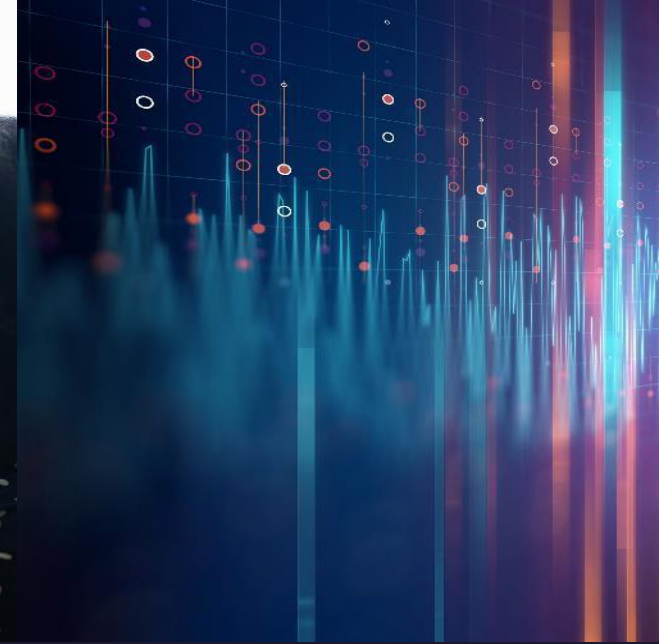


Machine Learning Proiect

Adriana Birluțiu

Daniela-Maria Cristea



Proiectul se va realiza individual!

Aplicația trebuie să fie implementată în limbajul Python si pentru a avea nota de trecere la examen, aplicatia trebuie sa functioneze.

Aplicatia include:

- Descrierea datelor și a tehnicilor de invatare automata folosite
- Software-ul si pachetele folosite
- Rezultate experimentale obtinute

Proiect. Informații generale

Agenda

Când primești un proiect ML (problema expusă)

Explorarea datelor

Procesarea datelor (transformarea tuturor
valorilor în numere)

Modele/Tipuri de învățare/

Măsurarea performanțelor (metrici)

Testare

Hyperparameters



Topic 1.

Când primești un proiect

ML (problema expusă)

- ce încerc să prezic
- ce caracteristici ar trebui să folosesc?
- Utilizați date brute!
- Utilizați date obiective (scală clară).
- Chiar am nevoie de ML?!

Topic 2

• Explorarea datelor

- Statistici descriptive : min, max, percentile, numărători, frecvențe, topX, medie, std
- Descriptive plots
- Boxploturi
- Histograme
- Grafice (Line plots)
- Diagrame (Scatter plots)
- Dendograme (eliminarea coloanelor duplicate)

- Împartirea setului de date Antrenare-Validare-Testare
- Gestionarea valorilor lipsă (np.NaN, np.inf, None)
- drop
- Imputation (mean, iterative, ..)
- Guardian values
- Feature encoding
- LabelEncoding (.., "a" -> 4, "b" -> 5...)
- One-hot-encoding ("a" -> 0001000, "b" -> 0000100)
- Normalizați datele continue
- StandardScaler / MinMaxScaler
- Remove the outliers
- Prin calculul $1.5 * IQR$ (boxplot method)
- Prin calculul scorului Z-score ($z = (x - \text{mean}) / \text{std}$) < -3 , > 3
- Eliminați funcțiile redundante (outliers prin corelarea caracteristicilor)
- Pearson / Spearman
- Balancing the dataset
- Undersampling / oversampling

Topic 3. Procesarea datelor

(transformarea tuturor valorilor în numere)

Topic 4

• Tipuri de învățare/Modele

LinearRegression

LogisticRegression

NaiveBayes, k-NN

DecisionTrees, RandomForests, GradientBoosting

Ensambling methods: Boosting, Model averaging

Topic 5. Modele. Măsurarea performanțelor (metrici)

- utilizați unele valori pentru a măsura setul de date de validare
- modele fictive pentru validarea setului de date
- utilizați o metodologie consecventă pentru compararea modelelor

Metrici de utilizat

- mean squared error, R2 score (MSE)
- F1-score, AuROC, Precision, Recall

(iterate)

începe cu modele liniare

Pipelines / Caracteristici polinomiale

Optimizarea hiperparametrilor (learning-to-learn)

GridSearchCV / RandomSearchCV

Topic 6. Testare

Overfitting vs. Underfitting

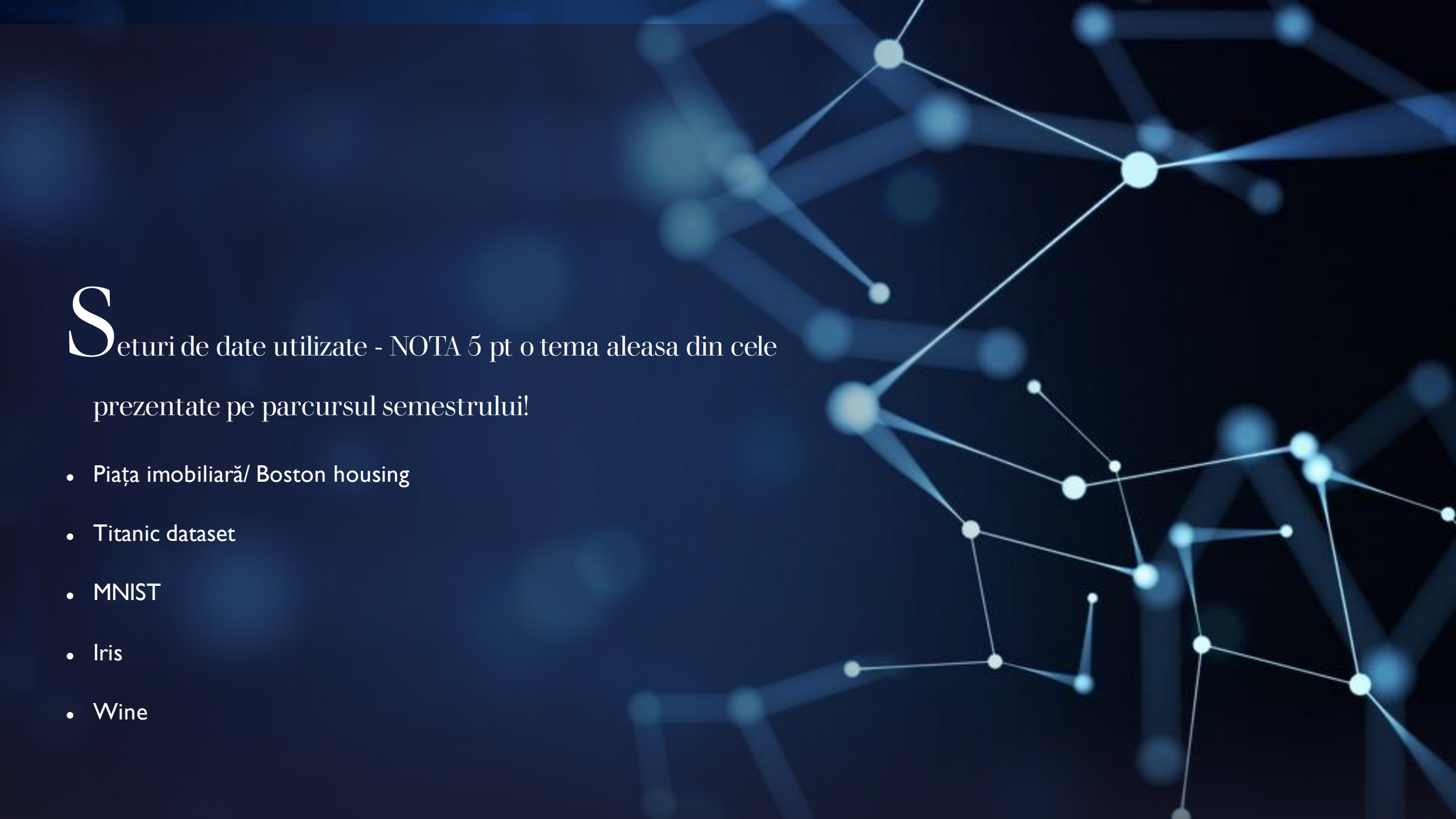
Regularisation (L1, L2)

Confusion matrix

Calculați importanța caracteristicilor

- folosind RandomForest / DecisionTree / model bazat pe arbore (.feature_importances_)
- folosind abordarea de amestecare aleatoare a caracteristicilor (feature shuffling), antrenează un model liniar și uită-te la weights

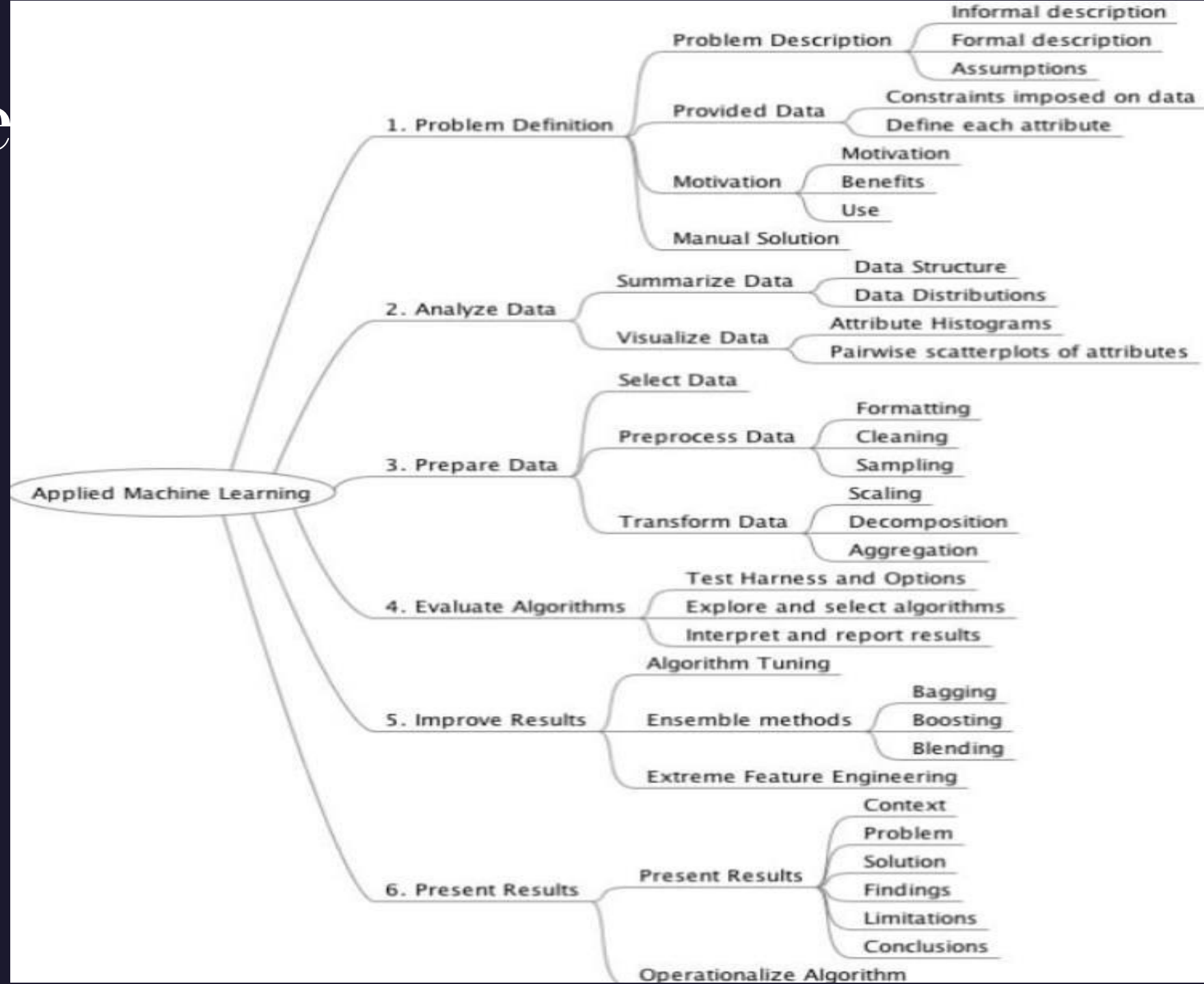
Vizualizați un DecisionTree



Seturi de date utilizate - NOTA 5 pt o tema aleasa din cele prezentate pe parcursul semestrului!

- Piața imobiliară/ Boston housing
- Titanic dataset
- MNIST
- Iris
- Wine

Aplicabilitate





Entry level

Data handling

- ☐ Small datasets
 - Cues
- ☐ Simple preprocessing
- ☐ Image data
- ☐ Audio data
- ☐ Time-series data
- ☐ Text data

Classic Machine Learning

- ☐ Regression
- ☐ Clustering
 - Cues
- ☐ SVMs

Networks

- ☐ Dense Neural Networks
- ☐ Convolutional Neural Networks
- ☐ Recurrent Neural Networks

Theory

- ☐ Mathematical notation
 - Cues
- ☐ Matrix operations
- ☐ Regression
- ☐ Clustering
- ☐ Convolution
- ☐ Simple metrics

General

- ☐ ~1 million parameters
- ☐ Learning the toolset
 - Cues
- ☐ Knowing the docs
- ☐ Data analysis
- ☐ Supervised data
- ☐ Working with metadata files
- ☐ Saving and loading models
- ☐ Callbacks

Intermediate level

Data handling

- ☐ Large datasets
 - Cues
- ☐ Imbalanced datasets
- ☐ Complex datasets
- ☐ Augmentations
- ☐ Normalization
- ☐ Generators
- ☐ Data collection
- ☐ Custom pipelines

Custom projects

- ☐ Custom image project
- ☐ Custom audio project
- ☐ Custom time-series project
- ☐ Custom text project

Networks

- ☐ Large networks
- ☐ Advanced layers:
 - Cues
- ☐ Custom layers
- ☐ Language models:
 - Cues
- ☐ Generative networks
 - Cues
- ☐ Siamese Networks

Training

- ☐ Transfer learning
- ☐ Fine-tuning
- ☐ Custom embeddings
- ☐ Custom callbacks
- ☐ Data-parallel training
- ☐ Multi-GPU training
- ☐ Custom training loops
- ☐ Training in the cloud
 - Cues
- ☐ TPU training

Advanced level

Data handling

- ☐ Huge datasets
 - Cues
- ☐ Multi-modal datasets
- ☐ Distributed pipelines

Custom projects

- ☐ Custom generative project

Training

- ☐ Custom tracking
- ☐ Learning rate scheduling
 - Cues
- ☐ Custom distributed training
- ☐ Mixed-precision training
- ☐ Model-parallel training
- ☐ Multi-worker training
- ☐ Multi-TPU training

Theory

- ☐ Advanced optimizers
 - Cues
- ☐ Reinforcement learning
 - Cues

General

- ☐ ~billion parameters
- ☐ Teamwork
- ☐ Efficient code
- ☐ Model deployment
 - Cues
- ☐ Reinforcement learning
- ☐ Custom inference
- ☐ Reading papers
- ☐ Staying up-to-date
 - Cues

Expert level

Data handling

- ☐ on-GPU pipelines

Theory

- ☐ Quantum deep learning
 - Cues
- ☐ Graph neural networks
- ☐ Evolutionary algorithms
- ☐ Beyond computer science
 - Cues
- ☐ Open-endedness

General

- ☐ ~trillion parameters
- ☐ Research
- ☐ Understanding papers
- ☐ Implementing papers
- ☐ Teaching
- ☐ Contributing to society at large
 - Cues
- ☐ (Degree)



Sugestii pentru proiect

Linkuri utile:

- <http://cs229.stanford.edu/projects2013.html>
- <http://archive.ics.uci.edu/ml/>
- www.kaggle.com