за кредитне картице користе duo mining са веома добрим ефектом. Коришћењем data mining-а банкари могу да прате стања рачуна неких фирми или појединаца и да претпоставе, након неког времена, колике су шансе да ће посматраној особи или појединцу бити потребан кредит. Када су подаци агената са шалтера додати у обраду података разумевање и резултати постали су још јаснији. На пример text mining-ом рачунар може да разуме разлику између „Ја ћу платити“, „Нећу платити“ и „Ја сам платио“ и да одреди склоност (број-проценат) да ће поменути клијент платити што може бити обрађено коришћењем data mining-а.

Коришћењем duo mining-а (комбиновањем data и text mining-а) фирме су добиле могућност да повећају своју зараду за око 20% на месечном нивоу.

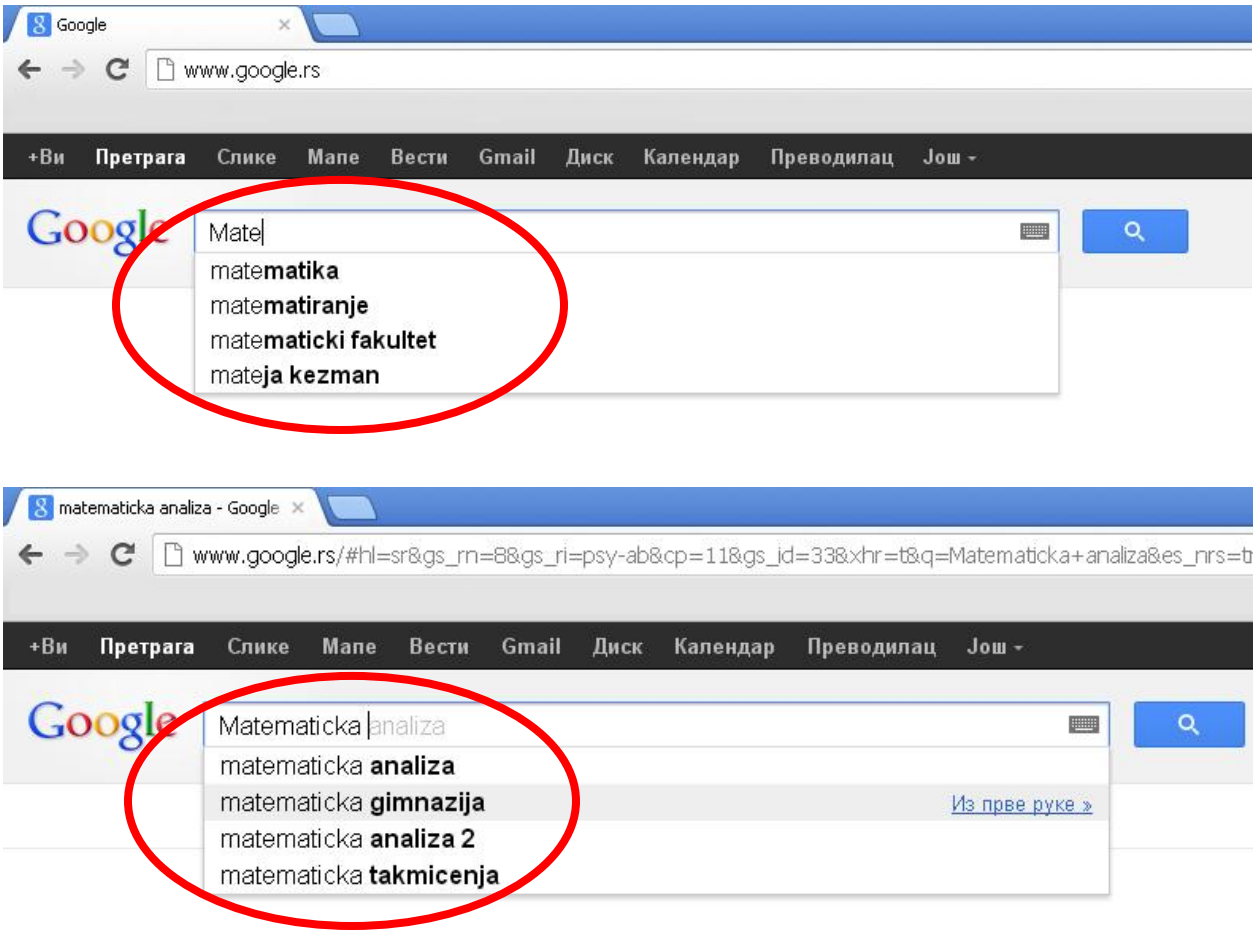
Остале области где се коришћење duo mining-а показало као веома корисно и исплативо јесу: анализирање листа жеља приликом куповине на интернету, истраживање одговора на анкете и приговоре клијената на рад неке фирме, и распоређивање тарифа код корисника неког оператера на основу њихових жеља.

У наставку рада приказаћемо вам једну веома интересантну област примене data и text mining-а, а то је моделовање природних језика.

6. Модели природних језика

Велики број области рачунарства и вештачке интелигенције, као што су машинско превођење, издвајање информација, креирање и употреба експертских система, разумевање текста, препознавање и генерисање говора и сличне, почивају управо на техникама обраде природних језика.

Свако од вас вероватно је користио интернет претраживаче, а самим тим и приликом уношења кључних речи приметио да му се као опције за унос нуде готови обрасци на које може да кликне и да скрати време уношења података за претрагу. Погледајмо следећу слику која приказује поменути случај.

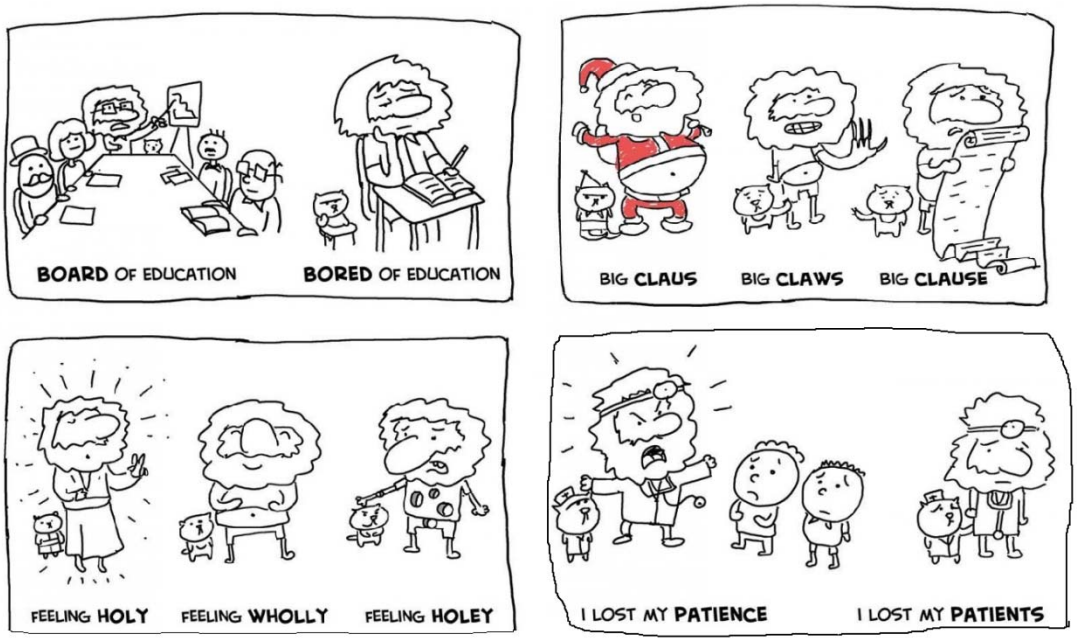


Слика 22. Пример претраге интернета

Аутоматско препознавање говора спада у највеће техничке изазове савременог доба и већ више од пола века заокупља пажњу истраживача широм света. Задатак аутоматског препознавања говора је да се говор преведе у текст, односно да се „препозна“ шта је особа изговорила.

Одређивање излазне реченице на основу улазног сигнала није нимало једноставно. На пример, реченице „Стигли смо у Бар“ и „Стигли смо у бар” ће звучати потпуно исто и није могуће само на основу улазног сигнала једнозначно одредити излаз.

Ове реченице зову се хомофони.



Слика 23. Хомофони у енглеском језику

Уколико је човек тај који слуша и препознаје, он ће користити нека додатна знања о језику и свету који нас окружује како би одредио право значење изговорене реченице. Међутим, уколико је рачунар тај који треба да препозна изговорено, ситуација се знатно компликује.

Проблем препознавања говора, а и многи други слични проблеми, решавају се употребом вероватноће како би се формирао *модел природног језика* – интерпретација неког језика помоћу које рачунар одређује који је највероватнији излаз за дати улазни сигнал.

Наиме, циљ је израчунати вероватноћу P(реченица) = P(реч_1 , реч_2 , реч_3 , … , реч_n) да ће се нека реченица или секвенца речи појавити у неком природном језику.

Овај проблем можемо посматрати и као проблем одређивања вероватноће следеће речи у реченици, када су познате претходне P(реч_5 | реч_1 , реч_2 , реч_3 , реч_4).

На пример, изговорена реченица „Покажи ми твој рад“ могла би да приликом проласка кроз систем звучи као: „Покажи ми твој рад“, „Покажи ми твој град“ , „Покажи ми твој јад“, ... Нека је L = {*покажи, ми, твој, рад, јад, ...*} скуп речи неког језика (или дела језика). Са L* обележићемо скуп свих могућих варијација елемената из L, односно:

L* = {*покажи, покажи ми, покажи ми твој, ми твој рад, покажи ми твој рад, покажи покажи ми*, ...}. Скуп L* има бесконачно много елемената. Међутим, нису сви његови елементи исправне реченичне конструкције. Неки елементи скупа L* ће се појављивати често у текстовима, неки од њих ће се појављивати ретко, а неки неће никада. Приликом креирања модела, из неког унапред расположивог скупа текстова на природном језику, управо се на основу броја појављивања неке секвенце речи рачуна њена вероватноћа јављања. Ова информација се касније користи за одређивање највероватнијег излаза система. Дакле, циљ је одредити функцију P која задовољава следеће:

$\sum_{x \in L^*} P(x) = 1, P(x) \geq 0 \ \ \forall x \in L^*$, а на основу формуле условне вероватноће важи да је:

P(Покажи ми твој рад) = P(Покажи)·P(ми|Покажи)·P(твој|Покажи ми)·P(рад|Покажи ми твој).

Функција P се одређује полазећи од неког унапред познатог скупа текстова којим је машина „тренирана“ односно који су скенирани и обрађени у рачунару.

Уколико бисмо покушали да интуитивно рачунамо вероватноће секвенци речи као:

$P(\text{Покажи ми твој рад}|\text{Покажи ми твој}) = \frac{\text{бројПојављивања(Покажи ми твој рад)}}{\text{бројПојављивања(Покажи ми твој)}}$ било би пуно секвенци које се нису ни једном појавиле у тексту намењеном за тренирање система, а сасвим су исправне и могуће у неком језику.

Није могуће имати тако велики скуп текстова који би обухватио све могуће реченице неког језика. Због тога се користе Марковљеви модели. Најједноставнији је униграм модел, који вероватноћу појављивања неке секвенце (реченичног дела) рачуна као производ вероватноћа појављивања речи те секвенце: $P(x_1 \dots x_n) = \prod_{i=1}^{n} P(x_i)$

Ако пак вероватноћа појављивања неке речи у реченици зависи само од вероватноће појављивања прве речи која јој претходи такав модел називамо биграм. Тада важи:

$P(x_1 \dots x_n) = \prod_{i=2}^{n} P(x_i|x_{i-1})$, где је $P(x_i|x_{i-1}) = \frac{\text{бројпојављивања}(x_{i-1}, x_i)}{\text{бројпојављивања}(x_{i-1})}$ (*).

На следећој страни приказан је практичан пример употребе модела природних језика.

Пример 2: Нека су из великог броја текстова који се уносе при „тренирању“ рачунара обрађене следеће речи приказане у две табеле на слици 4. Ознаке <s> и </s> означавају почетак и крај реченице.

	<s>	Ја	волим	да	једем	</s>
број јављања	1000	923	905	374	920	1000

	Ја	волим	да	једем	</s>
<s>	830	59	51	12	3
Ја	**0**	704	1	61	15
волим	**0**	**0**	253	**0**	23
да	21	57	1	321	1
једем	2	**0**	51	**0**	901

Слика 24. Табеле са неким речима скенираног текста

У првој табели приказане су појединачно неке речи из текста и њихов број јављања у тексту. У другој су приказани парови речи и њихов број јављања у тексту. Одредимо, на основу текстова унетих у рачунар, вероватноћу да ће се у српском језику појавити реченица „Ја волим да једем“.

Из формуле (*) следи да важи:

P(<s> Ја волим да једем </s>) = P(Ја|<s>)·P(волим|Ја)·P(да|волим)·P(једем|да) ·P(</s>| једем).

где је: P(Ја|<s>) = $\frac{830}{1000}$, P(волим|Ја) = $\frac{704}{923}$, P(да|волим) = $\frac{253}{905}$, P(једем|да) = $\frac{321}{374}$, P(</s>|једем) = $\frac{901}{920}$.

Одавде је P(<s> Ја волим да једем </s>) = 0.15.

На сличан начин можемо проширити модел на 3−граме, 4−граме, ... , n−граме, где су n−грами модели код којих вероватноћа јављања неке речи зависи од вероватноће појављивања n − 1 претходних речи. У пракси се заснивање модела језика на n−грамима показало као веома ефикасно. Приликом моделовања природних језика, секвенци која није прочитана додељују се неке веома мале вероватноће јављања тако да је свака реченица вероватна, али оне које се не срећу често имају веома мале вероватноће.

Добар пример коришћења модела природних језика јесу различити преводиоци на интернету који генерешу највероватнију секвенцу страног језика као резултат превода унетог текста.

7. Закључак и дискусија

Као што се може закључити, машинско учење, data, text и duo mining су својом неизмерно широком употребом заслужили место међу водећим дисциплинама данашњице. Само изучавање ових области захтева много времена и рада, делом због математичког апарата који се користи, а делом и због њихове широке примене.

Машинско учење, data, text и duo mining су технике и процеси за филтрирање података који ће се и у будућности све више развијати и имати све већу практичну примену и у другим дисциплинама.

Рад указује на велике могућности даљег истраживачког рада на овом подручју, усавршавања апликација за филтрирање података и њихова израда у другим програмским језицима.

8. Захвалност

Захваљујем се својој менторки Јелени Хаџи-Пурић са Математичког факултета за указану помоћ и подршку при изради овог матурског рада. Захваљујем се такође дефектологу и професору информатике за особе са сметњама у развоју Мири Стевановић и професорима Специјалне основне школе „Нови Београд“ на указаној помоћи при изради апликације Захвалио бих се и професоркама Математичке гимназије Невенки Спалевић и Јелени Поповић за указану помоћ при избору литературе и адекватних приручника на интернету који су ми помогли да направим апликацију.

Хвала вам свима на указаној помоћи!

9. Литература

[1.] I. Stanley Greenspan, M. D. Serena Wieder, R. Simsons, *The Child With Special Needs: Encouraging Intellectual and Emotional Growth,* A Merloyd Lawrence book, Massachusetts,1998;

[2.] A. Smola, S. V. N. Vishwanathan, *Introduction to Machine Learning,* Cambridge University Press, Cambridge, 2008;

[3.] E. Alpaydin, *Introduction to Machine Learning*, The MIT Press, Massachusetts, 2010;

[4.] I. H. Witten, E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques,* Morgan Kaufmann Publishers, San Francisco, 2005;

[5.] C. M. Bishop, *Pattern Recognition and Machine Learning,* Springer, New York, 2006;

[6.] R. Feldman, J. Sanger, The *Text Mining Handbook*, Cambridge University Press, Cambridge, 2006.

[7.] W. McKnight, *Building business intelligence: Text data mining in business intelligence*, y DM Review, Miller, 2005.

[8.] D. Jurafsky, M.H. James, *Speech and Language Processing: An Introduction to Natural Language processing, Speech Recognition and Computational Linguistics*, Prentice Hall, 2009;

[9.] A. Kao, S. Poteet, *Natural Language Processing and Text Mining,* Springer, New York, 2007.