Data analysis theoretical questions and tasks for colloquium

Higher School of Economics, Computer Science faculty

Module 3, 2020

The rules

You will be given two (almost) random questions from the list. You will have 15 minutes to prepare your answer without using any materials, except A4, prepared beforehand and hand-written personally by you (from two sides). You will be asked additional questions during your answer, during which you may also use your A4. Answers and discussion can be conducted in Russian.

Colloquium questions

1. Expected and empirical risk minimization. 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. How expected train/test loss changes with train set size and complexity of the model?
3. What is one-hot-encoding? Give feature normalization methods. Why all these feature transformations are important?
4. How do K-NN and decision tree methods change when they are applied in classification and in regression context?
5. Give definition of the following methods and explain why or why not feature scaling can affect their performance: linear regression estimated with least squares minimization, linear regression with regularization, CART decision trees, logistic regression without regularization.
6. Explain idea of weighted K-NN. Give examples of weights. Will feature scaling affect predictions of K-NN?
7. Give definition discriminant functions and margin for classification. What is its intuition?
8. Definition of decision tree. Definition of impurity function. Examples of impurity function. Splitting rule selection for CART trees.
9. Propose stopping rules for setting tree node to leaf node based on:
   - class distribution
   - number of samples in the leaf
   - impurity function
   - change in impurity function from splitting this node
10. Linear regression estimated with ordinary least squares - derive its solution. RIDGE and LASSO regularizations - write them out. Which of them selects features and why?
11. Definition of linear classifier for two and multiple classes.
12. Show that for linearly dependent features vector of coefficients of linear regression is not uniquely defined. What is dummy variable trap? Given two linearly dependent features which of them will be eliminated by LASSO?
13. How to estimate parameters of linear classifier? Write out the optimization task for different loss functions. Compare qualitatively typical loss functions.
14. Optimization criteria to find weight of binary linear classifier with L_1 and L_2 regularization on weights. Which of regularizer has feature selection capability? Why?
15. Multiclass logistic regression. How to find weights? What is Softmax function?

16. Give definition of gradient descent and stochastic gradient descent. Motivation for stochastic gradient. Write out pseudo code for both methods for one of the following losses:
    - $L(M) = [-M]_+$.
    - $L(M) = [1 - M]_+$.
    - $L(M) = \log(1 + e^{-M})$.
17. Definition of confusion matrix. How to calculate error rate, accuracy with it? What is the relationship between TPR, FPR, FNR and TNR?
18. Definition of confusion matrix in binary classification. How to calculate precision and recall. What is their intuition. What is F-measure?
19. Definition of ROC curve and AUC. Motivation for them. ROC for random classifier. Show that ROC-AUC defines the probability of random "positive" object having higher score than random "negative" object
20. Give constructive algorithm to calculate ROC curve. ROC curve of inverted binary classifier.
21. Describe different approaches to multi-classification measures aggregation. Show that micro Precision = micro Recall = Accuracy
22. How can you measure model quality for regression task? Write down the definition of RMSE, MAE, MAPE, RMAE, RMLSE. Describe their meanings, pros and cons.
23. Derive distance from point x to linear hyperplane $\{x : w^\top x + w_0 = 0\}$. What vector is orthogonal to linear hyperplane?.
24. Which of the following methods under certain conditions have a linear decision boundary? Give examples of such conditions
    (a) K-NN
    (b) Decision tree
    (c) Logistic regression
25. Give definition of the following models and answer how do the following parameters affect complexity of the model?

model	parameter
K-NN	K
K-NN	weight kernel
linear regression with L_1 regularization	Regularization coefficient
decision tree	min samples count in leaf
elastic net	Regularization coefficient and regularization type balancing
decision tree	max depth and minimum gain

26. Provide example, when adding a new feature improves accuracy of the model
27. Describe groups of feature selection approaches, their pros and cons
28. What is PCA? Provide derivation of principal components. Propose rules to select number of components.