Предавање 1

* Аналитика – процес на анализирање на податоците за да се извлечат вредни согледувања

1. Дискриптивна – опис на минатите настани
2. Дијагностична – разбирање на причините
3. Предиктивна – предвидување на идни резултати
4. Прескриптивна – препорака за акција
5. Когнитивна – предизвикување нешто да се случи

Data Science – комбинација од статистика/математика, комп. Науки и машинско учење за да се извлече знаење од податоците

Business Inteligence – процес на собирање, обработка и претставување на податоци за да се олесни донесувањето на одлуки

Machine Learning – подобласт на вештачката интелигенција која користи алгоритми за да се обучи модел со податоци и да се направи предвидување или одлука без експлицитно програмирање

Вештачка интелигенција – подрачје на комп. Науки што се фокусира на разбирање системи кои можат да извршуваат задачи што нормално ги бараат луѓето

Предавање 2

Data Science Process – итеративен процес

1. Поставување цел за истражување
2. Превземање податоци – внатрешни/надворешни
3. Припрема на податоци

* Прочистување
* Трансформација
* Комбинирање податоци

1. Истрањување со помош на графови, комбинирани графови...
2. Моделирање на податоците
3. Презентирање и автоматизација

Big Data - е концепт кој се однесува на огромни количини на податоци што се создаваат со голема брзина, од различни извори и во разл. Формати

Otlier – податок што е надвор од нормалната дистрибуција

* Можат критично да влијаат
* Најлесно за да се видат е ако користиме plot/histogram

Моделирање може со:

* Линеарна регресија
* KNN – прави врз база на соседи

Confusion matrix – покажува колку случаи биле точно класифицирани со споредување на предвидуваните вредности со реалните

Предавање 3

Карактеристики на денешните податоци

* Volume – колку има?
* Variety – колку разл. Податоци постојат?
* Velocity – со која брзина се генерираат?

Начина на собирање на податоци:

* API – сет на prebuild функции развиени од дадена компонента
* RSS – Rich Site Summary – сумаризира често ажуриран онлајн контент во стандарден формат
* Web Scraping – процес на автоматско собирање податоци преку софтвер, скрипти врз тоа што е прикажано, со извлечување HTML

Batch data – зачувани во даден интервал и претставуваат јавни податоци

Dark data – скриени податоци кои имаат потенцијал да креираат нови вредности (податоци кои ги има компанијата но не ги презентира како официјални податоци)

Stream - контирнуирана обработка во реално време

Lambda архитектура – архитектура за процесирање податоци која се справува со огромни количини на податоци преку batch i stream методите

Digital Twins – дигитална репрезентација на реалниот свет или систем

Data Lake – складиште на податоци што овозможуваат складирање на огромно количество податоци во нивна оригинална форма

Променливи:

* Квантитативни – дискретни (конечен бр. Вредности се ограничени во секој интервал)

-континуирани (безконечен бр. Во секој интервал)

- Категориски

Популација – целосен сет на објекти или настани кои се испитуваат

Примерок – дел од податоците кои се испитуваат

Предавање 4

KNN е едноставен и логички метод за предвидување

* одбери ја вредноста за к, број на најблиски точки
* за секоја точка во тестирачките податоци:

1. калкулирај ја сличноста меѓу тест и секој ред од тестирачките податоци
2. базиран на вредност на растојание, сортирај ги во растечки
3. земи ги првите к реда
4. доделете класа на тест-точката врз основа на најчестата класа на овие редови
5. обично се зима најблиското според Евклидово растојание

Типови на missingness

1. missing completely at random – податоците недостасуваат целосно на случаен начин без никаква врска со останатите променливи или самите вредности што недостасуваат

пр. Датата се оштетила и некои записи недостасуваат

1. missing at random – недостигот не е поврзан со некоја др. Променлива во податоците, но не и со самите вредности што недостасуваат

пр. Испитаници со понизок приход почесто избегнуваат да одговорат на прашања поврзани со тоа

1. missing not at random – најлошото сценарио / недостигот е директно поврзан со самите вредности што недостасуваат

пр. Луѓе со повисоки приходи не одговараат на прашања за приход затоа што тоа е чувствителна информација

Справување со недостиг на податоците

* отстранување на вредноста
* постави вредност на null
* вметни статичка вредност како 0 или просек
* вметни вредност од приближна или теоретска дистрибуција
* моделирај ја вредноста

Справување со outliers

* бришење
* трансформирање
* truncation

Векторизација е процес на претворање на податоците во нумерички формат (вектори) за да можат да се користат како влез во алгоритми за машинско учење

Encoding

* Label encoding – категориските податоци се претвараат во целобројни вредонсти (кога има природен редослед)

Пр. лошо = 0, добро = 1, одлично = 2

* One – hot encoding – секоја категорија се претвора во бинарен 1/0 формат (кога нема природен редослед)

Пр. имиња на градови

* Binary encoding – категориските податоци се претвараат прво во бројки, а потоа тие бројки во бинарен формат (кога има мн колони и One – hot би заземал мн просто)
* Proxy – ја заменува секоја категориска со нумеричка вредност базирана врз некоја proxy метрика

Пр. просечна цена, времетраење услуга...

Data SCHEMA denormalization – комбинирање податоци од различни дата извори во една табела

Нормализација – ги скалира сите нумерички варијабли во опсег 0-1

* Кога податоците немаат нормална распределба
* Кај KNN се користи

Стандардизација – ги трансформира сите податоци така што ќе имаат средна вредност 0 и стандардна девијација 1

* Кога податоците имаат нормална или приближно нормална распределба
* Се користи кај линеарна или логистичка регресија

Предавање 5

Типови на учења:

1. Супервизирано – учење од податоци кои содржат етикети (labels) или очекувани излези

Моделот учи од познати примери од минатото за да предвиди

Пр. линеарна регресија, невронски мрежи, decision tree

1. Несупервизирано – учење без labels каде целта е да се пронајдат шаблони или структури

Пр. кластеризација

1. Reinforcement learning – преку засилување-учење преу награди и казни базирани на низа на акции

Компоненти на ML:

1. Representation – дефинира форма и структура на моделот кој ќе биде користен за да се научат податоците и да се донесат предвидувања
2. Evaluation – се однесува на метриките и техниките кои се користат за да се процени како моделот ги извршил задачите

-за класификација: accuracy, precision, recall, F1 score

-за регресија: mean squared error, R2

3. Optimization – процес на подобрување на перформансите на алгоритмот

R squared – коефициент на детерминација

* Ако моделот дава скоро исти резултати како средната вредност враќа 0
* Ако моделот е перфектен враќа 1
* Ако моделот има полоши резултати од средната вредност враќа негативна вредност

Bootstraping – статистичка техника која се користи за проценка на својства на еден статистички показател преку создавање на повеќе примероци од одреден податочен сет

Типови грешки

1. Тип 1 грешка – отфрлање на нултата хипотеза иако е вистинита
2. Тип 2 грешка – не ја отфрламе нултата хипотеза иако е лажна

Тестови параметарски податоци со нормална дистрибуција

1. T – test – да се споредуваат процентите на 2 групи и да се утврди дали постои значајна разлика меѓу нив
2. Chi – squared – да се утврди дали постои врска меѓу две категориски променливи
3. ANOVA – да се споредат процентите на 3 или повеќе групи за да се утврди дали постои значајна разлика меѓу нив

Класификација – предвидување на категориски излезни променливи базирани на влезни податоци

* За небалансирано податочно множество се користи F1 score
* Confusion matrix – колку точно и погрешно моделот ги класифицирал примероците

-TP

-TN

-FP

-FN

- Precision – мери колку од податоците предвидени како позитивни се навистина позитивни

- Recall – мери колку од вкупно вистинските случаи моделот успеал да идентификува

- F1 score – комбинација на precision и recall

Предавање 6

Полиномална регресија – моделирање на нелинеарна регресија помеѓу зависната променлива (output) и еден или повеќе независни променливи (input) користејќи полиноми на влезната вредност

Overfiting – ако полиномалниот степен е превисок, моделот може да се навикне на тренинг податоците и да води до генерализација

Cross Validation – техника на евалуација на перформанси која го дели достапното податочно множество на повеќе под-множества за да се тестира моделот на невидени податоци

-се користи кога има мало податочно множество

K-fold

1. Податоците се делат на к еднакви делови
2. Моделот се обучува к-1 пати на различни делови и се тестира на преостанатиот дел
3. Се повтара к пати и конечниот резултат е просек од сите тестови

Bias – грешка што се јавува кога моделот е премногу едноставен за да ја долови сложеноста на податоците

Variance - грешка што се јавува кога моделот е премногу сложен и премногу добро се прилагодува на тренинг податоците

Regularization – техника да се спречи overfiting со додавање казна за комплексност на моделот во функција за загуба

Decision Tree – дрво структура за донесување одлуки врз основа на влезни податоци

1. Се делат податоците врз основа на некој критериум за да се созсдадат подгрупи

-на секој јазол се одбира најдобрата карактеристика

2. Секој јазол е поделен на гранки кои ги претставуваат можните вредности

3. Крајните точки се викаат лисја и ја претставуваат класата (класификација) или предвидената вредност (регресија)

-root node – корен

-information gain (IG) – колку се намалува ентропијата

- gini index – мерка за чистота на јазолот

ID3 алгоритам – работи само со дискретни вредности

* Работи според IG

Ентропија мери несигурност или хаос во даденото податочно множество

Предавање 7

Bagging – е техника на ансаблирање во машинското учење која се користи за подобрување на стабилноста и точноста на моделите. Тоа го намалува overfiting и ја зголемува робусноста на моделот преку комбинација на повеќе слаби модели во еден посилен продукт-модел

* Bootstrap – од дадено множество бираме помали множества
* Aggregate – агрегација на Bootstrap-от

1. Се генерираат подмножества од податоци со Bootstrap (повеќе различни)
2. Тренирање на повеќе модели каде за секое подмножество се тренира независен модел
3. Агрегација

-за регресија – се зема пресекот од сите предвидувања

-за класификација – се користи гласање на множества

Во Bagging податоците што остануваат неупотребени во тренирачкиот сет се нарекуваат Out-of-Bag податоци

Bagging користи ДРВА

Random Forest e алгоритам што комбинира повеќе дрва на одлучување за да креира помоќен и попрецизен модел

* Кај Random Forest се работи со случаен број на подмножества, но и случаен број карактеристики

Параметри:

* n-estimators
* max-depth
* max-featurs
* min-samples-split
* min-samples-leaf

Boosting – техника на ансамблирање која комбинира повеќе слаби учесници во еден силен

* за разлика од Bagging, во неа секој последователен модел се тренира на грешките од последниот модел

Два главни типа: - adaptive(AdaBoost) и -gradient(XGBoost)

AdaBoost функционира така што секој модел во ансамблот добива тежина пропорционална на неговата тежина

Gradient ги минимизира грешките со оптимизирање на функцијата на загуба преку градиентен спуст

* XGBoost – има регуларизација и спречува overfitting,
* CatBoost – оптимизиран за категориски променливи

Баесов класификатор – се користи во проблеми со класификацијата, особено кога податочните множества имаат високи диммензии

* Не му сметаат missing values

--------->R^2 -> R squared <-----------------

R^2 е метрика односно коефициент на детерминација која се користи кај модели кои користат регресија(линеарна регерсија).Покажува колку добро моделот ја објаснува променливата која ја предвидуваме Y врз основа на независната променлива X

Тоа е вредност од -1 до 1 каде 1 значи дека е перфектно,додека ако е под 0 тоа значи дека моделот е полош од просекот.0 е просекот и вредностите помеѓу 0 и 1 се сметаат за добри вредности за моделот.

--------->Класификација<--------------------

Класификација е модел кој се користи кај категорички променливи за предвидување на променливата Y врз основа на множеството на независни променливи X. Класификацијата може да биде бинарна доколку предвидуваме на пр. даден човек дали е маж или жена и може да е мултилабел каде имаме повеќе од 2 вредности.

--------->Важни мерки за перформанси на модели (Accuracy, Precision, F1-Score, Recall)<-----------------

Доколку предвидуваме модел кој е класификација односно предвидуваме променлива Y која е од категорички тип тогаш се користат следните метрики за дознавање на повеќе информации за тоа колку моделот е добар:

-Accuracy ни кажува колку моделот е точен,и тој се изразува со формулата TP/(TP+FP),каде TP значи колку точно позитивни случаи има врз бројот на сите позитивни(точно позитивни и грешка позитивни).Ни дава информација колку од предвидените позитивни случаи се точни навистина.

-Recall ни кажува колку моделот е чувствителен.Тој се исразува преку формулата TP/(TP+FN) односно сите точно позитивни случаи врз збирот на истите со грешка негативните податоци.Ни дава информација колку од вкупното вистински податоци моделот успеал да идентифицира.

-F1-Score e комбинација од двете погоре и од него можеме да видеме дали даден датасет е балансиран или не.

-Accuracy ни кажува колку моделот што сме го предвиделе е точен. Се изразува како (TP+TN)/allSamples

--------->Што е стандардизација, што е нормализација?<--------------

-Стандардизација е процес кој ги скалирање на податоците така што тие ќе имаат средна вредност=0 и стандардна девијација = 1. Се користи кога има податоците имаат нормална распределба или приближно нормална. Се користи кај линеарна агресија или логистичка.

-Нормализација за разлика од стандардизација се користи кога нема нормална распределба и се скалираат податоците во опсег од 0 до 1. Ова го правиме со цел да се осигураме дека карактеристиките ни се во ист опсег. Ова го правиме бидејќи многу алгоритми во машинското учење учат подобро и работат подобро кога карактеристиките се слични.

--------->Модел на линеарна регресија<--------------

-Моделот на линеарна регресија е модел кој се користи кога податоците кои сакаме да ги предвидеме односно променливата Y се нумерички односно континуирани податоци. Тука можеме да прикажеме дека има корелација на променливата Y и независните променливи X. Целта на овој модел е да ја предвиде зависноста меѓу променливата Y и независните променливи X. Oвој модел користи равенка на права линија за да ги предвиде вредностите.

пр. кога сакаме да предвидеме како плоштината на една куќа ќе влијае на нејзината цена.

--------->КNN модел за класификација<----------------

КНН односно K Nearest Neighbords е најчестиот користен модел кога се работи со податоци кои се категориски односно класи. На пример кога предвидуваме дали даден студент би имал оценка одличен, добар или доволен. КНН работи на принципот што одбира според даден број колку соседи околу него да земе. На пример ако n\_neighbоrds параметарот е 3 моделот ги зема 3 најблиски соседи до податокот кој ни фали и зема најчестата класа од растојанијата на тие 3 точки околу него. Растојанието се пресметува со помош на евклидово растојание.

КНН може да се користи и кај регресија со тоа што ќе зеде просек во примерот погоре 3те точки околу него.

---------->Модел на логистичка регресија<------------------

Моделот на логистичка регресија се користи кога имаме класификација односно кога променливата која сакаме да ја предвидеме е категоричка. Логистичката регресија се користи за предвидување на веројатноста за бинарен исход (на пример, „default = Yes“ или „default = No“).

На пример, за податоците за кредитно задолжување, логистичката регресија ја моделира веројатноста за неплаќање на долговите. Таа користи логистичка функција која претвора линеарна комбинација на независни променливи во

вредност помеѓу 0 и 1 со пресметување на логаритам. Ако резултатот е поголем од 0.5 вообичаено се класифицира како 1 (позитивен случај), ако е помал од 0.5 се класифицира како 0 (негативен случај)

Сепак, недостаток на овој модел е тоа што е чувствителен на аутлаер.

---------->Модел на полиномијална регресија<------------------

Полиномијална регресија за разлика од линеарната или логистичката се користи за нелинеарни податоци каде кривата на податоци е нелинеарна. Наместо тоа што кај линеарната регресија се користеше линеарна комбинација на независните променливи , да ги најдеме B0, B1.. тука имаме и хиперпараметар(метаподаток) кој ни дава можност да одбереме параметар(полином) кој што ќе ни го оформи моделот. Пример хиперпараметар кај КНН претставува бројот на соседи,

тука ни е x на некој степен (X^n)

Се користи за поубаво фитување на моделот кога линеарната регресија дава голема грешка.

---------->Decisiton Tree модел<--------------

Decision Tree модел е модел каде се користи дрво на одлучување за да се предвиде врз основа на влезните податоци.

Се состои од 3 дела :

1) коренот на дрвото(root) - ги содржи сите податоци

2) внатрешни јазли(internal nodes) - се создаваат врз основа на прашања или услови засновани на некој feature на податоците

3) листовите - конечна одлука или предвидување

како функционира моделот?

1) поделба на податоците (splitting)

дрвото одлучува како ќе ги подели податоците врз основа на некоја нивна карактеристика(feature) , тука се користат разни метрики како gini index(за колку е чист одреден јазол), ентропија, mean squered error(за регресија, мери разлика меѓу предвидените вредности и вистинските вредности )

2) рекурзија - се повторува процесот се додека не се исполни услов за стоп(пр. достигната е максимална длабочина на дрво, нема веќе податоци за делење, класите се чисти)

3) одлука - секој лист претставува крајна одлука, за класификација најчесто е некоја класа, за регреија е некоја просечна вредност

- се користи и за категориски и за континуирани променливи

- не бара претходна нормализација на податоците

Џини индекс е мерка која се користи кај decision trees за да одреди колкава е хомогеноста(чистотата) во еден јазол , помага да се одбере кој feauture е најдобар за да се избере при поделба на податоците

- колку е помал џини индексот толку е почист јазолот(node);

чист јазол - сите податоци припаѓаат на една класа (џини = 0)

нечист јазол - податоците се мешавина од повеќе класи

-------------Bootstraping-----------------

Bootstraping е статистички метод кој се користи во машинското учење каде даден датасет се дели на помали делови односно примероци(samples).Се користи за проценка на распоределбата на примероците и за добивање на доверливи интервлали или статистички проценки.Овој метод е корисен кога не се има претходна претпоставка за обликот на распределбата или кога работиме со мали примероци.

------------Overfitting--------------------

Overfitting (прекумерно прилагодување) се јавува кога моделот премногу се прилагодува на специфичните податоци за тренинг, што доведува до ниска грешка на тренинг сетот, но лоша способност за предвидување на нови, невидени податоци. Ова може да се случи кога моделот е премногу сложен и вклучува многу карактеристики или високи степени на полиномијални функции, што резултира со "приспособување" на секој мал варијабилитет во податоците. Може да се избегне со користење на крос-валидација, намалување на сложеноста на моделите, како и со зголемување на тренинг множеството со што моделот ќе може да се тренира на повеќе различни случаи.

----------Cross Validation(крос-валидација)--------------

Cross Validation(крос-валидација) e техника на евалуација која се користи во машинското учење со тоа што ги дели податоците на повеќе подмножества со тоа што за да се тестира моделот на невидени податоци

Има повеќе видови на ваква валидација:KFold,RepeatedKFold...

Процесот на валидација кај КФолд:

1.ги дели податоците на к еднакви дела(folds)

2.моделот се обучува к-1 пати на различни делови и се тестира останатиот дел

3.се повторува к пати и резултатот на крај е просекот од сите тестови

Оваа техника се користи за избегнување на overfitting кај повеќе модели(линеарна регерсија,логистичка регресија,КНН...)

----------Bias vs Variance---------------

Bias претставува разликата помеѓу просечната прогноза на моделот и вистинската вредност која сакаме да ја предвидиме. Висок bias значи дека моделот има погрешни претпоставки за природата на податоците и може да направи едноставни погрешни проценки. Моделот не успева да ја улови сложеноста на податоците.

Variance се однесува на променливоста на моделот кога се користат различни тренинг сетови. Модел со висок variance има способност да се приспособува премногу на малите флуктуации во тренинг податоците (шум), што значи дека ќе направи многу добри прогнози на тренинг сетот, но лоши на нови податоци.

Високиот variance значи дека моделот е премногу сложен и многу чувствителен на тренинг податоците.

---------Regularization, Ridge vs Lasso------------

Регуларизација е техника со која се спречува overfitting кај моделите со додавање казна за комплексноста на моделот во функција на загуба.

Lasso се користи за да се елиминираат нерелевантни карактеристики.

Ridge е метод за регресија кој добро работи ако имаме висока корелација меѓу променливите. Се користи за да редуцира на тежина на сите карактеристики.

----------Bagging------------------

Bagging e техника на ансамблирање во машинското учење која комбинира повеќе слаби модели во 1 посилен продуктивен модел.Се користи за подобрување на стабилнсота и точноста на моделите. Се состои од:

1.boostraping - делење на датасетот на помали подмножества

2.aggregate - тренирање на сите подмножеста на повеќе модели и одбирање на најдоброто односно кога има класификација се користи мода или најчесто повторуван податок, а кога има регресија се одбира просекот од сите предвидени.

Во Bagging податоците кои не се користат односно кога тренираме моделот се тренира на 70/80% и другиот дел останува за неискористен. Тука тој се искористува за тестирање и тој дел се вика Out-Of-Bag податок.

---------Boosting------------------

Boosting e техника која се користи во машинското учење која комбинира повеке слаби ученици во 1 силен. Boosting за разлика од bagging-от во секој послендователен модел се тренира на грешките на претходните модели со цел на крај да дојдеме до модел кој има многу мала грешка.

Има 2 вида на boosting:

1.adaptive(како на пример AdaBoost)

-во адаптирачките, секој модел добива тежина пропорционално на неговата тежина и така се учи на грешки

2.gradient(LightBoost,XGBoost,CatBoost...)

-ги минимализира грешките со оптимизирање на ф-јата на загуба преку градиентен спуст

XGboost e еден од најдобрите со тоа што има регуларизација и спречува overfiting и вклучува техники како паралелно пресметување.

---------Random Forest------------

Random Forest е алгоритам кој комбинира повеќе дрва на одлучување за да креира помоќен и попрецизен модел. Кај него се користи случаен број на подмножества, но и случаен број на карактеристики(features).

Има повеќе параметри како што се:

-n\_estimators - од колку дрва е составен(поголем број значи поголема точност,но и поголемо време на извршување)

-criterion - мери колкав е квалитетот на поделбата

-max\_depth - максималната длабочина на дрвото

-min\_samples\_split - минимален број на примероци потребни за поделба

-min\_samples\_leaf - минимален број на примероци потребни да бидат лист

----------Типови на учења-----------

Има повеќе видови на учења кај машинското учење:

1.супервизирано каде се учи од податоци кои имаат лабели или очекувани излези, и моделот учи од познати примери од минатото

пр. линеарната регресија, логистилка регресија, дрва на одлука...

2.несупервизирано - учење без лабели односно моделот учи сам и треба да најде некој шаблон, структура за да научи

пр. кластеризација

3.учење со поттикнување(reinforcement learning)-учење со засилување односно со доделување награди и казни базирани на акции кои биле спроведени

-------------Типови на MISSINGNESS-------------

Има 3 видови на недостиг на вредности:

1.missing completely at random - кога недостигот на податоците е целосно случаен и не е поврзан со другите променливи или со самите вредности што недостасуваат

пр. кога на дадена анкета даден испитаник заборавил случајно да пополни полето за хоби

2.missing at random - кога недостигот на податоци може да е поврзан со некоја од другите променливи, но и со самите податоци кои недостигаат

пр. дадени испитаници со понизок приход почесто пропуштаат да одговараат на прашања поврзани со плата

3.missing not at random - кога недостигот е поврзан директно со самите вредности што недостасуваат

пр. испитаници со повисоки примања не сакаат да одговараат на прашања поврзани за приход, поради тоа што сметаат дека е тоа чувствителна тема