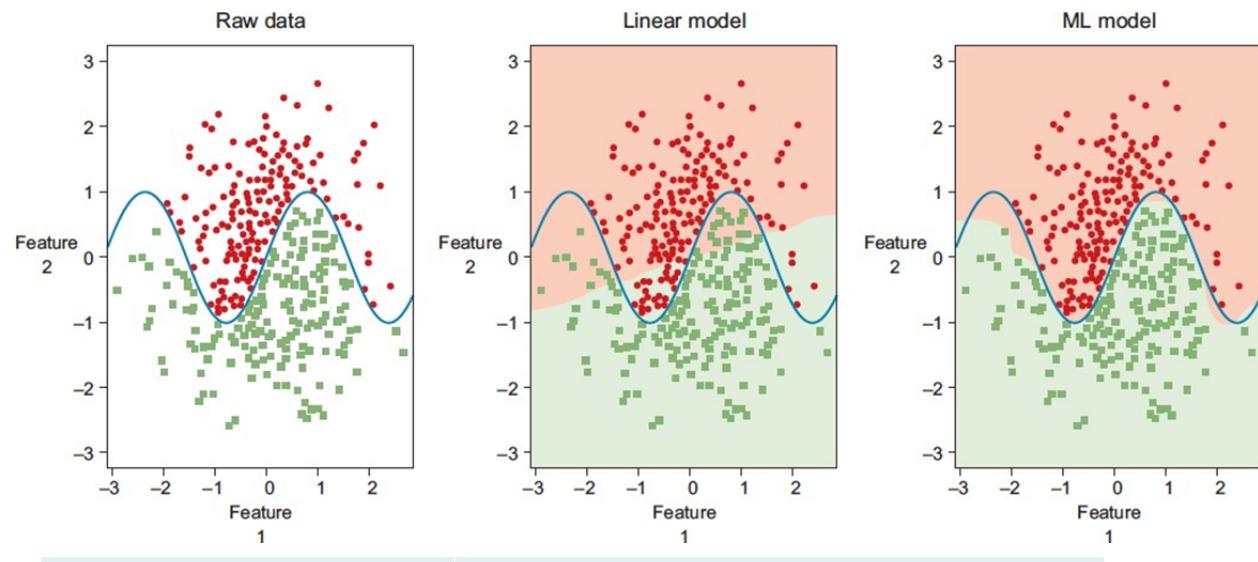


Introduction to Machine Learning

Representation -> How the learning occurs. How the learning process is structured (KNN, Linear Regression)

Evaluation -> How good our model is (using statistics)

Optimization -> choosing the best line



2. Ако податоците се означени кој ред во која класа припаѓа, тогаш станува збор за **Supervised learning**

Ако податоците не се означени со припадност по класи, тогаш станува збор за **unsupervised learning**.

Ако при обуката се добиваат само сигнали дали е нешто добро научено или не, тогаш станува збор за **reinforcement learning**.

Ако треба да се одреди припадноста на даден клиент во една од четирите **групи** на корисници, за каков вид на машинско учење станува збор?

- Класификација (Classification) - ТОЧНО

- Откривање на недостатоците (Anomaly Detection)
- Регресија (Regression)
- Учење со поттикнување (Reinforcement Learning)

Ако треба да се предвидува вредноста на температурата во даден пластеник во **текот на ноќта**, за каков вид на машинско учење станува збор?

- Откривање на недостатоци (Anomaly Detection)
- Учење со поттикнување (Reinforcement Learning)
- Класификација (Classification)
- Регресија (Regression)

Ако треба да се избере соодветна **акција** за одреден нов податок, за каков вид на машинско учење станува збор?

- Откривање на недостатоците (Anomaly Detection)
- Учење со поттикнување (Reinforcement Learning)
- Класификација (Classification)
- Регресија (Regression)

KNN

Advantages:

Only one parameter needs to be predicted

Disadvantage:

The only thing we know is the predicted output, but not how much each column affects the output

Only work with integers

Less than 10 different values per column

Linear Regression:

Advantage:

More descriptive

You can learn the dependencies

You can work with floats, time series, many different values...

Disadvantage:

More complex

R squared: Finding the best fit

It is used for regression

It describes how much the variance of dependent variables is explained by the independent variables

1 - unexplained variance/total variance

R²=0 ---> Average as a baseline model

Better than average + values

Worse - values

Above 0.5 ---> The model is good

Dictionary:

{'key1':value1, 'key2' : value2}

or dict constructor

```
dict([('sape', 4139), ('guido', 4127), ('jack', 4098)])
```

if inplace is false it returns a copy
false is default
axis = 0 -> row
axis = 1 -> column
If not given default is 0
so in this case it returns a copy and deletes the 2nd and 3rd row

Gradient descent: Used only for differentiable functions
Find the minimum point
Go right if slope is negative
Go left if slope is positive
It finds the global minimum, since the function has a lot of local minimums IRL

Confusion matrix:

	positive(predicted)	negative(predicted)	
positive(true samples)	TRUE POSITIVES (Ntp)	FALSE NEGATIVES (Nfn)	positives
negative(true sample)	FALSE POSITIVES (Nfp)	TRUE NEGATIVES (Ntn)	negatives

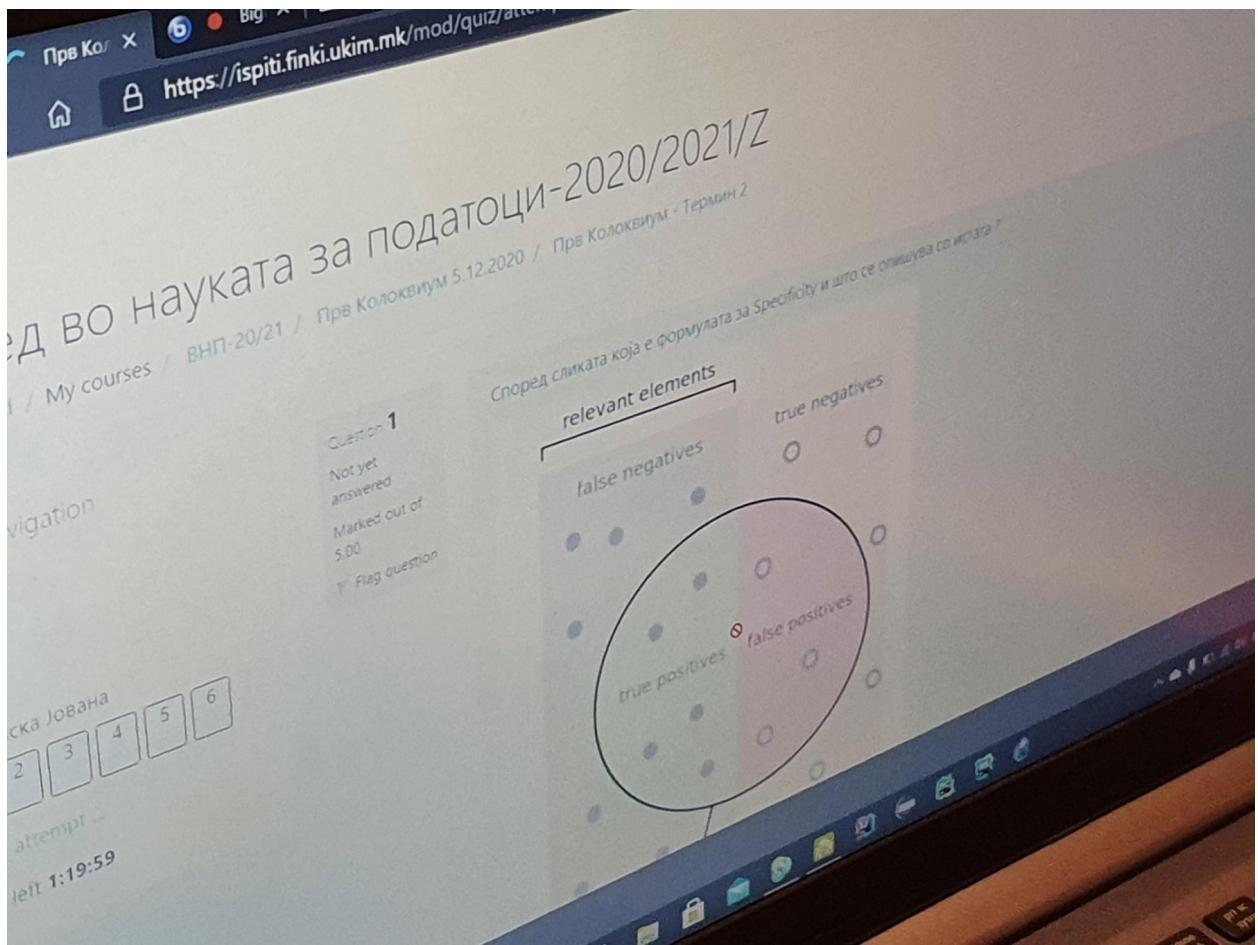
Precision: True positives between positives
 $Pr = Ntp / (Ntp + Nfp)$

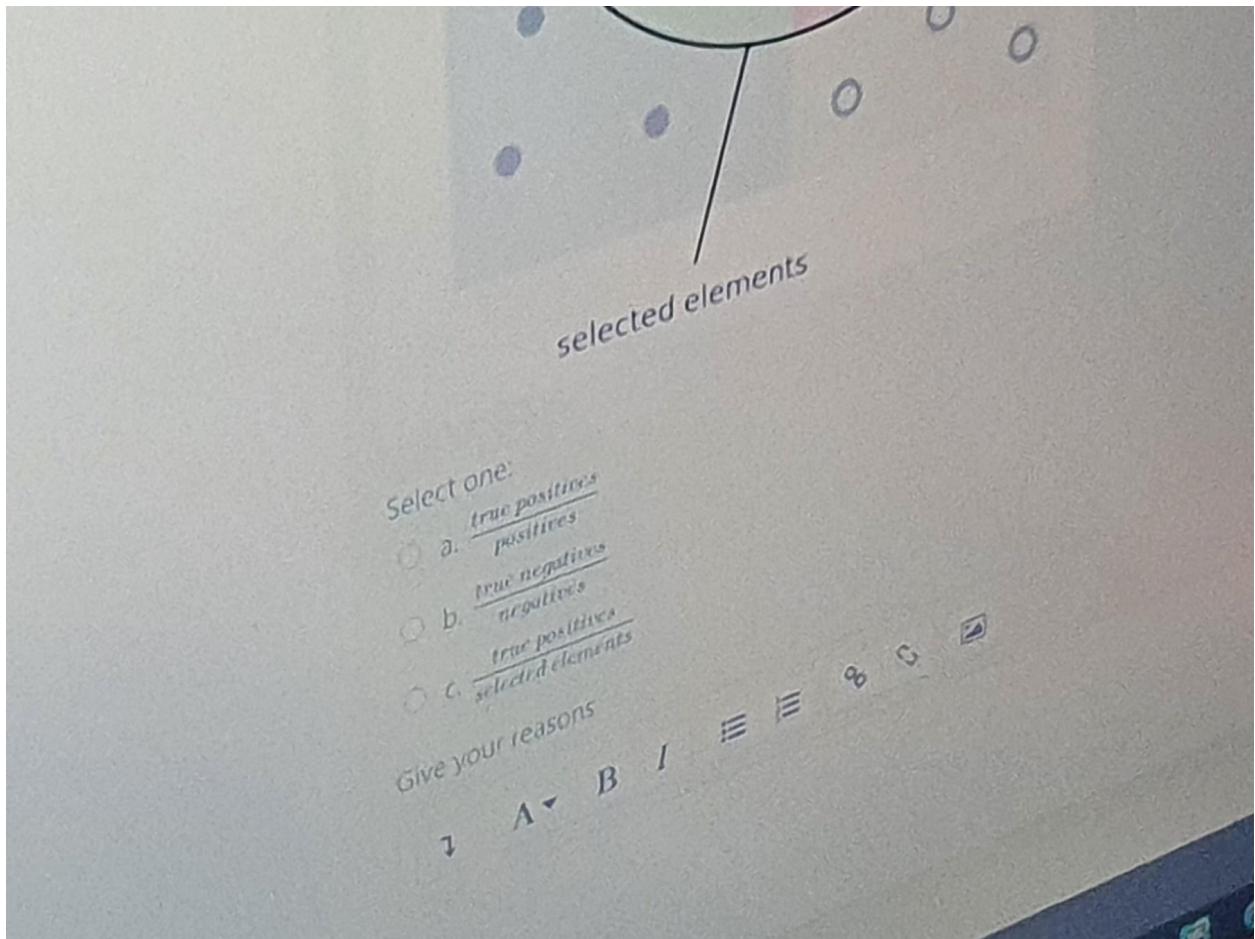
Recall on positives (Sensitivity): Correct guessed positives within one guess
 $Re = Ntp / (Ntp + Nfn)$

Recall on negatives (Specificity): Correct guessed negatives
 $Sp = Ntn / (Nfp + Ntn)$

Accuracy : How many correct we have all together
 $Acc = (Ntp + Ntn) / \text{all 4}$

F1 score → how good our model is
 $2 / (1/\text{recall} + 1/\text{precision})$





Data preparation for ML

KNN algorithm:

Distance as similarity measures.

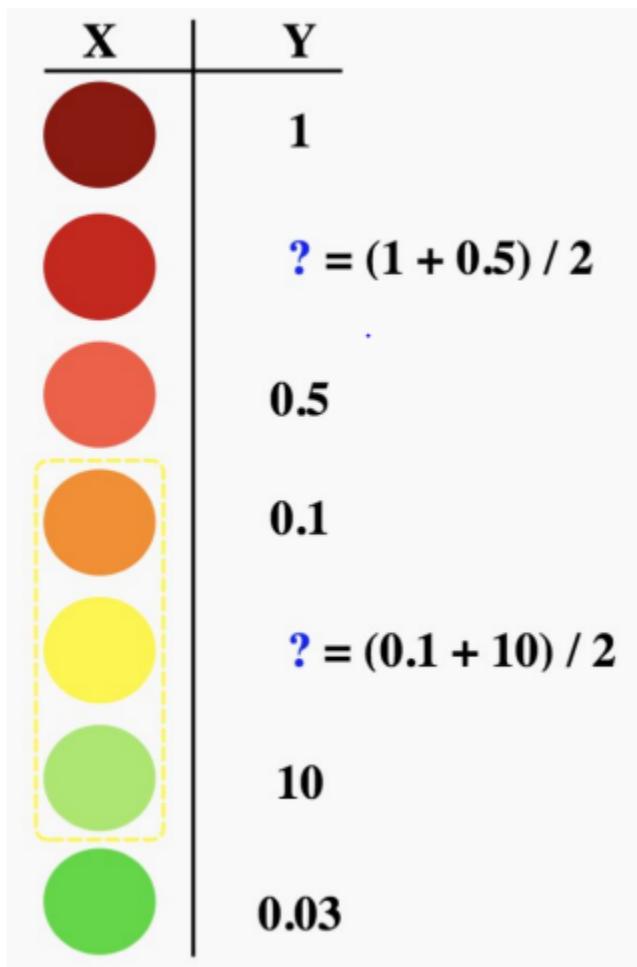
13. Нека се дадени следниве променливи X и Y

X	Y
1	3
2	3
Null	4
4	3
1	2

Ако се користи k-NN метод со k=2 за пополнување на податоците кои недостасуваат, со која вредност ќе се замени Null за променливата X.

- a. Неможе да се определи
- b. 3
- c. 4
- d. 2

imputation through modeling



Даден е модел на KNN класификација за предвидување дали куката ќе се продаде или не - 0 или 1 соодветно (class колоната), ако како влезни променливи се следниве:

1. местоположба на куката
2. број на спратови
3. површината на земјиштето

Што дефинира n_neighbors=2 за дадениот код:
classifier = KNeighborsClassifier(n_neighbors=2)

- a. Само за последните два примероци од dataset-от ќе се пресмета растојанието до примерокот чија класа ја предвидуваме
- b. За предвидување на class колоната на нов примерок ќе бидат земени двата најблиски примероци на кои е трениран моделот
- c. Само за првите два примероци од dataset-от ќе се пресмета растојанието до примерокот чија класа ја предвидуваме
- d. Само за соседните два примероци на примерокот чија класа ја предвидуваме ќе се пресмета растојанието

[Clear my choice](#)

Marked out of

Correct

Mark 15.00 out
of 15.00

Flag question

Edit
question

За дадениот датасет во табелата потребно е со помош на KNN класификација со $k = 3$, да се предвиди класа ќе припаѓа новиот тест примерок со **Id 5**

Id	Debt	Annual Income	Defaulted Borrower
1	1	3	No
2	0	4	No
3	2	2	Yes
4	3	5	Yes
5	4	2	?

Во следната табела пополнете го растојанието до примерокот со Id 5 пресметано со помош на Euclidean distance. (да се заокружи на 2 децимали)

Id	Defaulted Borrower	Distance
1	No	3.16 ✓
2	No	4.47 ✓
3	Yes	2.1 ✓
4	Yes	3.16 ✓

Примерокот со Id 5 ќе биде класифицран како

Question 2

Not yet
answered

Marked out of
20.00

Flag question

Time left 0:55:09

За дадениот датасет во табелата потребно е со помош на KNN класификација со $k = 3$, да се предвиди во која класа ќе припаѓа новиот тест примерок со **Id 5**

Id	Debt	Annual Income	Defaulted
1	6	3	No
2	5	4	No
3	4	2	Yes
4	3	3	Yes
5	2	2	?

Во следната табела пополнете го растојанието до примерокот со Id 5 пресметано со помош на Euclidean distance. (резултатите да се заокружат на 2 децимали)

Id	Defaulted	Distance
1	No	
2	No	
3	Yes	
4	Yes	

Примерокот со Id 5 ќе биде класифицран како

$$\text{First} \rightarrow \text{Normalization } x_{\text{new}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

Second → Calculate distance with euclidean formula for each point

$$\sqrt{(y_1 - y_2)^2 + (x_1 - x_2)^2}$$

Third → Sort in ascending order and choose from K first points the class

If k is even, we choose the closest ones → If same distance, it doesn't matter

A lot of Machine Learning techniques can't handle missing values!!!

Types of missigness:

- 1.Missing completely at random: Doesn't affect end result Od nisto nikogas ne zavisi
- 2.Missing at random: Od nesto zavisi nekogas
- 3.Missing not at random: Missing data means something.

Vo koja grupa pripagja odreduvame spored toa kolku vlijae toj faktor na target prashanjeto

Dealing with outliers:

- Removal
- Transformations
- Truncation

Encoding Categorical Data

Label encoding → replacing categories with numbers using dictionary

One-Hot encoding → convert each category to a new column and assign 0 or 1

Binary encoding → first convert to integer → then integer to binary number

Proxy encoding → find a value that has a meaning

Text vectorizer:

- Count vector
- You make a vocabulary
- Image→ already a vector

Data leak

Data schema denormalization

Normalization: $x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$ Range [0,1]

Standardization: $x_{new} = \frac{x - \mu}{\sigma}$ → $\mu=0$ and unit variance

Reduce the number of variables → Some unimportant variable can mess up all the statistics

Curse of dimension ! Data becomes sparse

If we have more data, we need more samples to make the model

1axis 100 samples → 100% full

2axes 100 samples → 1% full 10 000 100 → 1% of 10 000
3axes 100 samples → 0.01% full

Outliers: → easiest to see on box plot but also distributions

- Present false information (data errors, etc.) • Ruin predictions (increase error-proneness)
- Create new questions (new clusters, etc.) • Offer insights (anomalies, examples, etc.) • Announce issues (fraud, etc.)

Get and understand the data

Three Vs:

Volume -
Variety -
Velocity - At what speed is new data generated?

Ways to generate online data:

API: prebuilt by companies to access their services. Pay to use. Google Maps, Facebook, Twitter

RSS(Rich Site Summary): free to read, sites and blogs, XML format
Web scraping: What is contained in the HTML

Dark data: not easily available

Tabular data: CSV

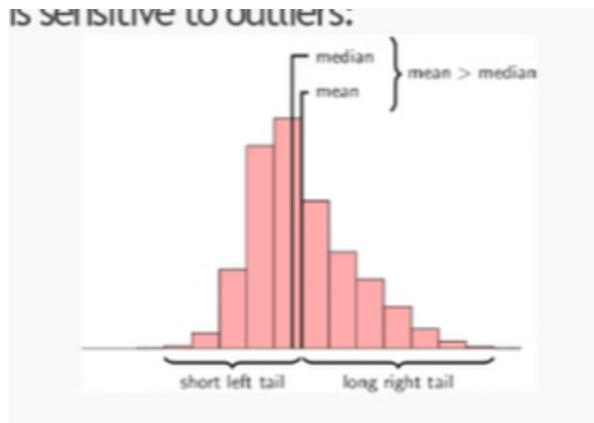
Structured data: json.dictionary,xml

Semi-structured: different key-value

Types of data:

Quantitative
Categorical

Median stays same, mean changes → when there are outliers



Right skewed: mean > median (outliers are at right)

Normal \rightarrow mean == median

Left \rightarrow mean < median

Variance: How much on average the sample values deviate from the mean

For continuous (Normally distribution):

Mean, std,min,max

Histogram, dot plots, box plots, scatter plots

For categorical:

frequency , pertentages

Bar charts

Anscombe data: Summary data doesn't always tell the story of the data. Outliers can mess up the statistics. Identical statistics, different distributions because of outliers.

Visualization:

Identify hidden patterns

Formulate/test hypothesis

Communicate results

Determine the next step

Types of visualization:

Distribution: how the variables are distributed over a range

Relationship: how the variables relate

Composition: subgroups

Comparison: how trends in multiple variables compare

Histogram: One dimensional data (Trends are sensitive to number of bins)

Pie chart: categorical variables

Scatter plots: relationship within two different attributes of multidimensional data

Axes numerical, but categorical can be added with color, size, shape of dot.

Stacked area: trend over time

Multiple histogram

Boxplot: data distribution of continuous variables, quantiles are shown, best for outliers.

Кои од наведените дескриптивни статистики е најдобро да се изберат ако податочното множество се состои од континуирани податоци и при тоа е потребно да се прикаже варијација на истите? - Ранг , Интер-квартална разлика, Стандардна Варијација

Data Science Process

1. Setting the research goal

2. Retrieving data → Internal (databases, data lakes, data marts, data warehouses)
External data

3. Data Preparation

Cleaning (Missing values, Errors (Inconsistencies (F or Female), Interpretation error (don't make sense)),
Transforming (Aggregation WorkWeek vs Week)

Combining

4. Data exploration

Graphical (simple graphs, combined graphs, link and brush)
Nongraphical (tabulation, clustering, modeling techniques)

5. Data modeling

Select modeling technique
Execute

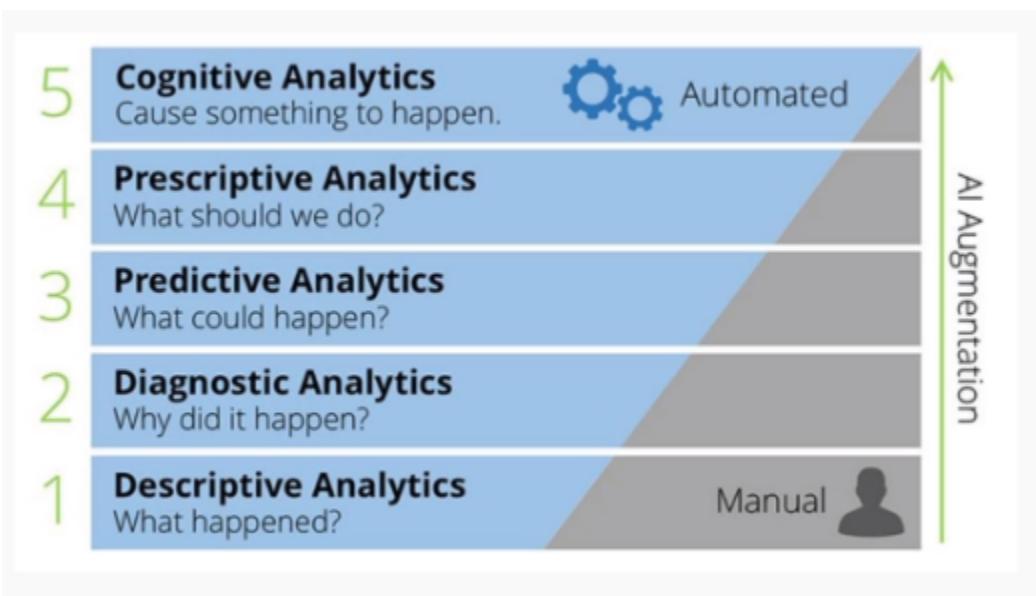
Compare with other models

6. Presentation and automation

Introduction Lecture

Machine Learning	Data Science
Develop new (individual) models	Explore many models, build and tune hybrids
Prove mathematical properties of models	Understand empirical properties of models
Improve/validate on a few, relatively clean, small datasets	Develop/use tools that can handle massive datasets
Publish a paper	Take action!

	Databases	Data Science
Data Value	"Precious"	"Cheap"
Data Volume	Modest	Massive
Examples	Bank records, Personnel records, Census, Medical records	Online clicks, GPS logs, Tweets, Building sensor readings
Priorities	Consistency, Error recovery, Auditability	Speed, Availability, Query richness
Structured	Strongly (Schema)	Weakly or none (Text)
Properties	Transactions, ACID*	CAP* theorem (2/3), eventual consistency
Realizations	SQL	NoSQL: Riak, Memcached, Apache River, MongoDB, CouchDB, Hbase, Cassandra,...



Machine Learning 2

□ Polynomial Regression

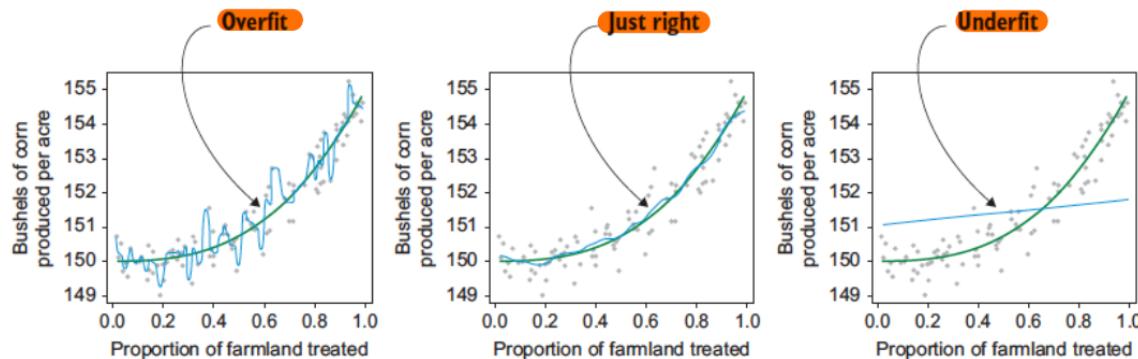
Simplest non linear model

Linear regression: $\beta_0 + \beta_1 x$

If polynomial passes in 6 points, then it has till $\beta_5 x^5 \rightarrow \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4 + \beta_5 x^5$

Prefer the simplest hypothesis that is consistent with the data to prevent overfitting.

Divide the data into train and test to see if your model overfits.



Training set error is usually lower than the test set error.

□ Cross validation:

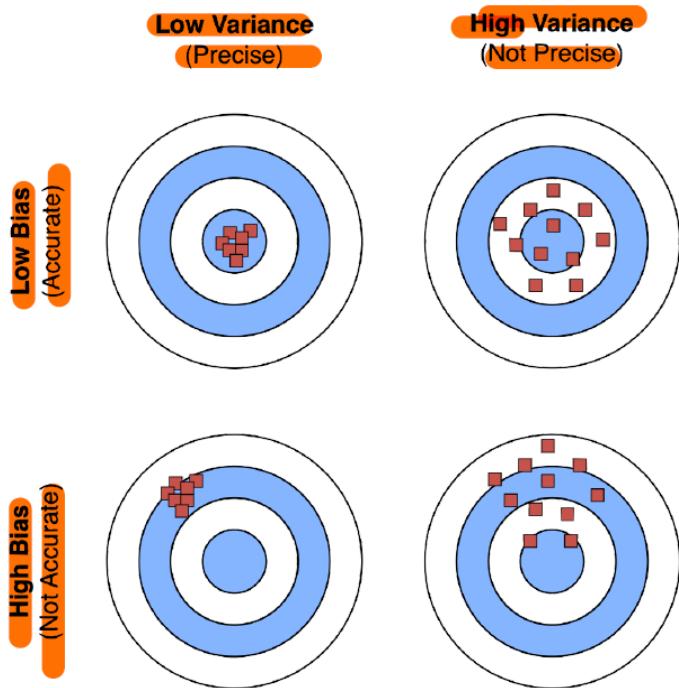
Randomly split the training set into training and validation multiple times but randomly creating these sets can create the scenario where important features of the data never appear in our random draws.

Therefore, we use K-fold cross validation.

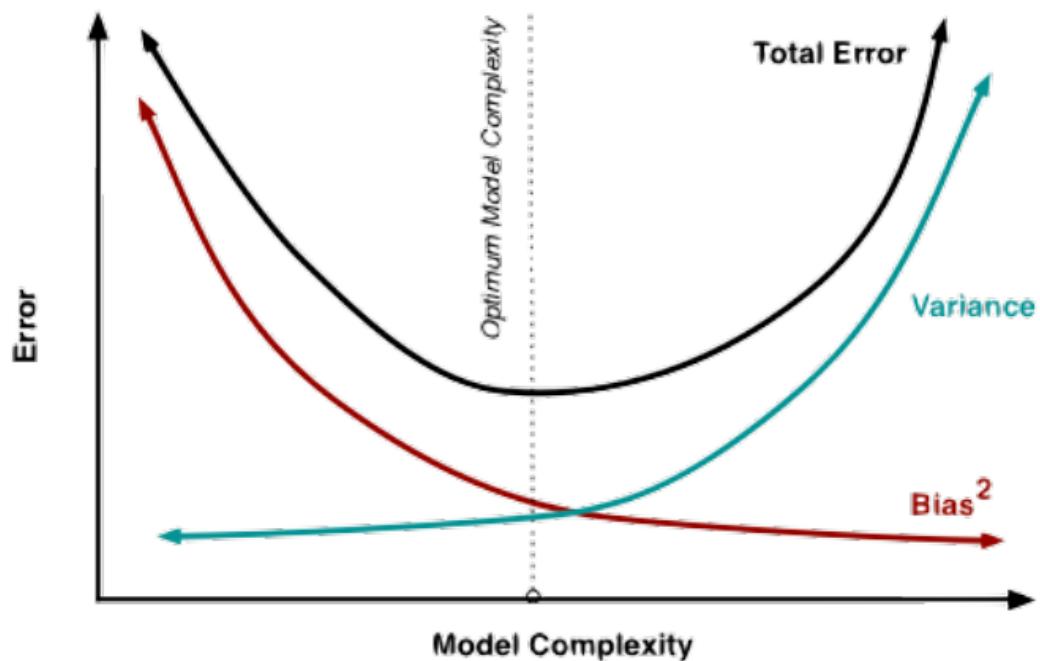
The data set is divided into k subsets, and the holdout method is repeated k times. Each time, one of the k subsets is used as the test set and the other k-1 subsets are put together to form a training set. Then the average error across all k trials is computed.

The variance of the resulting estimate is reduced as k is increased.

Cross validation is a must for a small dataset.



Both variance and bias introduce error to the model. When we have low bias and low variance the model is the best.



Polynomial regression has coefficients in billions.

Regularization: LASSO and Ridge:

The idea of regularization revolves around modifying the loss function L , we add a regularization item that penalizes some specific property.

$$L_{reg}(\beta) = L(\beta) + \lambda R(\beta) \text{ where } \lambda \text{ gives weight for the regularization term}$$

If it's close to 0, it is a regular MLE, if it is high MSE will be close to zero. We use cross validation.

LASSO: We chose MLE as our loss function and as R we chose the sum of the absolute value of the coefficients

RIDGE: We choose a regularization term that penalizes the squares of the parameter magnitudes.

Decision Trees

Logistic regression for building classification boundaries works best when:

- the classes are well-separated in the feature space
- have a nice geometry to the classification boundary

A decision tree model is one in which the final outcome of the model is based on a series of comparisons of the values of predictors against threshold values.

- the internal nodes of the tree represent attribute testing.
- branching in the next level is determined by attribute value (yes/no).
- terminal leaf nodes represent class assignments.

Given a training set, learning a decision tree model for binary classification means:

- producing an optimal partition of the feature space with axis aligned linear boundaries
- each region is predicted to have a class label based on the largest class of the training points in that region

ID3 Algorithm – works with discrete values only, if all the values are from the same class it stops, otherwise the division continues.

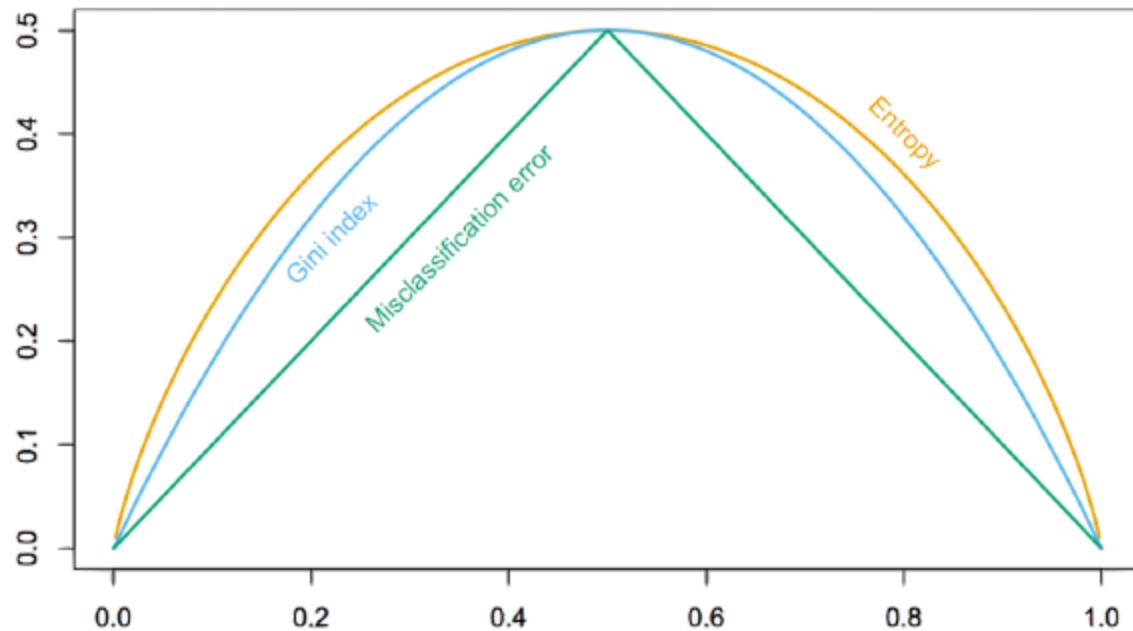
Classification error:

$$\text{Error}(i|j, t_j) = 1 - \max_k p(k|R_i)$$

where $p(k|R_i)$ is the proportion of training points in R_i that are labeled class k .

Gini index: smaller gini is better

Entropy:



Gini index and entropy punish more.

Stopping conditions: -min_samples_leaf

-max_leaf_nodes

-min_impurity_decrease when metric is gini index or entropy

-gain in purity

Alternative to stopping is pruning

$$C(T) = \text{Error}(T) + \alpha|T|$$

where T is a decision (sub) tree, $|T|$ is the number of leaves in the tree and α is the parameter for penalizing model complexity

$$\alpha = 0.2$$

Tree	Error	Num Leaves	Total (complexity score)
T	0.32	8	0.32+0.2*8=1.92
Tsmall	0.33	7	0.33+0.2*7=1.73

Even though it looks like the smaller tree has bigger error, overall it's better. (cost complexity pruning)

58. Потребно е да се креира модел за предвидување на дневната температура, ако како влезните податоци ги зема:

Question 3
Not yet answered
Marked out of 1.00
Flag question

Time left

Потребно е да се креира модел за предвидување на дневната температура, ако како влезни податоци ги зема:

1. влажноста на воздухот
2. температурата од претходниот ден
3. дали врне или не во моментот

Кој од дадените модели е точен?

a. `model = DecisionTreeClassifier()`

b. `model = DecisionTreeRegressor()`

Под а

Question 5
Not yet answered
Marked out of 1.00
Flag question

Time left 0:18:47

Кај Наивните Баесови класификатори, за атрибуутите A_i за дадена класа C може да претпоставиме:

a. Условна зависност меѓу атрибуутите, за таа класа

b. $P(A_1, A_2, \dots, A_n | C) = P(A_1 | C)P(A_2 | C) \dots P(A_n | C)$

c. Условна независност меѓу атрибуутите, за таа класа

d. $P(A_1, A_2, \dots, A_n | C) = P(A_1 | C) + P(A_2 | C) + \dots + P(A_n | C)$

Previous page

Next page

Најдобрите податоци

Question 12
Not yet answered
Marked out of 1.00
Flag question

Time left 0:17:07

Кога дистрибуцијата на податоците е како на сликата, какви се наклонетоста (bias) и варијансата кај овие податоци?



Select one:

a. Неможе да се заклучи од дадениот график

b. Голема наклонетост и мала варијанса

c. Мала наклонетост и голема варијанса

d. Голема наклонетост и голема варијанса

63. На кој начин би требало да го избереме параметартот ламбда во техниката за Регуларизација ?

Question 3
Answer saved
Marked out of 15.00
 Flag question

На кој начин би требало да го избереме параметартот ламбда во техниката за Регуларизација?

За да се избере ламбда, потребно е да се користи крос валидација бидејќи доколку се земе преголема или премала ($\lambda > 0$) вредност, би добивале отстапувања од стандардната грешка. За да се избегне тоа, се користи формула, така што се избира ламбда, а потоа се пресметува коефициент за корелација R^2 . Се зема во предвид најголемиот пресметан коефициент на корелација кој одговара и тоа ламбда што се користело во таа формула, тоа се користи како крајно.

64. Што може да се заклучи кога податочното множество има висока варијанса и висок баес?

Question 1
Answer saved
Marked out of 5.00
 Flag question

Што може да се заклучи кога податочното множество има висока варијанса и висок баес ?

Select one:

- a. Претпоставките за таргт функцијата нема да бидат со голема точност, воедно ќе се зголеми прецизноста поради големите промени кај податоците.
- b. Претпоставките за таргт функцијата ќе бидат со голема точност, воедно ќе се намали прецизноста поради големите промени кај податоците.
- c. Претпоставките за таргт функцијата ќе бидат со голема точност, воедно ќе се зголеми прецизноста поради големите промени кај податоците.
- d. Претпоставките за таргт функцијата нема да бидат со голема точност, воедно ќе се намали прецизноста поради големите промени кај податоците.

[Clear my choice](#)

Give your reasons



65. Кои се предности а кои недостатоци на Convolutional Neural Network ?

Question 2
Incomplete
Mark: 10.00 out of 10.00
 Flag question
 Edit question

Кои се предностите а кои недостатоците на Convolutional Neural Network

Во конволуциски мрежи предност е што секој слој е поврзан со константен број на единици од слојот пред него. Исто така единиците ги споделуваат тежините за овие поврзаности, односно конекции. Друга предност на овие мрежи е што нивните скриени единици се поврзани со одреден број на единици од слојот кој што следи после нив, со што се намалува бројот на параметри. Конволуциските мрежи како влезови прифаќаат матрици. Овие матрици се користат за процесирање на слики и текст. Недостаток или негативна страна на овој тип на мрежи е следниов: за да ние изградиме една мрежа за учење на различни врски меѓу зборови, ќе ни требаат многу филтери со различни висини. Друг недостаток е што конволуциското јадро не е доволно за да се детектираат различни видови на features, кои ќе бидат користени во класификацијата.

The screenshot shows a digital assessment interface. At the top, there's a header with 'Photos - Screenshot (585).png 188%' and various file management icons. Below the header, a table displays a single question:

5	16/02/21, 10:36	Manually graded 10 with comment:	Complete 10.00
---	-----------------	----------------------------------	----------------

The question details are as follows:

Question 4
Complete
Mark 9.00 out of 10.00
Flag question
Edit question

The question text asks: "Кои предности ги имаат Transformers моделите во однос на RNN Sequential моделите." The answer provided is: "Transformers моделите може да го решат проблемот на sequence transduction или neural machine translation, односно секој влез кој што е како секвенца, да го трансформираат во точна излезна секвенца. Пример speech recognition, text-to-speech transformation... На тој начин transformers моделите го решаваат проблемот на паралелизам, користејќи Convolution Neural Network(CNN) заедно со Attention models."

Below the question area, there's a comment section with the placeholder "Comment:" and the instruction "Make comment or override mark".

67. Објаснете го принципот на кој работат autoencoder-ите.

The screenshot shows a digital assessment interface. A question is displayed with the following details:

Question 3
Complete
Mark 9.00 out of 10.00
Flag question
Edit question

The question text asks: "Објаснете го принципот на кој работат autoencoder-ите." The answer provided is: "Аутоенкодерите работат на тој начин што автоматски пронаоѓаат најдобар начин за енкодирање на влезот, така што резултатот при декодирање да биде најблизок со оригиналниот влез. Енкодерите се претставници на supervised learning. Начинот на кој што тие работат е што најпрво правиме train на две мрежи преку минимизирање на loss function (cross entropy loss). Аутоенкодерите се користат за енкодирање на слики, видеа и аудио звуци, со цел да се намали нивната големина за да можат полесно да се испратат низ телекомуникациските мрежи. При кодирањето се губат податоци, и затоа постојат lossy и lossless енкодирање. Што поголема е разликата помеѓу оригиналниот податок и декодираниот податок, тогаш станува збор за lossy енкодирање. Овој тип на енкодирање најчесто се користи кај енкодирањето на слики, видеа, а поретко кај звуци и текстови. Бидејќи ако премногу ја намалиме големината на еден текст тогаш ќе ја изгубиме смислата на самиот текст, ќе изгубиме важни податоци."

1. Да се објасни за Gradient boosting и зошто е тој ефикасен?

2. Да се објасни за концептот positonal encoding кај трансформерите.

79. 3. Објасни ги разликите помеѓу GINI и Entropy метриките.

Question 1

Correct

Mark 5.00 out of
5.00

Flag question

Edit
question

Што се случува кога ентропијата за одредена колона кај даден датасет тежне кон нула?

Select one:

- a. Податоците се добро поделени.
- b. Податоците се несредени т.е немаат добра поделба.

Give your reasons

Кога ентропијата тежне кон нула, податоците се полесни за претпоставка и имаат поистакнати "пикови".

Your answer is correct.

The correct answer is: Податоците се добро поделени.

Comment:

Make comment or override mark

Прашање 2.2

- При градење на дрво за одлучување се избира атрибутот со
 - A) најголем број различни вредности
 - B) најмал број различни вредности
 - C) најголема информациска добивка
 - D) најмала информациска добивка

Прашање 2.3

- Дрвата за одлучување:
 - A) се лесни за разбирање
 - B) претставуваат black-box
 - C) се тешки за визуелизација
 - D) не се многу прецизни

Question 6
Not yet answered
Marked out of 1.00
Flag question

Ako се подели податочното множество на повеќе делови и потоа се остава едно за тестирање, а другите се користат за обука, за која техника на машинското учење се користи за избор?

Select one:

a. Ласо регуларизација (LASSO Regularization)
 b. Вертична валидација (Cross Validation)
 c. Регуларизација по ортог (Ridge Regularization)
 d. Бигропнија

Previous page Next page

Announcements

Jump to... Cetaja 1 ←

5 Time left of

Кои од наведените можат да се користат како критериуми за прекин на понатамошното деление на јазлите кај дрвата за одлучување (Stopping Conditions)?

Select one or more:

a. Ако бројот на примероци што припаѓаат на дадена класа го надмине дозволениот број.
 b. Ако сите примероци во јазелот припаѓаат на истата класа
 c. Ако бројот на циклуси надмине даден праг
 d. Ако бројот на примероци во под-јазлите спадне под даден праг (`min_samples_leaf`)
 e. Ако бројот на јазлите во дрвото надмине даден праг.

Question 11
Not yet answered
Marked out of 1.00
Flag question

Кој случајните шуми, кои хипер-параметри можат да се нагодуваат

- a. Бројот на атрибути кои се избираат случајно при секоја подделба.
 b. Вкупниот број на дрва во ансамблот.
 c. Сите наведени.
 d. Претходните веројатности (apriori) за дадените ознаки на некоја класа.

Previous page

Question 9
Not yet answered
Marked out of 1.00
[Flag question](#)

Кога дистрибуцијата на податоците е како на сликата, какви се наклонетоста (bias) и варијансата кај овие податоци?



Select one:

a. Мала наклонетост и мала варијанса

[Previous page](#) [Next page](#)

[Termin 2 - пријател](#) [BigBlueButton - Секција 7](#) [Course Вовед во науката за податоци](#) [ukim.mk](#) [https://www.ukim.mk/pluginfile.php?fileid=247089&contextid=12087&page=9](#)

Андонов Лазар

RINKI испити

Вовед во науката за податоци-2021/2022/Z

Dashboard / My courses / ВНИ121/22 / При Колоквум 5.12.2020 / Термин 2 - пријат

Quiz navigation

1	2	3	4	5	6
7	8	9	10	11	12

Question 10
Not yet answered
Marked out of 1.00
[Flag question](#)

За дадениот модел:
model = DecisionTreeClassifier()
Кој параметар треба дополнително да се додаде како аргумент во заградите за да се користи ентропија како метрика за поделба на дрвото на одлука.

a. metric = "entropy"
 b. splitter = "entropy"
 c. criterion="entropy"

Time left 0:17:55

[Previous page](#) [Jump to...](#) [Next page](#)

Да се определи колку изнесува Чини индексот за првата редица (R1) од дадената табела каде колоните ја означуваат класата, а редиците регионот.

Class 1 Class 2

R1	2	5
R2	6	4

Select one:

- a. 0.5
- b. 0.168
- c. 0.282
- d. 0.45

Аудито | X | □

quiz/attempt.php?attempt=102471&cmid=5803&page=1

За дадениот датасет во табелата потребно е со помош на Classification Error да се одреди следната колона по која ќе се прави раздружене на дрвото на одлука (Defaulted Borrower е target колона т.е по неа се врши класификацијата!)

Id	Marital Status	Annual Income	Defaulted Borrower
1	Single	High	No
2	Married	Low	No
3	Divorced	Low	Yes
4	Married	Medium	Yes
5	Divorced	High	No
6	Single	Low	No
7	Divorced	Medium	Yes
8	Divorced	High	

(Закружи ги децималните места ако се повеќе на втората децимала)

Classification Error за колоната Marital Status изнесува:

Classification Error за колоната Annual Income изнесува:

За следна поделба на дрвото на одлука се избира колоната

question 2
not yet
answered
Marked out of 5.00
Flag question

За даденото податочно множество во табелата потребно е, со помош на индексот Џини, да се одреди следната колона по која ќе се врши разграничување на дрвото на одлука. (Вид_на_овошје е целна колона т.е според неа се врши класификацијата)

Овошје	Слаткост	Киселост	Вид_на_овошје
Лимон	Многу ниска	Висока	Кисело
Цитрон	Многу ниска	Висока	Кисело
Портокал	Ниска	Висока	Кисело
Малина	Ниска	Средна	Кисело
Чреша	Ниска	Средна	Благо
Банана	Висока	Ниска	Благо
Лубеница	Висока	Ниска	Благо

(Засокржете ги децималните места, ако се повеќе, на втората децимала)

Просечниот индекс Џини (со тежински фактор) за колоната Слаткост изнесува:

Просечниот индекс Џини (со тежински фактор) за колоната Киселост изнесува:

За следна поделба на дрвото на одлука се избира колоната

Give your reasons

1 A B I %

Question 4
Partially correct
Mark 7.00 out of 15.00
Flag question

За дадениот датасет во табелата потребно е со помош на Classification Error да се одреди следната колона по која ќе се врши разграничување на дрвото на одлука. (Fruit_type е таргет колона т.е по неа се врши класификацијата)

Fruit	Sweetness	Sourness	Fruit_type
Lemon	Extremely Low	High	Sour
Grapfruite	Low	Medium	Sour
Orange	Low	Medium	Sour
Raspberry	Medium	Medium	Sour
Cherry	Medium	Medium	Sweet
Banana	High	Low	Sweet
Watermelon	High	Low	Sweet
Mandarin	Extremely Low	Medium	None

(Засокржи ги децималните места ако се повеќе на втората децимала)

Classification Error за колоната Sweetness изнесува: 0.33

Classification Error за колоната Sourness изнесува: 0.37

За следна поделба на дрвото на одлука се избира колоната Sourness

Give your reasons

Задачата имајќи во видувањето на датасетот јасно е дека таа е балансирана. Сепак, просметките се предвидени со помош на индексот Џини. Сепак, просметките се предвидени со помош на индексот Џини. Сепак, просметките се предвидени со помош на индексот Џини.

12:30 PM
12/2/2021

Кои од наведените може да се користат како критериум за поделба (Splitting Criterion) на јазлите кај дрвата за одлучување?

- Грешка при класификација , Ентропија , Индексот Цини

Machine Learning 3

Bagging: We divide the data into multiple different sets via bootstrapping(we divide the data in the train and test part differently every time, so we have different validation data. The bootstrapping technique uses sampling with replacements to make the selection procedure completely random.) and for each data set we build a separate classifier (decision tree). We combine the different classifiers at the end and we choose the class with the majority of the voting or for regression the average output (aggregation). Bagging has high expressiveness and low variance. There can be overfit or underfit if the tree is too large or too shallow. The major drawback is that we can not track the logic of the averaged model.

Ensemble method: a method of building a single model by training and aggregating multiple models.
→ out-of-bag error (For each point in the training set, we average the predicted output for this point over the models whose bootstrap training set excludes this point. We compute the error for this point. Then average this for the whole training set.

Calculate the total amount that the MSE (for regression) or Gini index (for classification) is decreased due to splits over a given predictor, averaged over all B Trees. → variable importance

Random Forests

Random Forest is a modified form of bagging that creates ensembles of independent decision trees. What is different from Bagging is that we randomly select a set of J' predictors from the full set of predictors. From amongst the J' predictors, we select the optimal predictor.

Hyperparameters: (all though in theory each tree is full size), if the OOB (error) is big change them

- the number of predictors to randomly select at each split ($\sqrt{N_j}$ for classification, $\frac{N}{3}$ for regression)
- the total number of trees in the ensemble
- the minimum leaf node size

Record the prediction accuracy on the oob samples for each tree.

Randomly permute the data for column j in the oob samples the record the accuracy again.

The decrease in accuracy as a result of this permuting is averaged over all trees, and is used as a measure of the importance of variable j in the random forest.

When the number of predictors is large, but the number of relevant predictors is small, random forests can perform poorly.

Boosting

Sequential algorithm where at each step, a weak learner is trained based on the results of the previous learner.

Adaptive Boosting: Reweighting datapoints based on performance of last weak learner. Focuses on points where previous learner had trouble. It puts bigger weights on missed classes.

Example: AdaBoost.

- Gradient Boosting: Train new learner on residuals of overall model. Constitutes gradient boosting because approximating the residual and adding to the previous result is essentially a form of gradient descent. It doesn't focus on the missed classes, instead it tries to improve the loss function.

Example: XGBoost.

If we have 1. T^0 , set $T = T^0$, 2. then we have T^1 , 3. set $T \leftarrow T + \lambda T^{(1)}$ 4. Compute residuals

$$\text{set } r_n \leftarrow r_n - \lambda T^i(x_n), n = 1, \dots, N$$

If we want to easily reason about how to choose λ and investigate the effect of λ on the model T , we need to formulate gradient boosting as a type of gradient descent.

For a constant learning rate, λ , if λ is too small, it takes too many iterations to reach the optimum. If λ is too large, the algorithm may 'bounce' around the optimum and never get sufficiently close. If λ is a constant, then it should be tuned through cross validation, else choose dynamic lambda that is small around the optimum and larger when further from the optimum.

Naive Bayes Classifier • **Bayes Theorem:** $P(C|A) = \frac{P(A|C)P(C)}{P(A)}$

$$P(A_1, A_2, \dots, A_n|C) = P(A_1|C) P(A_2|C) \cdots P(A_n|C)$$

$X = (\text{Refund} = \text{Yes}, \text{Status} = \text{Single}, \text{Income} = 80K)$

- For the class $C = \text{'Evade'}$, we want to compute:
 $P(C = \text{Yes}|X)$ and $P(C = \text{No}|X)$
- We compute:
 - $- P(C = \text{Yes}|X) = P(C = \text{Yes}) * P(\text{Refund} = \text{Yes} | C = \text{Yes})$
 $* P(\text{Status} = \text{Single} | C = \text{Yes})$
 $* P(\text{Income} = 80K | C = \text{Yes})$
 - $- P(C = \text{No}|X) = P(C = \text{No}) * P(\text{Refund} = \text{Yes} | C = \text{No})$
 $* P(\text{Status} = \text{Single} | C = \text{No})$
 $* P(\text{Income} = 80K | C = \text{No})$

$$P(C = c) = \frac{N_c}{N}$$

Discrete attributes:

$$P(A_i = a | C = c) = \frac{N_{a,c}}{N_c}$$

$N_{a,c}$: number of instances having attribute $A_i = a$ and belong to class c

N_c : number of instances of class c

If one of the conditional probability is **zero**, then the entire expression becomes zero

Laplace Smoothing:

$$P(A_i = a | C = c) = \frac{N_{ac} + 1}{N_c + N_i}$$

– N_i : number of attribute **values** for attribute A_i

Given a Test Record:

With Laplace Smoothing

$X = (\text{Refund} = \text{Yes}, \text{Status} = \text{Single}, \text{Income} = 80\text{K})$

naive Bayes Classifier:

$P(\text{Refund}=\text{Yes} \text{No}) = 4/9$ $P(\text{Refund}=\text{No} \text{No}) = 5/9$ $P(\text{Refund}=\text{Yes} \text{Yes}) = 1/5$ $P(\text{Refund}=\text{No} \text{Yes}) = 4/5$ $P(\text{Marital Status}=\text{Single} \text{No}) = 3/10$ $P(\text{Marital Status}=\text{Divorced} \text{No}) = 2/10$ $P(\text{Marital Status}=\text{Married} \text{No}) = 5/10$ $P(\text{Marital Status}=\text{Single} \text{Yes}) = 3/6$ $P(\text{Marital Status}=\text{Divorced} \text{Yes}) = 2/6$ $P(\text{Marital Status}=\text{Married} \text{Yes}) = 1/6$ For taxable income: If class=No: sample mean=110 sample variance=2975 If class=Yes: sample mean=90 sample variance=25

- $P(X | \text{Class}=\text{No}) = P(\text{Refund}=\text{No} | \text{Class}=\text{No}) \times P(\text{Married} | \text{Class}=\text{No}) \times P(\text{Income}=120\text{K} | \text{Class}=\text{No}) = 4/9 \times 3/10 \times 0.0062 = 0.00082$
 - $P(X | \text{Class}=\text{Yes}) = P(\text{Refund}=\text{Yes} | \text{Class}=\text{Yes}) \times P(\text{Married} | \text{Class}=\text{Yes}) \times P(\text{Income}=120\text{K} | \text{Class}=\text{Yes}) = 1/5 \times 3/6 \times 0.01 = 0.001$
 - $P(\text{No}) = 0.7, P(\text{Yes}) = 0.3$
 - $P(X | \text{No})P(\text{No}) = 0.0005$
 - $P(X | \text{Yes})P(\text{Yes}) = 0.0003$
- => Class = No

Without :

- We compute:
 - $P(C = \text{Yes} | X) = P(C = \text{Yes}) * P(\text{Refund} = \text{Yes} | C = \text{Yes}) * P(\text{Status} = \text{Single} | C = \text{Yes}) * P(\text{Income} = 80\text{K} | C = \text{Yes}) = 3/10 * 0 * 2/3 * 0.01 = 0$
 - $P(C = \text{No} | X) = P(C = \text{No}) * P(\text{Refund} = \text{Yes} | C = \text{No}) * P(\text{Status} = \text{Single} | C = \text{No}) * P(\text{Income} = 80\text{K} | C = \text{No}) = 7/10 * 3/7 * 2/7 * 0.0062 = 0.0005$

Computing the conditional probabilities involves multiplication of many very small numbers. (so we use the logarithm of the conditional probability). Naïve Bayes is commonly used for text

classification.

Function	XGBoost	CatBoost	Light GBM
Important parameters which control overfitting	1. learning_rate or eta – optimal values lie between 0.01-0.2 2. max_depth 3. min_child_weight : similar to min_child_leaf; default is 1	1. Learning_rate 2. Depth - value can be any integer up to 16. Recommended - [1 to 10] 3. No such feature like min_child_weight 4. L2-leaf-reg : L2 regularization coefficient. Used for leaf value calculation (any positive integer allowed)	1. learning_rate 2. max_depth : default is 20. Important to note that tree still grows leaf-wise. Hence it is important to tune num_leaves (number of leaves in a tree) which should be smaller than $2^{(max_depth)}$. It is a very important parameter for LGBM 3. min_data_in_leaf : default=20, alias= min_data , min_child_samples
Parameters for categorical values	Not Available	1. cat_features : It denotes the index of categorical features 2. one_hot_max_size : Use one-hot encoding for all features with number of different values less than or equal to the given parameter value (max – 255)	1. categorical_feature : specify the categorical features we want to use for training our model
Parameters for controlling speed	1. colsample_bytree : subsample ratio of columns 2. subsample : subsample ratio of the training instance 3. n_estimators : maximum number of decision trees; high value can lead to overfitting	1. rsm : Random subspace method. The percentage of features to use at each split selection 2. No such parameter to subset data 3. iterations : maximum number of trees that can be built; high value can lead to overfitting	1. feature_fraction : fraction of features to be taken for each iteration 2. bagging_fraction : data to be used for each iteration and is generally used to speed up the training and avoid overfitting 3. num_iterations : number of boosting iterations to be performed; default=100

Кои од наведените параметри се дел од хиперпараметрите за тренирање на XGBoost моделот?

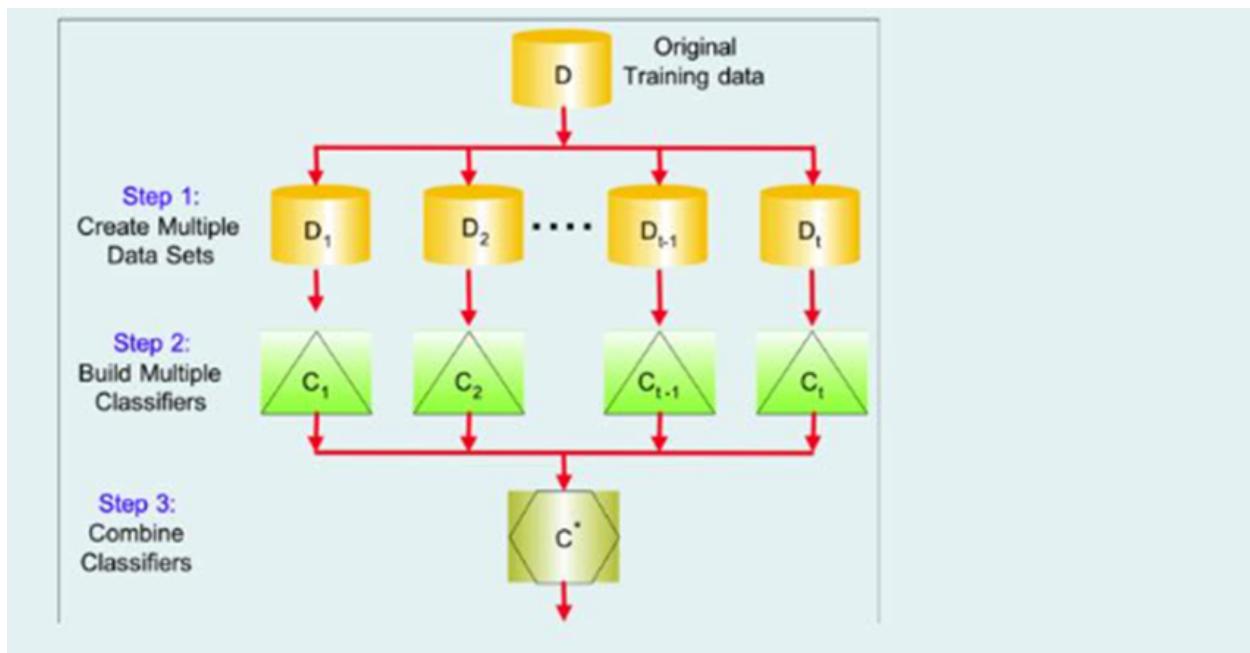
- a. n_estimators
- b. min_depth
- c. learning_rate
- d. max_depth

Што претставува хиперпараметарот n_estimators = 5 во XGBoost моделот?

- a. 5 процесори да се искористат за тренирање на моделот
- b. 5 внатрешни јазли во дрвото на одлука
- c. 5 дрва на одлука кои паралелно ќе се изградат
- d. 5 листа на дрвото на одлука

Question 1
Not yet
answered
Marked out of
20.00
Flag question

Описете ја техниката на машинското учење со ансамбли која се нарекува Bagging.



Што претставувашемата на сликата? Bagging

Прашање 3.1

- Наивниот Баесов класификатор претпоставува:

- A) дека атрибутите се слични помеѓу себе
- B) дека атрибутите се различни помеѓу себе
- C) дека атрибутите се независни помеѓу себе
- D) дека атрибутите се зависни помеѓу себе

Пред табелата колку е $P(\text{ветер} = \text{слаб} | \text{можност за игра} = \text{да})$

Опис	Температура	Влажност	Ветер	Можност за игра
Сончево	Жешко	Голема	Слаб	Не
Сончево	Жешко	Голема	Јак	Не
Облачно	Жешко	Голема	Слаб	Да
Дождливо	Умерено	Голема	Слаб	Да
Дождливо	Студено	Нормална	Слаб	Да
Дождливо	Студено	Нормална	Јак	Не
Облачно	Студено	Нормална	Јак	Да
Сончево	Умерено	Голема	Слаб	Не
Сончево	Студено	Нормална	Слаб	Да
Дождливо	Умерено	Нормална	Слаб	Да
Сончево	Умерено	Нормална	Јак	Да
Облачно	Умерено	Голема	Јак	Да
Облачно	Жешко	Нормална	Слаб	Да
Дождливо	Умерено	Голема	Јак	Не

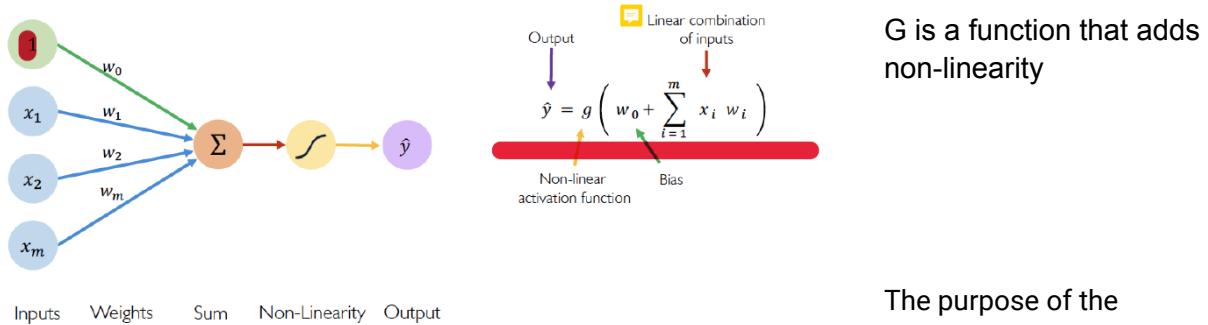
- A) 6/9
- B) 2/8
- C) 6/14
- D) 6/8

Artificial Neural Networks and Deep Learning

Processing information in ANN: a single neuron (processing element - PE): has inputs, weights, summation, transfer function and outputs.

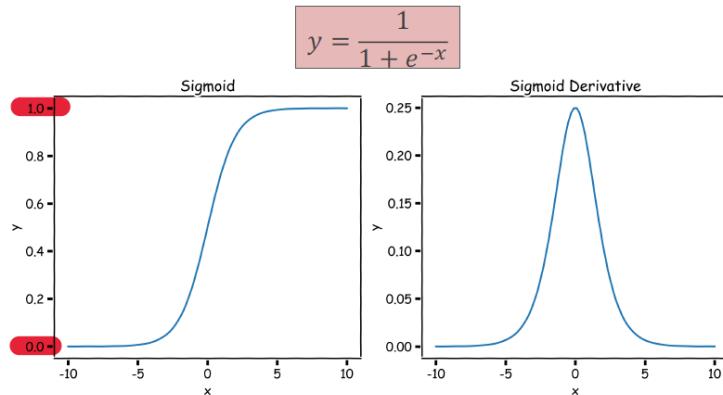
Deep Learning: Extracting patterns from data using neural networks. Multiple stages of the feature learning process.

In a deep network, high levels can express combinations between features learned at lower levels



The purpose of the activation function is to add non-linearities to the network and to ensure that the gradients remain large through hidden unit.

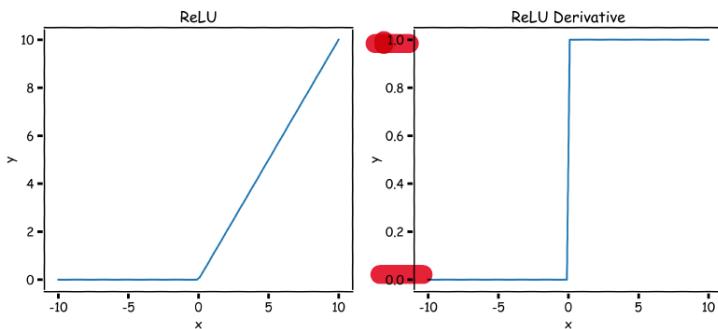
Sigmoid (aka Logistic)



Derivative is **zero** for much of the domain. This leads to “vanishing gradients” in backpropagation.

Rectified Linear Unit (ReLU)

$$y = \max(0, x)$$



Two major advantages:

1. No vanishing gradient when $x > 0$
2. Provides sparsity (regularization) since $y = 0$ when $x < 0$

Leaky ReLU tries to fix ReLU so the derivative is nonzero everywhere (pod ReLU)

Gen ReLU (nand ReLU)

Maxout: Max of k linear functions. Directly learn the activation function.

Swish: better than ReLU on deeper models

Loss Function: Takes all of these results and averages them and tells us how bad or good the computer or those weights are.

Output Type	Output Distribution	Output layer	Cost Function
Binary	Bernoulli	Sigmoid	Binary Cross Entropy
Discrete	Multinoulli	Softmax	Cross Entropy
Continuous	Gaussian	Linear	MSE (Mean Squared Error)
Continuous	Arbitrary	-	-

Cross-entropy bara najcisti klasi.

One hidden layer is enough to represent an approximation of any function to an arbitrary degree of accuracy. But shallow networks need exponentially more width and tend to overfit more.

If we have 96.0 accuracy that is two times better than 92.0 accuracy, since in the first one we have 4 mistakes in 100 trials and in the second 8 mistakes.

Exploding and Vanishing gradients: Exploding gradients lead to cliffs

Good practice to normalize features before applying a learning algorithm, therefore we limit the function and the gradients. Normalization can reduce expressive power

Norm penalty: penalizes on the weights of the affine transformations and leaves biases unregularized. This is because of the fact that biases require lesser data to fit accurately than the weights.

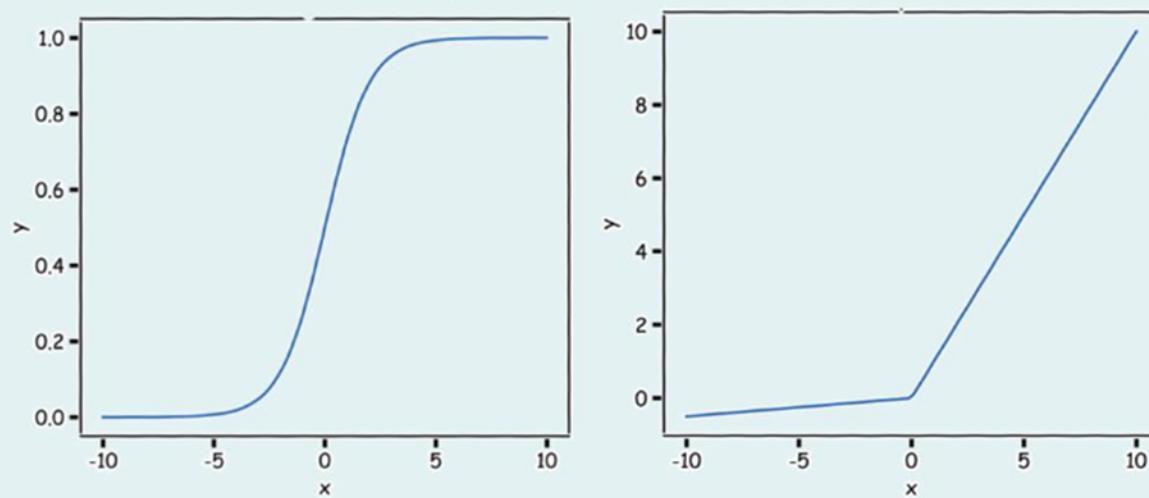
Early stopping: terminate while validation set performance is better

Noise robustness: Adding noise means that the network is less able to memorize training samples because they are changing all of the time, resulting in smaller network weights and a more robust network that has lower generalization error

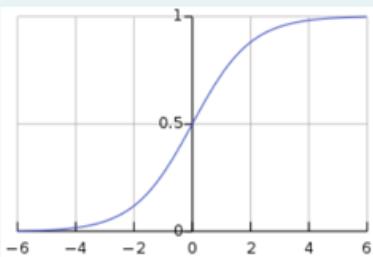
Dropout: Randomly set some neurons and their connections to zero.

Dropout weight scaling: Reducing weight with some probability $p \rightarrow w = w^*p$

На сликата се прикажани кои активациски функции ?



Која активациска функција е представена на графикот?



- a. relu
- b. sigmoid
- c. linear

Што претставува поимот отфрање (dropout) во контекст на невронски мрежи?

Select one:

- a. Бришење од меморијата при тестирање.
- b. Случајно поставување на активацијата и тежините на врските на некои неврони на нула.
- c. Трајно бришење од меморијата.
- d. Откривање на недостатоци и нивно отфрање.

датоци-2021/2022/Z

Question 9
Not yet answered
Marked out of 1.00
Flag question

Во кој случај би било најдобро да се употреби ReLU као излезно ниво ја невронската мрежа.

- a. Кога влезовите во мрежата се дискретни предности.
- b. Кога како мрежа за пресметка на загуба во мрежата се користи MSE (Mean Squared Error)
- c. Кога бројот на влезови е поголем од бројот на излези во невронската мрежа
- d. Кога сакаме да добиеме по број прогресарне на резултатите на GPU
- e. Кога имаме бинарна класификација

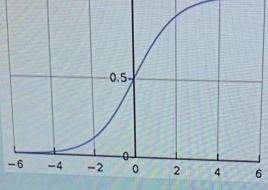
Previous page

Announcements

Листа 10...

Question 9
Not yet answered
Marked out of 1.00
Flag question

Која активацијска функција е претставена на графикот?



The graph shows the sigmoid function, also known as the logistic function. It is a curve that passes through the point (0, 0.5). The x-axis ranges from -6 to 6, and the y-axis ranges from 0 to 1. The curve is symmetric about the point (0, 0.5).

- a. relu
- b. linear
- c. sigmoid

v/mod/quizz/attempt.php?attempt=277790&cmid=12856&page=8#

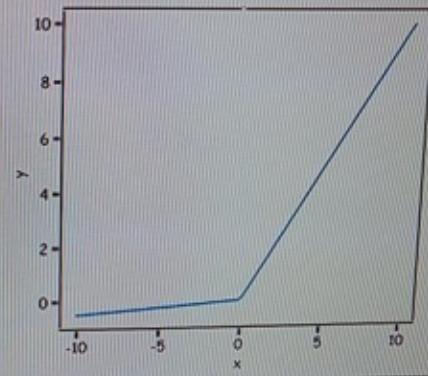
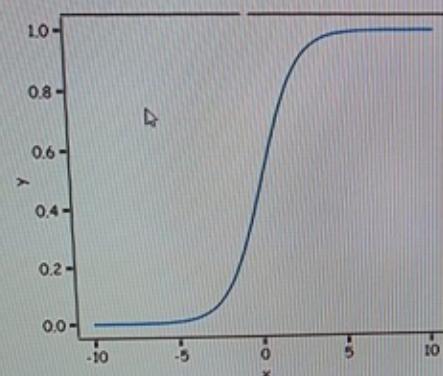
Question 8

Not yet
answered

Marked out of
1.00

Flag question

На сликава се прикажани кои активацијски функции?



Hyperbolic Tangent (Tanh)

Leaky ReLU

Softplus

Rectified Linear Unit (ReLU)

Swish

Sigmoid (Logistic)

Previous page

атоци-2021/2022/Z

/ 12 - предел

Question 2

Not yet
answered

Marked out of
1.00

Flag question

Кои карактеристики треба да ги има активацијата функција кај невроните мрежи?

Select one or more:

- a. Да има некаква нелинеарност.
- b. Да овозможи градиентите да останат доволно големи и преку неколку скриени слоја.
- c. Да дава активација само за позитивни влезови.
- d. Да е заоблена.

Previous page

Announcements

Jump to...

Колку често можат да се ажурираат тежините кај невронските мрежи?

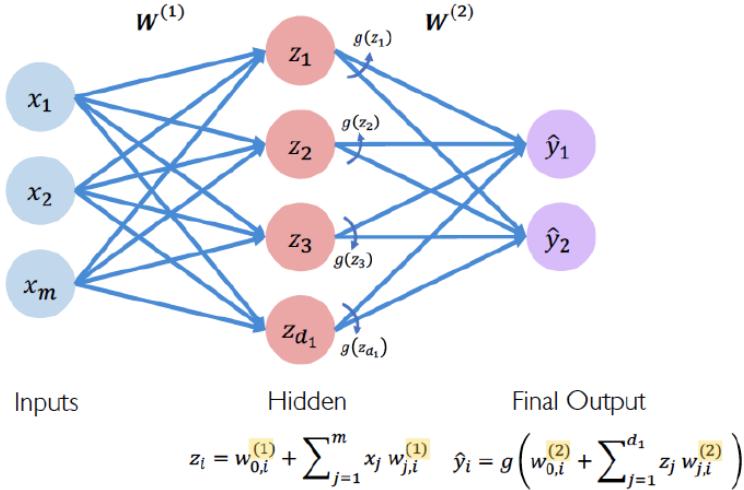
Select one or more:

- a. Ажурирање во серии (batch)
- b. Ажурирање во случајно расфрлани мини-серии (mini-batches)
- c. Ажурирање после секој примерок во множеството за обука
- d. Ажурирање во моменти
- e. Ажурирање во конволуции

Во кој случај би било најдобро да се употреби Softmax како излезно ниво кај невронските мрежи

- a. Кога имаме класификација во повеќе од две класи
- b. Кога сакаме да добиеме по брзо процесирање на резултатите на GPU
- c. Кога како мрека за пресметка на загуба во мрежата се користи MSE (Mean Squared Error)
- d. Кога бројот на влезови е поголем од бројот на излези во невронската мрежа
- e. Кога имаме длабока невронска мрежа

Advanced Neural Networks



Softmax layer as the output layer: applied on all neurons in the layer (like probabilities, all sum up to 1)

We should move in the direction 180 degrees with the gradient or opposite to the gradient for maximum loss reduction

Algorithm

1. Initialize weights randomly $\sim \mathcal{N}(0, \sigma^2)$
2. Loop until convergence:
3. Compute gradient, $\frac{\partial J(\mathbf{W})}{\partial \mathbf{W}}$
4. Update weights, $\mathbf{W} \leftarrow \mathbf{W} - \eta \frac{\partial J(\mathbf{W})}{\partial \mathbf{W}}$
5. Return weights

But this is complex since by applying the chain rule, every weight in the network uses gradients from later layers;

We should set a learning rate η (not too big or not too small)

Adaptive Learning Rate Algorithms: Momentum, Adagrad, Adadelta, Adam, RMSProp

It is very complex to compute the gradient each time. With picking one point it is inaccurate, so we choose to pick batch of B data points → faster to compute

How often are the weights updated?

- Updating after every sample in the training test: this is like cross validation, but it can be time consuming and outliers can throw off the model
- Updating in batch : this is like a random forest. Dividing training samples into several large batches, running and then calculating backpropagation on all the samples together. Training is performed iteratively on each of the batches. This makes the model more resistant to outliers and variance in the training set
- Randomized mini-batches: combination of previous two. This avoids a biased selection of samples in each batch, which can lead to the local optimum. They give us more accurate estimation of the gradient, smoother convergence and allow for larger learning rate. They can parallelize computation and achieve significant speed increases of GPU.
Take batch → update → next batch → update (we do this for each epoch)

Large batch size yield worse performance. Smaller batch size means more updates in one epoch. Gradient descent never guarantee global minima. (if the learning rate is too small it can be stuck)

Movement = Negative of a gradient + Movement (parameters are saved after each check point)

In a fully connected layer each unit is connected to all units of the previous layers.

In a convolutional layer, each unit is connected to a constant number of units in a local region of the previous layer. CNN need much less parameters.

Furthermore, in a convolutional layer, the units all share the weights for these connections, as indicated by the shared line-types.

Na primer koga iame slika so pikseli nema potreba da se povrzani pikselite so pikselite vo drugiot kjos, tuku so sosednite samo

Three Stages of a Convolutional Layer:

- Convolution stage
- Nonlinearity: a nonlinear transform such as rectified linear or tanh
- Pooling: output a summary statistics of local input, such as max pooling and average pooling

CNN for images : input 2D array

CNN for text: 1D convolution, in which context is the word seen. when a convolutional kernel is applied to different sets of similar words, it will produce a similar output value!

Pooling: Max pooling, average pooling

Maxpooling operation, forces the network to retain only the maximum value in a feature vector, which should be the most useful, local feature.

To set up a network so that it is capable of learning a variety of different relationships between words, you'll need many filters of different heights

Autoencoders: An Autoencoder is a feedforward neural network that learns to predict the input itself in the output.

The input-to-hidden part corresponds to an encoder

The hidden-to-output part corresponds to a decoder.

Deep Autoencoder: extra hidden layers, gradient becomes too small since it passes many different levels.

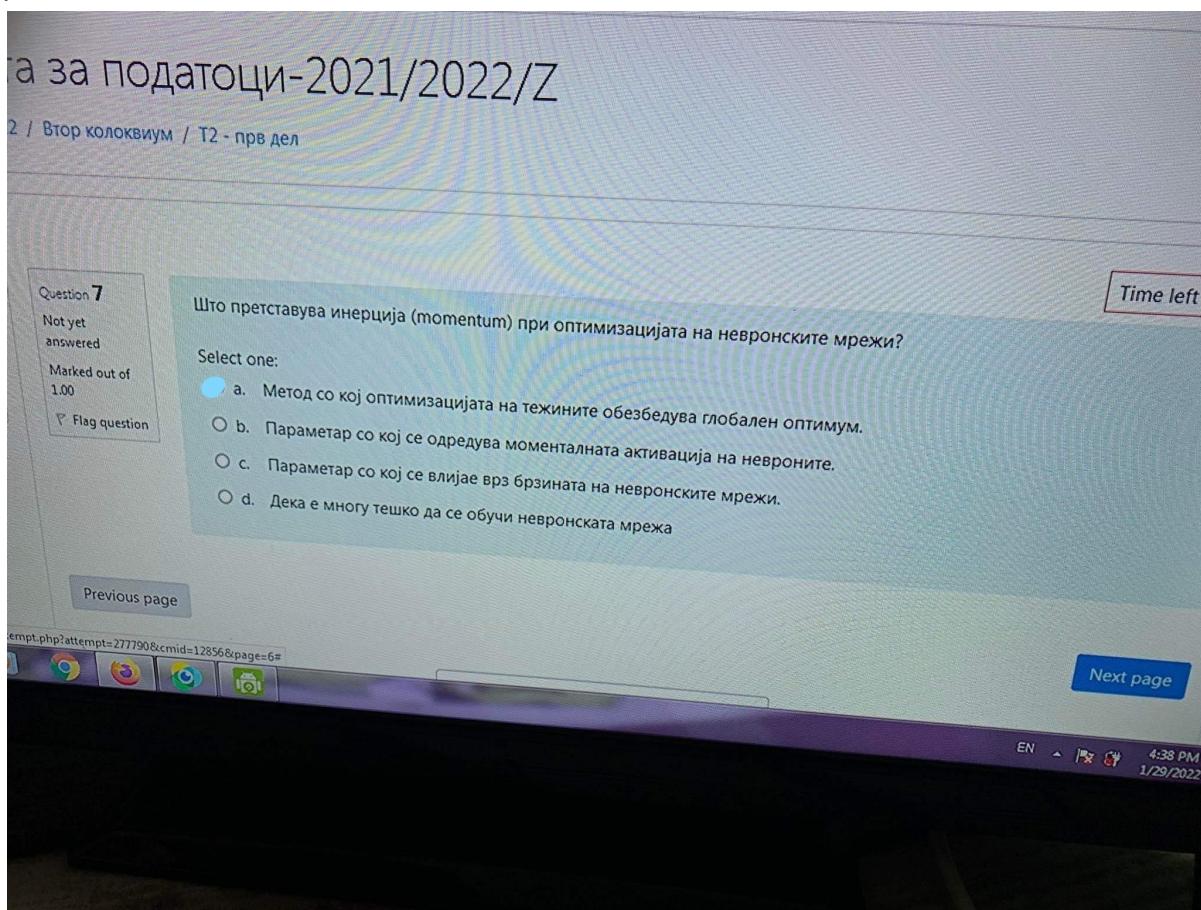
Neural Network with Memory: Named Entity Recognition, Sequence Modelling

RNN: The output of hidden layer are stored in the memory. Memory can be considered as another input. Backpropagation through time. Not easy to learn

Long Short-term Memory LSTM : Input gate, Output gate, Memory cell, Forget Gate

LSTMs contain information outside the normal flow of the recurrent network in a gated cell.

The cells learn when to allow data to enter, leave or be deleted through the iterative process of making guesses, back-propagating error, and adjusting weights via gradient descent. -> 4 times of parameters



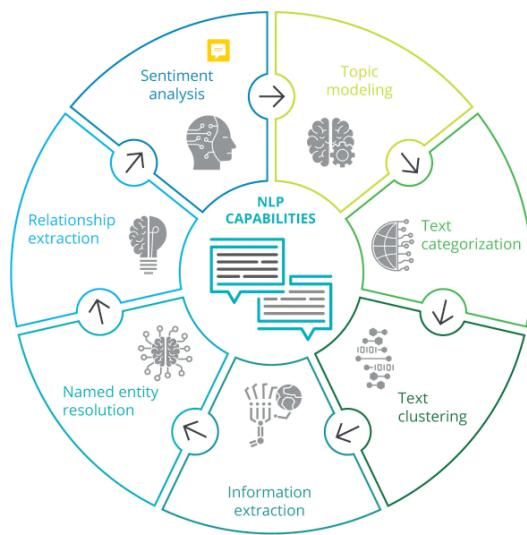
Кои од следниве се називи на алгоритми за оптимизација кај невронските мрежи?

- Adam , Adagrad

Natural Language Processing

Natural language processing (NLP) describes the interaction between human language and computers. Ability to identify the appropriate word, phrase, or response by using context clues, the same way as humans.

Seven key technical capabilities of NLP:



These capabilities allow organizations to recognize patterns, categorize topics, and analyze public opinion.

Topic modeling is a method based on statistical algorithms to help uncover hidden topics from large collections of documents. Topic models are unsupervised methods of NLP

Named Entity Recognition

Challenges in Natural Language Processing

- Languages often seem to behave in arbitrary ways and forms (cabz, cats)
- Ambiguity, sarcasm and irony are often not apparent from purely textual information
- Domain-specific terms and phrases that may not even be grammatically correct

Bag-of-words (BoW): a featurization that uses a vector of word counts (or binary) ignoring order.
(most common since provides numerical values) We should discard the rarest words
But information isn't in the most common words, so compromise.

N-gram: a consecutive sequence of words in a text.

Skip-grams: displacements of ..., -3, -2, -1, +1, +2, +3, ... in each sentence where the word occurs.

Parse trees

We use WordNet's list of synonyms since with one hot encoder hotel is different than motel
A word's meaning is given by the words that frequently appear close-by

Word vectors are also called word embeddings or (neural) word representations. They are a distributed representation.

Vector Embedding of Words: Word embeddings depend on a notion of word similarity. Similar words occur in similar contexts.

Word2vec uses words a few positions away from each center word ("skip-grams.") It considers all words as center words, and all their context words. Word2vec optimizes a softmax loss for each word. The softmax function maps arbitrary values x_i to a probability distribution p_i . It captures the similarity of words

Transfer learning: train model on large amount of general data (example Wikipedia)

Transformers typically undergo semi-supervised learning involving unsupervised pre training followed by supervised fine-tuning. Pretraining is typically done on a much larger dataset than fine tuning, due to the restricted availability of labeled training data.

Каква димензионалност треба да е влезното тренирачко множество кај LSTM невронската мрежа?

- a. 2D - матрица
- b. 1D
- c. 3D

Што претставува поимот отфране (dropout) во контекст на невронски мрежи?

Select one:

- a. Бришење од меморијата при тестирање.
- b. Случајно поставување на активацијата и тежините на врските на некои неврони на нула.
- c. Трајно бришење од меморијата.
- d. Откривање на недостатоци и нивно отфране.

Каков вид на учење се реализира кај Автоенкодерите?

Select one or more:

- a. надгледувано (supervised)
- b. полу-надгледувано (semi-supervised)
- c. само-надгледувано (self-supervised)
- d. со поттикнување (reinforcement)

Што претставува Parts of Speech Tagging

- Процес на означување на збор во текст што одговара на категоријата зборови (или , поопшто лексички единици) кои имаат слични граматички својства (именки глаголи)

WordVec како основа за креирање на Embeddings користи:

Select one:

- a. n grams
- b. part of speech tagging
- c. Skip grams
- d. one-hot embeddings

when to use softmax

Natural Language Processing and Transformers

A Language Model represents the language used by a given entity.

A Language Model estimates the probability of any sequence of words

A **Language Model** is useful for:

Generating Text

- Auto-complete
- Speech-to-text
- Question-answering / chatbots
- Machine translation

Classifying Text

- Authorship attribution
- Detecting spam vs not spam

Language Modeling : **Bigrams**→ probability of two consecutive words (NOT GOOD)

Neural networks: First, each word is represented by a word embedding. Words that are more semantically similar to one another will have embeddings that are proportionally similar/ (pre-existing word embeddings are used)

Neural Approach #1: Feed-forward Neural Network: using windows of words, predict the next word

FFNN STRENGTHS

- No sparsity issues (it's okay if we've never seen a segment of words)
- No storage issues (we never store counts)

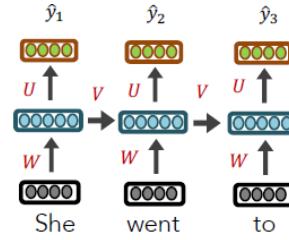
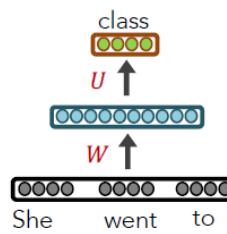
FFNN ISSUES

- Fixed-window size can never be big enough. Need more context.
- Increasing window size adds many more weights
- Requires inputting entire context just to predict one word

Neural Approach #2: Recurrent Neural Network (RNN)

- Can handle infinite-length sequences (not just a fixed-window)
- Has a "memory" of the context (thanks to the hidden layer's recurrent loop)
- Same weights used for all inputs, so word order isn't skewed (like FFNN)
- Slow to train (BPTT)
- Due to "infinite sequence", gradients can easily vanish or explode → if close to zero vanish, or easily go to infinity

$$P(\text{went}|\text{She}) = \frac{\text{count}(\text{She went})}{\text{count}(\text{She})}$$



n-grams

- Basic counts; fast
- Fixed window size
- Sparsity & storage issues
- Not robust

FFNN

- Kind of robust... almost
- Fixed window size
- Weirdly handles context positions
- No "memory" of past

RNN

- Handles infinite context (in theory)
- Robust to rare words
- Slow
- Difficulty with long context

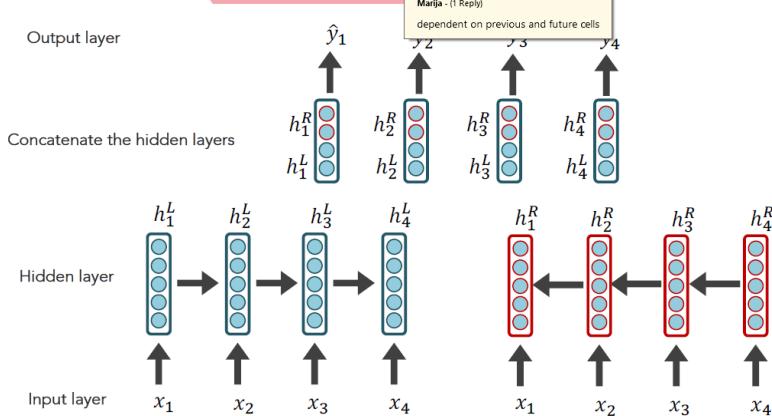
Long short-term memory (LSTM)

A type of RNN that is designed to better handle long-range dependencies

In "vanilla" RNNs, the hidden state is perpetually being rewritten. Here additionally we have a dedicated memory cell for long-term events. Old memories are forgotten.

- Captures long-range dependencies shockingly well
- Has more weights to learn than vanilla RNNs;
- Requires a moderate amount of training data
- Can still suffer from vanishing/exploding gradients

RNN Extensions: **Bi-directional LSTMs (review)**



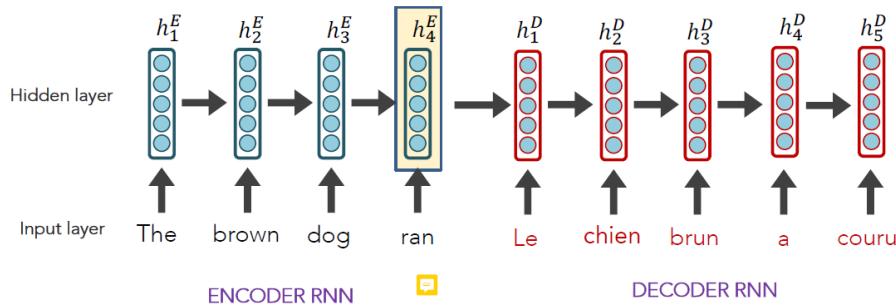
- Usually performs at least as well as uni-directional RNNs/LSTMs
- Slower to train
- Only possible if access to full data is allowed

Sequence-to-Sequence (seq2seq)

Seq2seq models are comprised of 2 RNNs: 1 encoder, 1 decoder

The final hidden state of the encoder RNN is the initial state of the decoder RNN. Loss from the decoder outputs is calculated and weights are updated all the way to the beginning of the encoder. Decoder generates outputs one word at a time.

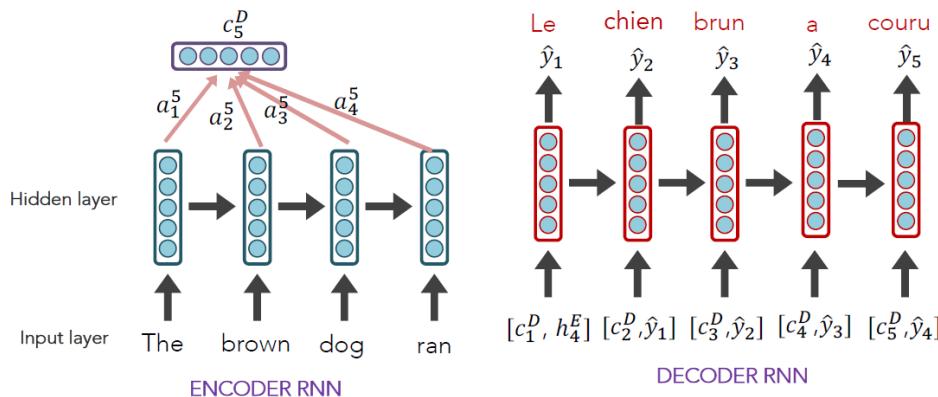
All data needs to be contained in h_4^E --> encoder part and the encoder is based on the last layer. The decoder takes the whole sequence at once when it is in the last layer of the encoder because there is no sense in word by word translation. It is challenging that everything is stored in the last layer, but to make a fully connected network is very complex.



Seq2seq + Attention → instead of fully connecting the network. We make a matrix with all the dependencies between the words. This connects everything to everything and we can go back fast because we **only have to access the matrix O(1)**. This greatly improves seq2seq results and allows us to visualize the contribution each word gave during each step of the decoder

seq2seq + Attention

NOTE: each attention weight a_i^j is based on the decoder's current hidden state, too.



From the matrix we get weights for the row that needs to be translated. This allows us to visualize the contribution each word gave during each step of the decoder

Challenges with RNNs

- Long range dependencies
- Gradient vanishing and explosion
- Large # of training steps
- Recurrence prevents parallel computation

Transformer Networks

- Facilitate long range dependencies
- No gradient vanishing and explosion
- Fewer training steps
- No recurrence that facilitate parallel computation

Transformers: It is faster. It uses encoder and decoder is based on attention.

It has layer normalization (0 mean and 1 variance) and positional embedding(the position of a word in a sentence affects its meaning). If we write them as binary numbers they look like sine or cosine functions.

Stack of encoders and stack of decoders

More parameters → more powerful

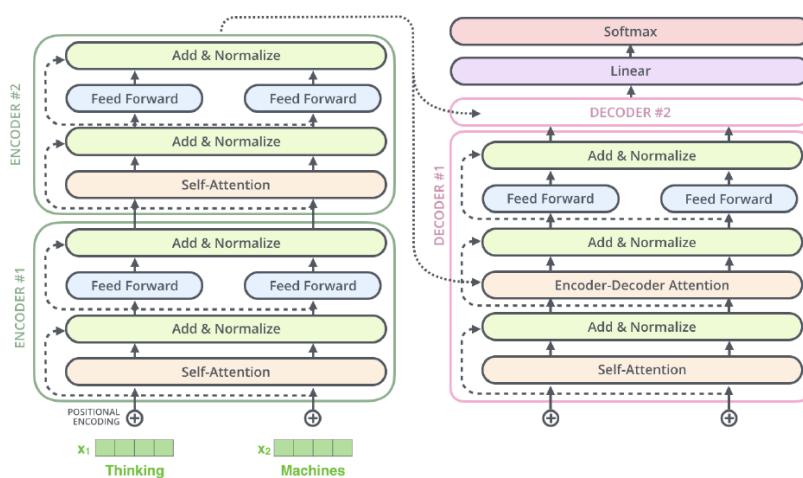
In the encoder layer we have Self-Attention, Normalization and FFNN.



Attention and Transformers

The complete transformer

The encoder-decoder attention is just like self attention, except it uses several inputs from the top of encoder output, plus its own input



Vectorot moze da se shifta, vo zavisnost od contextot i taka ucime

Separation focus of development:

So separation with BERT(to encode,train the model) and GPT(to process text,decode)

Question 4
Answer saved
Marked out of 1.00

Kaj Обработката на природни јазици се среќаваат следниве задачи:

Select one or more:

- a. Категоризација на теми
- b. Препознавање на векторски претстави на зборовите (word embeddings)
- c. Извлекување на контекстни зборови (skip-grams)
- d. Препознавање на именувани нешта

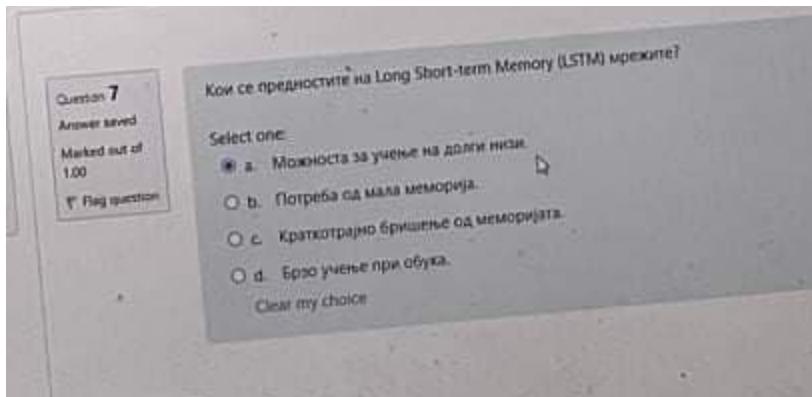
Што е точно за моделот seq2seq?

Select one or more:

- a. Крајниот скриен слој на енкодерскиот дел е влезен слој за декодерскиот дел.
- b. Обуката се одвива како и кај другите Рекурентни невронски мрежи.
- c. Предноста на seq2seq е што целото значење на реченицата е претставено во крајниот скриен слој на енкодерскиот дел.
- d. При тестирањето се генерираат збор по збор, сè додека не се добие на излез знак за крај на реченицата.

Кои од наведените карактеристики се новитети кај Трансформер моделите?

- a. Positional embeddings
- b. Self Attention layer
- c. Feedforward Network
- d. Tokenization



Question 6
Not yet answered
Marked out of 1.00
Flag question

Time left

Нека е дадена реченицата:
"It was a bright cold **day** in April, and the clocks were striking"
Skip-gram со големина на прозорец три за зборот **day** е:

- a. a bright cold
- b. in April, and
- c. was bright cold April clocks were
- d. a bright cold in April and

Објаснете зошто Attention нивото ги подобрува резултатите кај seq2seq моделите

Кои од следниве репозиториуми/библиотеки се користат за едноставно споделување на претренираните NLP модели.

- **HuggingFace transformers library, PyTorch Hub, TensorFlow-Hub**

Question 1
Complete
Mark 15.00 out of 15.00
Flag question
Edit question

Објаснете го концептот на трансформер моделите, што е тоа што ги прави подобри од сите останати модели?

Трансформер моделите се базирани на transformer learning. Составени се од input, output, encoder и decoder. Повеќе енкодери се споени во еден енкодер (поминувајќи низ повеќе нивоа моделот подобро учи) и повеќе декодери се споени во еден декодер. Извесот од енкодерот се практикано влез на декодерот. Енкодерот се обидува да ја разбере реченицата (секвенца од зборови од коишто добиени соодветни embeddings) и потоа тоа што го разбрал го препрака на декодерот. Исто така има и self-attention кој помага да се најде главниот збор според околните зборови и word position кој никажува до каде сме во реченицата (до кој збор). Трансформерите

Еден од најдобрите јазични модели GPT-2 се потпира на трансформер архитектурата. Кој дел од трансформер архитектурата се користи во GPT-2?

- a. Decoder+Encoder
- b. Првите 9 нивоа од Decoder делот
- c. Decoder
- d. Encoder

За што се користи Latent Dirichlet Allocation (LDA) алгоритмот

Select one:

- a. Topic modeling
- b. Part-of-Speech (POS) tagging
- c. Named Entity Recognition
- d. Open Information Extraction

Кои особености ги има Преносното учење (Transfer Learning)?

Select one or more:

- a. Врши пренос на моментите во друга невронска мрежа.
- b. Овозможува подобрување на перформансите.
- c. Може да го научи преносното значење на зборовите.
- d. Врши пренос на испуштените јазли (drop-out) во друга невронска мрежа.
- e. Користи означени податоци од други или сродни области.

Февруари Испит / Т1 - прв дел (16.02.2022)

Question 4

Not yet
answered

Marked out of
1.00

Flag question

Кои од следните репозиториуми/библиотеки се користат за едноставно споделување на претренирани NLP модели.

Select one or more:

- a. HuggingFace Transformers library
- b. PyTorch Hub
- c. GitHub
- d. TensorFlow-Hub

Question 8

Never Answered

Marked out of
1.00

Flag question

Кои се предностите на Двонасочните LSTM мрежи (Bi-directional LSTMs)?

Select one or more:

- a. Обично се подобри од еднонасочните рекурентни и LSTM мрежи.
- b. Побрзи се при обукувањето.
- c. Го зимаат предвид поширокото значење на контекстот.
- d. Не бараат пристап до сите податоци однапред.

Next page

Unsupervised learning

Clustering : Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups

Important distinction between hierarchical and partitional sets of clusters

- Partitional Clustering: each data object is in exactly one subset
- Hierarchical clustering: A set of nested clusters organized as a hierarchical tree

Centroid: the average of all the points in the cluster

Types of Clusters

- Well-separated clusters: all points are similar between each other
- Prototype-based clusters: all points are most similar to this center comparing to centers of other clusters
- Contiguity-based clusters: a point in a cluster is closer to one or more other points in the cluster than to any point not in the cluster.
- Density-based clusters: a dense region of points, which is separated by low density regions, from other regions of high density (used to eliminate outliers)

K-means Clustering (initial points are assigned randomly, and their choice affects the result)

Euclidian distance used to calculate centroids, but this doesn't work in more dimensional space

- 1: Select K points as the initial centroids.
- 2: **repeat**
- 3: Form K clusters by assigning all points to the closest centroid.
- 4: Recompute the centroid of each cluster.
- 5: **until** The centroids don't change

Empty clusters might appear (not to get this update centroids after each point)

Limitations of K-means: sizes, densities, non-globular shapes

Hierarchical Clustering

Cluster Distance:MIN, MAX, Group Average, Distance Between Centroids

Strength of MIN: Can handle non-elliptical shapes

Limitations of MIN: Sensitive to noise and outliers

Strength of MAX: Less susceptible to noise and outliers

Limitations of MAX: Tends to break large clusters, Biased towards globular clusters

Ward's Method: Similarity of two clusters is based on the increase in squared error when two clusters are merged

Agglomerative Clustering Algorithm

- Most popular hierarchical clustering technique
 - **Key Idea:** Successively merge closest clusters
- Basic algorithm is straightforward
 1. Compute the proximity matrix
 2. Let each data point be a cluster
 3. **Repeat**
 4. Merge the two closest clusters
 5. Update the proximity matrix
 6. **Until** only a single cluster remains

Density Based Clustering: regions of high density that are separated from one another by regions on low density

DBSCAN

Core point, border point, noise point

- 1: Label all points as core, border, or noise points.
- 2: Eliminate noise points.
- 3: Put an edge between all core points within a distance Eps of each other.
- 4: Make each group of connected core points into a separate cluster.
- 5: Assign each border point to one of the clusters of its associated core points

Strengths: • Resistant to Noise
• Can handle clusters of different shapes and sizes

Limitations: • Varying densities
• High-dimensional data

Internal Measures: **SSE** good for comparing two clusterings or two clusters can also be used to estimate the number of clusters

Кои мерки може да ги користиме за сличност помеѓу два кластера?

Select one or more:

- a. Бројот на елементи кои се наоѓаат во кластерите.
- b. Сличноста помеѓу два случајно избрани елементи од двата кластера.
- c. Најмалата различност помеѓу два елементи од кластерите.
- d. Сличноста помеѓу центроидите на двата кластера.

за податоци-2021/2022/Z

Фролоквиум / Т2 - прв дел

Question 1
Not yet answered
Marked out of 1.00
Flag question

Со кој од дадените алгоритми за кластерирање може да се добијат шест кластери како што се прикажани на сликата



- a. K-means
- b. Hierarchical Clustering
- c. DBSCAN
- d. K-means++

Announcements

Jump to...

You are logged in as Matevska Ana Marinko (Log out)
БИО-21/22

На кои од наведените модели за кластерирање потребно е да се наведе бројот на кластери ?

- a. DBCAN Clustering
- b. K-means Clustering
- c. AffinityPropagation Clustering
- d. Agglomerative Clustering