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