Можни прашања за ВНП есејски прво колк

Data science process - Предавање 1

1. **R-squared**

R-squared статистичка мерка која ја претставува пропорцијата на варијансата во зависните променливи од независните во регресискиот модел. Служи за оценување на нашиот модел поточно regression model. Ако нашиот модел е добар како средната вредност тогаш R^2 = 0. Ако пак нашиот модел е перфектен тогаш R^2 = 1. R може да биде и негативна вредност ако нашиот модел е полош од средната вредност тоа може да се случи ако ги оценуваме моделот во тест сетот.

2. **Внатрешни податоци (Internal data)**

Овие податоци може да се складираат во официјални складишта на податоци како на пример база на податоци, маркети на податоци, складишта на податоци и data pools

3.**Надворешни податоци(External data)**

Ако податоците не се достапни во внарешноста на организацијата побарај ги надвор.

4.**Cleansing data (Чистење на податоци)**

Овој процес е подпрецес на науката за податоци и е процес кој се фокусриа на отстранување на грешки во нашите податоци како грешки при внес, непотребни празни места, невозможи вредности, вредности кои недостасуваат итн..., така што нашите податоци стануваат вистинити и конзистентни.

5.**Confusion matrix**

Ни покажува колку случаеви се точно класифицирани и колку се неточно, спроредувајќи ги предвидувањата со вистинските вредности.

Прдедавање 2 understand data

1.**Што се податоци**

Податоци се единечно мерење на нешто со висина на тоа да е разбирлибо и на личноста што го врши мерењето и на личноста што го чита

2.**Начини да се соберат податоци online**

Има повеќе начини да се приберат податоци. Преку API користејќи веќеизградени функции направени од страна на компанија за да се пристапат нивните податоци и сервиси. Како google facebook, twitter. Rich site summary RSS сумаризи секојденвни ажурирања на онлајн содржини во стандарден формат. Бесплатни за читање ако самиот сајт дозволува на пример сајтови за вести. Web scraping користејќи софтвер и скрипти или пак рачно вадење на податоци кои се претставени

на некоја страна или пак се дел од HTML на страната. Се прави ако некои сајтови немаат свои API's за пристапување на податоци.

Streaming data

Податоци кои што се генерираат во реално време со одредена фреквенција на генерирање. Разликите кај streaming data, оценките во iknow не се streaming data, додека цените на берзата се streaming data.

3.**Dark data**

Некористена и прикриена дата која има потенцијал да креира нови вредности. Тоа се информации кои се организациите ги собираат, процесираат и ги стават во регуларни бизнис активности но генерално не се искористат.

4.**Lambda architecture**

Тоа е дата процесирачка архитектура дизајнирана да се справи со масовна големина на податоци. Таа се обидува да ја балансира латентноста, пропустноста и толеранцијата на грешки. Кај lambda architecture сите податоци кои излегуваат од stream се снимаат и се стават за batch обработка, наместо до сега stream обработка, и се обработуваат на крајот. Целта е паралелно да се обработуваат податоците, да се има прецизноста на сложените алгоритми, и да се има брзиот feedback на податоците кои се генерираат со голема фреквенција.

5.Digital twins

Тоа е дигитална репрезентација на ентитет од реалниот свет или систем. Овој концепт постоел како концепт во симулација, каде што за да се пресметува нешто за реалните објекти, се прави нивна симулација и може да се види што се случува со дадениот објект.

6.Data lake

Тоа е масивен репозиториум за податоци, базиран на нискобуџетна технологија, која ги подобрува справувањето, подобрувањето и истражувањето на необработените податоци. Се справува со голема количина на структурирани и неструктурирани податоци и повеќето податоци се непрепознатливи. главниот концепт е концептот на Schema on Read. Schema on Read значи, кога се става податоци не не интересира што има во податоците и какви се тие.

Популација

е цела група на единки која сакаме да ја моделираме. Од таа популација имаме **Sample** кој го избираме и кој ни е составен дел од нашите податоци. Поради тоа што не ја земаме целата популација, а е невозможно да се земе цела популација, имаме некој **Bias** предизвикан од земањето на одреден sample. Постојат **Selection bias** кој се појавува поради тоа што едни единки поверојатно ќе се изберат од други. **Volunteer/nonresponse bias** каде единките кои не се лесно достапни, не се репрезентирани.

Нормална дистрибуција

е резултат од природни процеси кои се случуваат во природата. Има интересни особини кои покажуваат дека варијабли кои се генерирани со било каков процес, кога зборуваме за нивна сума или средна вредност на резултатите од таков процес, резултира во нормална дистрибуција.

7.**Каде кои статистики да се употребат**

-<u>за континуирани(нормално дистрибуирани податоци)</u>

N, средна вредност, стандардна девијација, мин, мах

Хистограм, dot-plots, box-plots, scatter plots

-<u>за континуирани(искривени skewed) податоци</u>

N, медиана, квартили, мин, мах

Histogram, dot plots, box plots, scatter plots

-<u>за категориски податоци</u>

Фреквенција, проценти

One way tables, two-way

Bar charts

Хистограм – начин да се нацрта како еднодимензионален податок е дистрибуиран низ одредени вредности.
Pie chart – начин да се нацрта групирањето на податоците од некоја променлива.
Scatter plot – начин да се нацрта односот помеѓу дведимензионални атрибути на повеќедимензионални податоци.
Stacked area graph – начин да се нацрта како композициите(групите на податоци) се менуваат со тек на време.
Multiple histogram – начин да се нацрта како различни променливи се споредуваат.
Boxplot – упростена визуелизација за споредба на квантитивни променливи во рамките на групи. Се покажуваат опсег, квартили, медијана, итн. Точките кои се цртаат надвор од boxplot, претставуваат отстапувања и не треба да се земаат како податок за цртањето на графот.

<u>Data preparation - предавање 3</u>

8.**K-nearest neighbors (KNN)**

Поедноставен и логичен метод на предвидување, кој произведува многу конкурентни резултати. Работи така што треба да избереме број К кој е бројот на соседи кои ги гледаме. За секоја точка што сакаме да ја предвидиме, правиме сличност на податоците помеѓу точката за споредба и нејзините соседи. Најчесто мерката за сличност е дистанцата до соседите. Според оваа мерка, ги сортираме сите точки, колку се слични со точката за споредба. Ги избираме најсличните и според тоа што најмногу доминира во таа група одлучуваме дека класата на примерокот ќе биде класата

на тие што доминираат во групата. Се користи за multiclass класификација исто така може да се употреби за справување со вредности кои недостасуваат

1-((x1-x2)^2 + (xn-xn+1)^2)

9.Типови на податоци кои недостацуаат

Missing completely at random (MCAR) - веројатноста на недостасувањето во податок е иста за сите units.

Missing at random(MAR) - веројатноста на недостатоците зависи само од достапни информации.

Missig not at random (MNAR) - веројатноста на недостатоците зависи од информации кои не биле зачувани.

10.**Справување со категориски податоци**

Категориски податоци може да земаат само лимитирани или фиксни бројки на можни вредности.

Многу модели се алгебарски. Одосно внесените вредности треба да се нумерички. За да се користат овие модели, категориите треба да се трансформираат во броеви прво пред да се вклучат во алгоритмот за учење односно да се енкодираат со соодветни енкодери кој ги нуди python.

11.**Енкодирање со лабели**

Основен метод, кој ги заменува категориите со посакуваните бројќи користејќи речник кој содржи мапирачки бројќи за секоја категорија.

Друга метода е да се енкодират категориските вредности со техника наречена label encoding која дозволува да се конвертира секоја вредност во колоната во бројка. Нумеричките лабели се секогаш помеѓу 0 и 1.

12.**One-hot encoding**

Стратегијата е да се конвертира секоја категорија во нова колона на која се задава вредност 1 или 0 (True/False)

13.**Binary encoding**

Во оваа техника прво категориските се енкодираат како редовни, потоа тие интеџери се конвертираат во бинарен код, потоа бројќите од бинарниот стринг се поделени во оделни колони. Оваа ги енкодира податоците со помалку колони.

14.**Стандардизација**

Е препроцесирачки чекор во дата анализата и машинското учење кое вклучува трансформирање на одреден датасет на ист степен на податоци, односно да ги доведе различните вредности до степен каде тие ќе можат да бидат споредливи. Се трансформираат податоците така што средната вредност да е 0 додека пак стандарната девијација е 1.

15.**Нормализација**

Препроцесирачки чекор во анализата на податоци кој се стреми да ги трансформира податоците на заеднички степен без да има искривување или промени во опсегот на вредности. Тоа е типично скалирање помеѓу 0 и 1.

Machine learning intro - Предавање 4

16.**Што е машинско учење**

Тоа е гранка на вештачка интелигенција која се занимава со дизајн и составување на алгоритми кои овозможуваат компјутерите да развиваат однесувања базирани на емпириски податоци.

Поговорка -> погледни ги податоците, обиди се да направиш нешто. Добиваш точен резултат? Не ? Погледни ги пак податоците, направи нешто различно. Подобро? Да ? Направи го пак.

17.**Типови учења**

Supervised learning (супервизирано) - лабелите се дадени , тренирањето податоци вклучува посакуван излез, предвидување на иднината, учи од претходни примери во минатото за да ја предвидиш инината.

Unsupervised learning - лабелите не се дадени, не се вклучува саканиот излез, треба да се разбере минатото, се учи структурата на податоците.

Задачи на алгоритмите за машинско учење:

Класификација – има податоци кои треба да ги класифицираат.
Регресија – предвидување на одредени бројки
Кластеринг – има податоци и сакаме да ги групираме заедно
Аномалија – дали постојат елементи кои се чудни во подмнож.

18.**Машинско учење in a nutshell**

Репрезентација: Decision trees, sets of rules, instances, graphical models, neural networks, support vector machines, model ensambles, etc...

Евалуација: точност, прецизност, squared error, истоликост, веројатност, цена, margin, entropy, различност

Оптимизација: greedy search, gradient descent, linear programming

19**.Донесување на одлука**

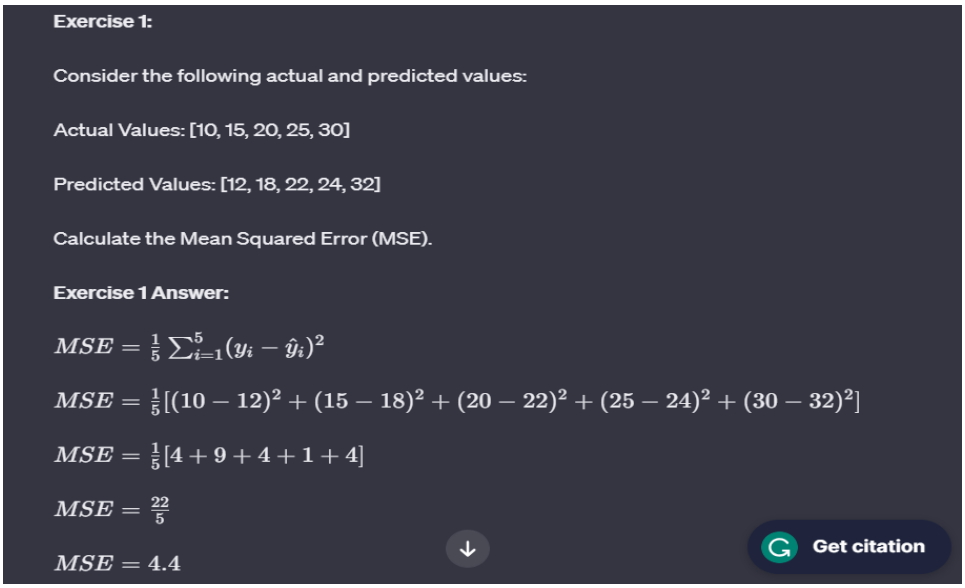
Машинското учење ги учи оптималните одлуки директно од податоците без има hardcore одлучување. И твојата одлука ќе биде многу по точна и ќе напредува во текот на времето додека има повеќе податоци.

20.**Предвидувања vs Проценка**

За некои проблеми важно ни е да го добиеме f. за некои проблеми ние не сакаме да се замараме за специфичната форма на f туку сакаме само да го предвидиме y што поблиску до набљудуваните вредности на y.

21.**Mean squared error**

Тоа е метрика за проценување на перформансот на регресиониот модел. Тој ја квантификува средната вредност помеѓу предвидените податоци и податоците од самиот датасет.



22.**Линеарни модели**

Тоа се класи на стаитстички и модели на машинско учење кои прават предвидување базирано на врската помеѓу внесените вредности и таргетираните вредности.

23.**<u>Bootstraping</u>**

Боотстрапингот е праксата на проценување на својства на еден проценувач со мерење на тие својства со земање на пример од одредени податоци кои ги надгледуваме. Bootstraping најчесто се користи да се предвиди дистрибуцијата на примероци на статистика кога аналитичките методи се многу комплексни

24.**Класификација**

Кога вредностите се повеќето категориски тогаш станува збор за класификациски проблем. Целта е да се проба да се класифицира секое набљудување во категорија исто така позанто како class cluster.

25.**logistic regression**

Таа се користи за бинарна класификација (0,1) каде променливата на исходот има само две категориски класи или може да се подели на две категориски класи.

27.Performance measurements

Precision со ова ги бараме true positives помеѓу сите примероци кои класификаторот ги лабелирал како позитивни

Recall ова е предвудувањето дека позитивни примери ќе бидат точно проценети од класификаторот.

Accuracy со ова велиме дека го мериме перформансот на моделот каде предметите се точно класифицирани.

Предавање 5 – Machine learning 2

28.**Polynomial reggression**

Тоа е тип на регресија каде врската помеѓу независната вредност и зависната вредност (таргетото) е моделирана како n-ти степен на полином. Односно дава можност на врските на вредностите да следат полиномна форма. Најдобро е да се користи кога имаме нелинеарни податоци односно кога хистограмот ќе покаже линија која не е права туку со искривеност.

29.**Overfitting**

Тоа е концепт од машинскот учење и статистичкото моделирање каде моделот ги учи податоците предобро, така што почнува да ги зима и шумовите и случајните промени кои не се толку битни но ќе влијаат лошо врз проценката. Тој модел се извршува добро на веќе тренираните податоци но нема да предвиди добро на податоци кои се нови.

30.**Избор на модел**

Изборот на модел е всушност примена на принципален метод за да се утврди покомплексниот модел. Бирање на субсет на предвидувачи, бирање на степенот на полиномниот модел и така натаму... многу важен мотив за да избереме точен модел е тоа што треба да избегнеме overfitting. Затоа секогаш треба да гледаме **бројот на параметри да биде помал од бројот на влезни податоци**.

31.**Крос валидација**

Тоа е техника користена во машинското учење за да се оцени перформансот и способноста за генерализација на моделот. Вклучува делење на датасетот на повеќе подсетови, тренирање на моделот на некој од тие сетови, и правење на предвидувања и тестирања врз останатите сетови. Целта на техинката е да се добие непристрасно оценување на перформансот на моделот.

32.**K-fold крос валидација**

Тоа е популарна техника за оценување на перформанс на моделот и таа евалуација да биде робустна. Оваа техника го дели датасетот на К субсетови кои се еднакви меѓусебе. Моделот е трениран К пати, секој пат користејќи К-1 субсетови за тренинг а останатите се користат за валидација. И овој процес се повторува К пати. Ова е корисно кога датасетот ни е ЛИМИТИРАН

33. Regularization

Е техника на машинско учење за справување со overfitting и за подобрување на способноста за генерализација на моделот и постојат **LASSO И RIDGE**

34.**LASSO**

Тоа е техника која ни помага да се справиме со overfitting со додавање на казнен член со објективната функција. LASSO го потикнува моделот да донесе ретки решенија со што тој ги доведува некој од коефициентите точно до нула. За да биде корисен тој претендира да исклучи некои податоци кои не се толку потребни од моделот.

35.**RIDG**

Исто како LASSO но тој главно се користи кога има многу корелациски податоци кои може да имаат некоја зависност меѓу себе.

LASSO VS RIDGE

Главна разлика помеѓу овие две техники е во тоа што LASSO користиме кога податоците немаат голема зависност со таргет податокот и тој претендира да ги донесе тренирачките податоци до нула додека пак RIDGE го користиме кога имаме некоја зависност помеѓу податоците и таргет множеството но тој не преферира да ги донесе стрикно до нула туку околу нула.

36.**Дрва на одлука**

Тие се широко користени во оваа област и можат да се употребат во класификација и во регресија. И овие модели се интерпретабилни односно предвидувањата се јасно воочливи и јасни зошто се направени од страна на моделот. Кај нив финалниот резултат се базира на серија на проценки на вредности и се разгранууваат тие одлуки како дрво.

37.**Gini index**

е друг пристап, кој наместо да се гледа грешката, се гледа колку се чисти групите кои сме ги поделиле. Не е битно колкава е грешката, туку дали групата е чиста, односно дали во една група има само една класа.

38.**Entropy**

Тоа е мерка за нарушување и нечистотијата на податоците во датасетот

39.Pruning

Тоа вклучува отргање на делво од дрвото кои не придонесуваат многу во одлучувачката моќ.