

Welcome to "How to win a data science competition"

Competition mechanics

Recap of main ML algorithms

Software/Hardware requirements

- ✓

Video: Software/Hardware Requirements
5 min
- ✓

Practice Quiz: Software/Hardware
3 questions
- ✓

Lab: Pandas basics
- ✓

Programming Assignment: Pandas basics
3h
- ✓

Quiz: Graded Soft/Hard Quiz
4 questions
- ✓

Reading: Explanation for quiz questions
10 min
- ✓

Reading: Additional Material and Links
10 min

Feature preprocessing and generation with respect to models

Feature extraction from text and images

Final project

Survey

Explanation for quiz questions

Question 1

Which library provides the most convenient way to perform matrix multiplication?

Correct answers:

- Numpy. It's a linear algebra library.

Incorrect answers:

- Pandas. This one is about data reading/processing, not algebra.
- SkLearn. This one is about ML algorithms.
- XGBoost. This one is about GBDT.

Question 2

Which libraries contain implementations of linear models?

Correct answers:

- SkLearn. This one is about ML algorithms including linear models

Incorrect answers:

- Pandas. This one is about data reading/processing.
- Numpy. This one is about linear algebra. It can be used for implementing linear model, but this is not working out-of-box.
- Matplotlib. This one is about visualizations.
- tsne. This one is about dimensionality reduction.

Question 3

Which library (or libraries) are used to train a neural network?

Correct answers:

- PyTorch. This one is about neural networks, so yes, this is the right answer.
- Keras. This one is about neural networks, so yes, this is the right answer.
- TensorFlow. This one is about neural networks, so yes, this is the right answer.

Incorrect answers:

- Numpy. This one is about linear algebra. It can be used for implementing NN, but you should hardcode all gradient calculations.
- Maptlotlib. This one is about data visualization.
- T-SNE. This one is about dimensionality reduction

Question 4

Select the correct statements about the RandomForest and GBDT models.

Correct answers:

- Trees in RandomForest can be constructed in parallel (that is how RandomForest from sklearn makes use of all your cores). Right, since each tree is independent from other trees.
- In GBDT each new tree is built to improve the previous trees. True, the idea of boosting is to correct errors of previously learned models..

Incorrect answers:

- Trees in GBDT can be constructed in parallel (that is how XGBoost makes use of all your cores). No, we need to build trees in sequential manner. In XGBoost multiple cores are used to build single tree.
- In RandomForest each new tree is built to improve the previous trees. No, every tree is independent.