Data analysis theoretical questions and tasks for colloquium

Higher School of Economics, Computer Science faculty

Module 3, 2018

Exam rules

You will be given two (almost) random questions from the list. You will have 20 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. What is curse of dimensionality? Why it applies to K-NN and does not apply to linear models?
6. Give definition of the following methods and explain why or why not feature scaling can affect their predictions: linear regression estimated with least squares minimization, linear regression without regularization, CART decision trees, logistic regression without regularization, SVM.
7. Explain idea of weighted K-NN. Give examples of weights. Will feature scaling affect predictions of K-NN?
8. Give definition discriminant functions and multiclass margin. What is its intuition? How margin simplifies for binary classification?
9. Definition of decision tree. Definition of impurity function. Examples of impurity function. Splitting rule selection for CART trees.
10. 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
11. Definition of discriminant functions. Are they uniquely defined? Definition of margin for multiclass and binary class cases, its intuition.
12. Parzen window method. Nadaraya-Watson regression, its derivation. Under what selection of $K(\cdot)$ and $h(\cdot)$ these methods reduce to K-NN?
13. Linear regression estimated with ordinary least squares - derive its solution. RIDGE and LASSO regularizations - write them out. Which of them selects features and why?
14. Definition of linear classifier for two and multiple classes.
15. Show that for linearly dependent features vector of coefficients of linear regression is not uniquely defined. What is dummy variable trap?

16. How to estimate parameters of linear classifier? Write out the optimization task for different loss functions. Compare qualitatively typical loss functions.

17. 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?

18. Multiclass logistic regression. How to find weights? What is Softmax function?

19. 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:

    - $L(M) = [-M]_+$.
    - $L(M) = [1 - M]_+$.
    - $L(M) = \log(1 + e^{-M})$.

20. Definition of confusion matrix. How to calculate error rate, accuracy with it?

21. Definition of confusion matrix in binary classification. How to calculate precision and recall. What is their intuition. What is F-measure?

22. Definition of ROC curve and AUC. Motivation for them. ROC for random classifier.

23. Give constructive algorithm to calculate ROC curve. ROC curve of inverted binary classifier.

24. How can you measure model quality for regression task? Write down the definition of RMSE, MAE, MAPE, RMAE, RLMSE. Describe their meanings, pros and cons.

25. Derive distance from point x to linear hyperplane $\{x : w^\top x + w0 = 0\}$. What vector is orthogonal to linear hyperplane?.

26. Give intuition behind SVM. Write optimization problem for linearly separable SVM and linearly non-separable SVM.

27. Write out optimization problem for linearly non-separable SVM. What linear classification loss function and regularization corresponds to this method?

28. Kernel trick. How it works for K-NN and for SVM?

29. How parameters of polynomial and RBF kernels and parameter C control complexity (flexibility) of kernel SVM? Why?

30. What is a Mercer kernel? Mercer theorem.

31. Multiclass classification with binary classifiers: one-vs-all and one-vs-one schemes. Compare their complexity.

32. Which of the following methods under certain conditions have a linear decision boundary? Give examples of such conditions

    (a) K-NN
    (b) SVM without kernel
    (c) SVM with RBF kernel
    (d) Decision tree
    (e) Logistic regression

33. Consider $\langle x, x' \rangle^M$ and $(\langle x, x' \rangle + 1)^M$ kernels. Why these are Mercer kernels?

34. Linear and constant kernels. Which operations produce new kernels out of existing ones? Link between kernels and norm.

35. Give definition of the following models and answer how do the following parameters a ect complexity of the model?

model	parameter
K-NN	K
linear regression with L_1 regularization	Regularization coefficient
decision tree	min samples count in leaf

36. Give definition of the following models and answer how do the following parameters a ect complexity of the model?

model	parameter
decision tree	max depth
decision tree	minimum purity gain

37. Give definition of the following models and answer how do the following parameters a ect complexity of the model?

model	parameter
elastic net	Regularization coefficient and regularization type balancing
decision tree	Feature count

38. Give definition of the following models and answer how do the following parameters a ect complexity of the model?

model	parameter
SVM	C
SVM with RBF kernel	γ

39. What is feature selection? Why is it useful? Describe types of feature selection procedures.

40. Provide example, when adding a new feature improves accuracy of the model.