Summer 2024

- KDD Process
- Tasks of Data Mining
- What is Deep Learning
- How are gradients updated?
- What happens if we have a very high learning rate
- Frequent ItemSet Mining Algorithm
- Lots of Questions about Vanishing/Exploding Gradients
- What loss functions do we use?
- Tasks of Unsupervised Learning
- 4 V's of Data
- What is Big Data
- Naive Bayes
- RNNs
- Explain DBSCAN
- What are ReLUs? Why are they better than a sigmoid function
- Weight Initialization?
- Object detection + segmentation
- Attention and transformers
- Which classification algorithm is better for spam detection
- Vanishing/exploding gradients in RNNs
- How does LSTM work
- What are the two tricks that ADAM uses
- What are precision, recall, accuracy?
- How can you evaluate a ML model
- Hierarchical clustering
- How to evaluate an unsupervised algorithm
- Explain Stable Diffusion
- How to build a decision tree + asked to draw decision boundaries
- MLP questions
- When do we define an itemset as frequent
- How do we define Association Rules?
- How do we actually find frequent items with association rules?
- Which clustering algorithms do we have?
- Alternatives to K-means
- How can we find medoids in k-medoids
- Explain agglomerative clustering
- What's the difference between HPC and cluster computing
- How does cluster computing work?
- What's the purpose of using data replication?
- How is data between workers passed?
- How is Big Data related to NN?
- Explain what the output of Image segmentation is

Summer 2023

- KDD Process, role of data mining
- 4 V's of big data
- Favorite classification algorithm
- Gradient descent, momentum
- MLP, why activation functions needed
- Why is gradient descent nice
- Explain DBSCAN, noise handling
- What is horizontal scaling
- What are SVMs and what do they do

- Which losses do we use in neural networks
- Talk about precision/recall/F1 score
- What did you learn in KDDM?
- Noise handling in k-means, what to do to make it more robust
- Draw sigmoid activation function, problem with gradients for large/small values
- Vanishing/exploding gradients in RNNs
- Difference between HPC and cloud cluster systems
- What is MapReduce
- What is frequent itemset mining
- Explain lossy counting
- What is supervised learning? Why split data into train/test? Explain cross-validation
- Why do we use CNNs? New trainable parameters, filter, size, stride, padding
- What is output dimension of a convolution
- How do we shrink dimension in CNN
- Problem with stacking dense layers in NNs
- Complexity of SVM, scalar product in dual form
- Benefits of Spark/Hadoop
- Which process in data mining takes a lot of time
- General idea of PCA
- Explain how decision tree works
- Why decision trees still popular
- How to make semi-structured data usable for classification/clustering
- What is bag of words
- How autoencoder works? Common loss functions
- What are GANs and how do they work? Why adversarial
- How Naive Bayes Classification works? Explain assumption
- Explain KNN for classification
- What are outliers? Ways to detect outliers? Use KMeans for detection?
- Explain HDFS architecture
- CAP theorem. Which properties desirable for HDFS
- Explain CNNs. Convolution, Padding, Stride, Pooling
- Why Leaky ReLU? Is ReLU differentiable?
- Why frequent itemset mining not for streams
- Supervised vs unsupervised learning, evaluation, metrics
- Steps before running classification algorithm on dataset
- How does maxpool work
- Processes on images with NN
- What is segmentation, optimizing loss
- What is object detection, optimizing loss
- How to make ReLU differentiable
- What is dying ReLU
- Assumptions for gradient descent
- What are cloud clusters, connections, why Ethernet
- SQL variety solved by NoSQL
- K-fold vs bootstrapping for small dataset
- What is word2vec
- How divisive hierarchical clustering works
- How agglomerative hierarchical clustering works, problem with single-linkage
- Time complexity of SVM vs NN
- Why NNs better than kernels
- How self-attention layers in transformers work
- Explain OPTICS

Summer 2022

- Definition of KDD and KDD process

- Frequent itemset mining, Frequent itemset
- Batch systems, Hadoop vs Spark, Map Reduce (detail), master slave system
- EM? EM vs K-means? output and convergence
- Frequent itemset mining in streams, lossy counting + why Apriori not suitable
- Apriori in detail
- Evaluation of classification: precision, recall, F1
- Visualization techniques/visualization of results
- 4 V's of Big Data
- Classification setting and methods - linear, SVM, non-linear, Bayes
- Sequential data, distance, algorithms
- Word2vec and preprocessing sequential data (DFT, DWT)
- Streaming systems, Flink advantages
- Naive Bayes Algorithm
- Data mining tasks (classification, clustering, FPM, PM, outlier detection)
- DBSCAN vs K-means (convex, noise points)
- NoSQL
- DFT - feature reduction for numerical sequences, why it works
- HDFS
- Clustering, clustering methods
- Frequent Itemset Mining - Apriori and hash table working
- FP Growth
- Clustering on data streams
- CAP Theorem in detail with examples
- Distributed File Systems (DFS)
- Replication - consistency problem in HDFS
- Horizontal vs vertical scaling
- Outlier detection
- Difference between outliers and noise
- Clustering methods best suited for outlier detection
- OLAP, OLTP
- DBSCAN vs OPTICS, epsilon definition
- DBSCAN convergence criterion
- Silhouette formula and explanation
- PCA - power iteration, SVD
- Relational DB (data type, big data limits)
- Sparse data, high dimensional data
- Feature selection vs dimensionality reduction
- Approach to handle stream data
- Sparse vectors
- Process mining - supervised/unsupervised, conformance checking
- Train test split (why), cross validation
- Buffer overflow handling

Summer 2020

- Define big data
- What you learned in KDD
- Idea of PCA and implementation
- Difference between knowledge discovery and machine learning
- Frequent itemset, association rules
- Frameworks for K-means, cluster centroid updates
- Lossy counting in stream mining
- Spectral clustering
- Semi-structured data handling
- Formal supervised learning task
- Vertical/horizontal scaling, which scales horizontally

- Decision Tree Algorithm
- Why learn BDMA
- Semi-structured data processing
- Frequent Itemset extraction
- Comparison between clustering algorithms, DBSCAN outcome
- Difference between HDFS and Spark, why Spark better
- How Flink works, differences
- SVMs in detail (primal, dual, soft-margin)
- Streaming aspects to keep in mind
- Sequential data vs streams
- Classification formally, favorite algorithm
- HDFS/NoSQL solving size problem
- Density-based algorithm
- DFT and DWT
- Data Mining vs Knowledge discovery
- Spectral clustering algorithms
- DBSCAN and k-means
- OPTICS, DBSCAN, spectral clustering comparison
- GPU limits for big data
- Stream data setting peculiarities
- Kernel idea, Gaussian RBF
- Sparse vs semi-structured data
- EM algorithm
- MapReduce with word count example