Data analysis theoretical questions and tasks for final exam

Higher School of Economics, Computer Science faculty

Module 4, 2018

Exam rules

You will be given with one random questions from the *detailed questions* list. You will have 15 minutes to prepare your answer using ANY materials. During the answer and further on you CANNOT you any materials and other sources of information. During your answer you also will be asked one or several questions from the *short questions* list and *theoretical minimum* list. To get appropriate grade you have to answer questions from *theoretical minimum* list. Answers and discussion can be conducted in Russian.

Note 1: When describing methods in *detailed questions* list you need to state the problem being solved, the method itself, its advantages/disadvantages. *Short questions* do not require deep analysis.

Note 2: Questions are grouped in topics. You need to answer particular questions.

1 Detailed questions

Dimensionality reduction and feature selection

1. Feature selection. Wrapper and embedded feature selection methods. Recurrent feature elimination. Decision tree feature importances.
2. Principal Component Analysis. Two definitions and their equivalence. Construction of components.
3. Principal Component Analysis. Its connection to Singular Value Decomposition. How to select the number of components
4. Multidimensional scaling and T-SNE algorithm. Their differences. Description of T-SNE algorithm. Possible pitfalls.

Ensemble methods

1. Ensemble learning. Use cases. Standard integration schemes. Blending and Stacking.
2. Ensemble learning. Use cases. Sampling schemes and random forests.
3. Ensemble learning. Additive models and adaboost algorithm. Formulae derivations.
4. Ensemble learning. Gradient boosting algorithm. Modification for trees. Partial dependency plot.

Neural networks

1. Definition of feed-forward neural network. Structure. Activation functions. Pitfalls of NN learning, ways to solve them.
2. Learning of neural-networks. Back-propagation algorithm. Regularization techniques.
3. Definition of Convolution (filter). Convolutional NN. Dropout layes. Key specs and structure.

Clustering

1. General idea behind clustering. K-means algorithm. Key factors. Its connection with EM-algorithm for Gaussian Mixtures.
2. General idea behind clustering. Agglomerative clustering. Cluster quality evaluation.
3. General idea behind clustering. DBSCAN. Cluster quality evaluation.
4. General idea behind clustering. Gaussian Mixture Model and EM-algorithm. Cluster quality evaluation.

Recommender system

1. RecSys idea and challenges. User-based collaborative filtering. Quality evaluation.
2. RecSys idea and challenges. item-based collaborative filtering. Quality evaluation.
3. RecSys idea and challenges. Latent Factor Model. Quality evaluation.
4. RecSys idea and challenges. SVD based algorithm. Quality evaluation.

2 Short Questions

2.1 Dimensionality reduction and feature selection

1. Definition of correlation and mutual information. Intuition behind them.
2. Recurrent feature elimination.
3. How does decision tree feature importance is calculated?
4. Definition of PCA. How to perform transformations?
5. Definition of SVD. Its connection to PCA.
6. Idea behind T-SNE algorithm.

2.2 Neural Networks

1. Definition of multi-layer perception. Possible activation functions.
2. Idea behind back-propagation algorithm.
3. Why conv-NN are more preferable to simple NN for image analysis?

Ensemble methods

1. What is Bagging?
2. Describe Random Forest algorithm.
3. Describe boosting. How it differs from bagging and random forest?
4. What is blending and stacking?

2.3 Clustering

1. Agglomerative clustering. Possible distance between clusters.
2. K-Means. Possible initializations of centroids.
3. K-Means. Ways to estimate number of clusters.
4. DBSCAN. Types of points.
5. DBSCAN. Pitfalls of the method.
6. Cluster quality and validity measures.

Recommender system

1. RecSys idea and challenges.
2. Baseline predictions. Motivation
3. User-based collaborative filtering.
4. Items-based collaborative filtering.
5. Ways to calculate similarity measures for collaborative filtering.
6. Use of SVD in recsys domain.
7. Idea behind latent variable approach.

Theoretical minimum

1. Discriminant functions. Write out discriminant functions for multiclass linear classifier and K-NN.
2. Describe model evaluation with train/test sets, cross validation and leave-one-out techniques. Over-fitting and under-fitting.
3. What is one-hot-encoding? Give feature normalization methods. Why all these feature transformations are important?
4. Give definition discriminant functions. Discriminant function for K-NN, linear models and decision tree?
5. L1 and L2 regularizations. Reasons to use them.
6. What is multicollinearity. What is dummy variable trap?
7. Give definition of gradient descent and stochastic gradient descent. Motivation for stochastic gradient.
8. Definition of confusion matrix in binary classification. How to calculate precision and recall. What is F-measure?
9. Definition of ROC curve and AUC. Motivation for them.
10. How can you measure model quality for regression task? Write down the definition of RMSE, MAE, MAPE, RMSLE.
11. Give intuition behind SVM. Write optimization problem for linearly separable SVM.
12. Kernel trick. How it works for K-NN and for SVM?
13. Multiclass classification with binary classifiers: one-vs-all and one-vs-one schemes.
14. What is feature selection? Why is it useful? Describe types of feature selection procedures.
15. What is dimensionality reduction. Why it should be used?
16. Bias-Variance decomposition of error.
17. Describe idea behind PCA algorithm.
18. Definition of multilayer perceptron.
19. Describe idea behind backpropagation algorithm.
20. Describe idea of convolution filter in conv-NN.
21. What is the general idea behind clustering. Why is it unsupervised learning task?
22. How to measure quality of clustering
23. Describe K-means algorithm.
24. Describe Agglomerative clustering.
25. Cluster quality and validity measures.
26. What is ensemble learning? Bagging.
27. What is ensemble learning? Random Forest.
28. What is ensemble learning? Gradient Boosting.
29. Idea of Collaborative Filtering.