Thesis changes log

Name of Candidate: Alexander Fonarev

PhD Program: Computational and Data Science and Engineering

Title of Thesis: Matrix Factorization Methods for Training Embeddings in Selected Machine Learning Problems

Supervisor: Prof. Ivan Oseledets

Chair of PhD defense Jury Prof. Andrzej Cichocki ***Email**: a.cichocki@skoltech.ru*

Date of Thesis Defense: 19 September 2018

The thesis document includes the following changes in answer to the external review process.

Reviewer: Andre Uschmajew

1. Section 1: on page 21 it is stated that the embedding concept was introduced, but in my opinion this introduction was a bit too vague, no mathematical notation was given. While I see some arguments for this sort of conceptual introduction, some more concrete statements good be helpful. For instance, in the three word example on page 14, onw could give an example of an actual embedding.

   Answer: Thank you for your comment. I have added the formal definition of embeddings.

2. page 24, middle: columns of U, V are ’dominant’ eigenvectors corresponding to the singluar values.

   Answer: Thank you for your comment. I have corrected this.

3. page 25, top: does this complexity refer to full SVD? Since truncated SVD has been considered on page 25 ...

   Answer: Thank you for your comment. I have corrected this.

4. page 25: could you specify where in reference [109] to find result that Md is a smooth embedded manifold

   Answer: Thank you for your comment. I have corrected this part of the text.

5. page 25: ’The key difference between ... is the non-requirememt of explicit factors ...’

   Answer: Thank you for your comment. I have corrected this.

6. page 26: ’standard Riemannian gradient method’ ... I think this part could be a little expanded. The word standard should maybe replaced by simplest. In fact the Riemannian gradient method depends

on the choice of t6he Riemannian metric, and the described method is obtained by choosing the trivial metric from ambient space.

Answer: Thank you for your comment. I have replaced the word “standard” with the word “simplest”. Since there was not a goal to dive deeply in this topic, I kept this as ot was.

7. page 26: reference [116] is for a sepcific cost function and under restricted isometry prop- erties that perhaps do not hold here. I think the citations should be more precise. There are many other references containing partial convergence results for Riemannian methods on low-rank matrices.

   Answer: Thank you for your comment. I have reformulated this part of the text to be correct

8. page 28: is it needed that Si in (28) is diagonal? The algorithm does not seem to suggest this.

   Answer: Thank you for your comment. I have corrected this.

9. page 28: what is Si+1 in the first step? Is it needed? Please explain the notation QR.

   Answer: QR means QR-decomposition. Indeed, S_{i+1} in the first line is not obligatory, but it is kept to make the algorithm easier to read (hope, it works)

10. page 29: interception ! intersection, what is Q in (22)?

    Answer: Thank you for your comment. I have corrected this.

11. page 30, middle: ’searching for a row set’ ... do you mena column set? What is Q in (23)?

    Answer: Thank you for your comment. No, I meant a row set. However, I mixed the notation, not it is corrected

12. page 30: ’it is an effective approach’ ... it is an effective approach for what purpose?

    Answer: Thank you for your comment. I meant performance — I have rewritten this part

13. page 31, line -4: should end with full stop and start new sentence ’We recommend ...’

    Answer: Thank you for your comment. I have corrected this.

14. page 32: a lot of embedding methods.

    Answer: Thank you for your comment. I have corrected this.

15. section 3.2.2: there is not much information provided here about what decision tree models actually are. Since random forests are used later, one should expand on this.

    Answer: Thank you for your comment. I have explained random forests in more details and added more references on decision trees.

16. page 35 top: The number of layers, the number of neurons ...

    Answer: Thank you for your comment. I have corrected this.

17. page 35, first sentence of sec. 2.3.1: features of objects are in R ... does it mean in such a case an embedding is not needed? I assume one would still do the factorization for dimensionality reduction. It maybe would help to explicitly state in the introduction on embddedings that the problem is relevant even if objects are already given as vectors.

    Answer: Thank you for your comment. The need to train embeddings for objects described by real values is a different problem (which is not always needed to be solved). That is why, I focus on objects described by categorical features – there I see a big potential.

18. page 36: effcient transformation ... do you mean meaningful transformation; same sen- tence: into real valued vectors

    Answer: Thank you for your comment. I have corrected this.

19. page 36: section on websearch is so short, one need at least a reference.

    Answer: Thank you for your comment. I think that deeper dive into the websearch might unfocus the flow of the text, so I kept the current version.

20. page 37, first sentence: this work ...

    Answer: Thank you for your comment. I have corrected this.

21. page 37, bottom: decision trees are again introduced, but without much information how they work.

    Answer: Thank you for your comment. As I mentioned above, I detailed description of random forests and added references on decision trees.

22. page 38: A very popular method.

    Answer: Thank you for your comment. I have corrected this.

23. page 39: since dummy encoding is just mentioned but not discussed, I would remove this word from title of subsection

    Answer: Thank you for your comment. I have corrected the wording

24. page 39: 'the following class will be predicted' what do you mean by class?

    Answer: Thank you for your comment. I meant one of two classes that are being used in the binary classification problem

25. page 39: what is I in (29)?

    Answer: Thank you for your comment. I have corrected the notation

26. page 41, bottome: here it is not yet clear why c \in V_C and what is V_C?

    Answer: Thank you for your comment. As mentioned in the text, V_C is a set of context words. I have corrected the wording where it could be possible.

27. page 42: Is D the set of all word-context pairs or a set of some ...

    Answer: Thank you for your comment. Actually, D is multiset because it is important to know a number of seeing each pair within a corpus

28. page 42: distribution

    Answer: Thank you for your comment. I have corrected the wording

29. page 42: 'd is a hyperparameter' ... I can't see d, where has d been introduced?

    Answer: Thank you for your comment. I have detalized this part in the introduction section

30. page 43: what is k in (36)?

    Answer: Thank you for your comment. I have corrected the notation

31. page 43: why is the method called skip-Gram? Is there a Gram matrix invovled?

Answer: Thank you for your comment. No, it is not related to Gram matrix — it comes from “n-gram” widely used in computational linguistics

32. page 43: It would be good to mention that finding w and c correspinds to finding the maps W and C in (31).

    Answer: Thank you for your comment. I have corrected this

33. page 43: notation lw,c in (38) is not good because lwc has been used in (36).

    Answer: Thank you for your comment. I have corrected the notation

34. page 44, first sentence of 2.4.4: which factorization task? Please write down an explicit matrix problem at the end of 2.4.3 that one has to solve.

    Answer: Thank you for your comment. I have corrected the wording

35. page 49: sentence starting with ’Among them ...’ what do you mean? Among Greedy methods?

    Answer: Thank you for your comment. I have corrected the wording to make it clear

36. page 50, sectionj 2.5.8. Is this still on rating elicitation methods?

    Answer: Thank you for your comment. I have corrected the wording to make it clear

37. page 53: call this formalization a framework; we also develop...

    Answer: Thank you for your comment. I have corrected the wording

38. page 59: X \in N^{n*m} vs. all features are categorical. Please explain the relation. In particular, in 2.3.3 I got the impression enumerating categorical data is not a good idea.

    Answer: Thank you for your comment. I used a bit confusing notation — now it should be clear

39. page 61, first sentence: why can it be approximated with low-rank d? Do you have to factorize a matrix G for every entry of Z? Does (47) imply d_qj1 and d_qj2 for all j1,j2? How to ensure it, who you choose d?

    Answer: Thank you for your comment. G is factorized for each pair of features in the input dataset. Selection of d was clarified in the introduction section.

40. page 66, sec. 4.3.3: the proposed method is most effective, but also most expensive?

    Answer: Thank you for your comment. I have corrected the wording to make it clear that the SVD-based method, which works fast, is also an improtant contribution

41. page 67: Random Forest predictor is mentioned, but nowhere explained.

    Answer: Thank you for your comment. I have explained this in the introduction

42. page 68, tabel 3: is this the precentage of correct classification?

    Answer: Thank you for your comment. As mentioned in the text, this is the AUC-ROC performance measure, which is standard for binary classificaiton problems

43. page 69, section 4.4: ’ ... methods that can handle categorical features out of the box ...’ your method can also handle them. What exactly do you mean? Do these methods work very differently (not using embeddings)?

Answer: Thank you for your comment. By this methods, I meant real-valued methods. But my method can be used only together with a real-valued binary classifier (of course, if the goal is to solve the binary classification problem)

44. page 72/73, rank-d constraint: is there a smart way to automatically choose rank in Riemannian optimization? Maybe here you good at least give an impression on the expected size of d.

    Answer: Thank you for your comment. The embedding dimensionality is fixed by the standard problem setting.

45. page 73, eq (55): please repeat definition of fw,c

    Answer: Thank you for your comment. I have done this

46. page 77: do you want to indicate that alternating optimization works equally well as projector splitting? Is it at least more expensive?

    Answer: Thank you for your comment. No, Aternating approaches (as well as SGD) work worse than projector-splitting

47. page 85: Algorithm 2, please refere to Sec. 2.5.6

    Answer: Thank you for your comment. I have done this

48. page 86: Are you using the notation from Section 2 here again?

    Answer: Yes, I have made it more clear now

49. What is f? Do you mean d?

    Answer: Thank you for your comment. Yes, you are right, I have corrected this

50. page 87, bottom: need boldface k in Q[k, :]T

    Answer: Thank you for your comment. I have corrected this

51. page 94, eq (86): C must be boldface

    Answer: Thank you for your comment. I have corrected this

52. page 97: 'It means that we are interested in the small values of C ...' – I think you should emphasize that we are interested in small values of C outside the set k, since, as far as I see, in k the columns of C are unite vectors, and the corresponding columns in the error terms in (90) cancel.

    Answer: Thank you for your comment. C contains all coefficients, including ones that correspond to k.

53. page98, Thm1: Should it be S=Q[k,:]^T? In the proof, what is Lemma 1?

    Answer: Thank you for your comment. I have corrected the text and added reference to the Lemma 1

54. page 98, bottom: 'Theorem 1 deomstrates ..' – it is a bit misleading. You should add ' at least for the ideal case that one can starts from a dominant submatrix ...'

    Answer: Thank you for your comment. I have corrected the wording

55. page 100: We use ...

    Answer: Thank you for your comment. I have corrected the text

56. page 101: 'For every size of the seed set we used the rank that gives best performance...' What does it

mean? How did you select ranks?

Answer: Thank you for your comment. I have selected them using additional train-test split of the data (it is described in Section 6.6)

57. page 101, bottom: Figure 6.7 ... there is no such figure, same on page 103.

    Answer: Thank you for your comment. I have corrected the formatting

58. page 105: typo: Conclusion

    Answer: Thank you for your comment. I have corrected this

59. page 105, item 1: ’Unfortunately, such techniques are usually underestimated by the com- munity’.. I suggest using a less strong and more positive phrase, like ’they have not been studied as intensively ....’ or just delete it.

    Answer: Thank you for your comment. I have reformulated this

60. A recurring topic suitable for discussion is the choice of ranks and wheter there is an adaptive way of learning it. Another set of questions is on convergence of methods under suitable assumptions.

    Answer: Thank you for your comment. I have made it more clear in Chapter 1

Reviewer: Alexander Tuzhilin

61. I have corrected several minor mistakes in English and removed wrong statements about the lack of space in the thesis.

Reviewer: Dmitry Ignatov

62. As it is illustrated on Figure 2, – in Fugure

    Answer: Thank you for your comment. Corrected

63. where the embeddings are be used. – are being used.

    Answer: Thank you for your comment. Corrected

64. close to each other in the representation space and, that is why

    Answer: Thank you for your comment. Corrected

65. a separate dataset which contains level of similarity between some

    Answer: Thank you for your comment. Corrected

66. of objects could be used as a benchmark to measure an embedding

    Answer: Thank you for your comment. Corrected

67. algorithm performance. – this piece should be definitely rewritten.

    Answer: Thank you for your comment. Corrected

68. There are a large number of papers – the papers are in the focus

    Answer: Thank you for your comment. Corrected

69. [106] explores using word embeddings

Answer: Thank you for your comment. Corrected

70. [58] explores items embeddings in recommender embeddings, – the

Answer: Thank you for your comment. Corrected

71. last embeddings term should be systems, is not it?

Answer: Thank you for your comment. Corrected

72. when initial matrix A – when the original (or input) matrix A

Answer: Thank you for your comment. Corrected

73. using alternating approache

Answer: Thank you for your comment. Corrected

74. the gradient step an the retraction

Answer: Thank you for your comment. Corrected

75. After this, the algorithm – After that, or After this step.

Answer: Thank you for your comment. Corrected

76. Decision tree-based methods vs Decision tree based methods

Answer: Thank you for your comment. Corrected

77. Although single decision trees [15] usually do not show a good

Answer: Thank you for your comment. Corrected

78. performance, and ensembling decision trees often provide state-ofthe-art

Answer: Thank you for your comment. Corrected

79. results. – Although single decision trees [15] usually do

Answer: Thank you for your comment. Corrected

80. not show a good performance, while ensembling decision trees often

Answer: Thank you for your comment. Corrected

81. provide state-of-the-art results.

Answer: Thank you for your comment. Corrected

82. In case of the binary classification – which one do you mean,

Answer: Thank you for your comment. Corrected

83. the is a void identifier here.

Answer: Thank you for your comment. Corrected

84. so-called log-loss – the is missing

Answer: Thank you for your comment. Corrected

85. A neural network decision-making process consists of several

Answer: Thank you for your comment. Corrected

86. steps (layers), and each of these steps takes outputs from the

Answer: Thank you for your comment. Corrected

87. previous step (neurons), applies a non-linear transformation to

Answer: Thank you for your comment. Corrected

88. them and multiplies them by some weight matrix in order to generate

Answer: Thank you for your comment. Corrected

89. outputs. – The sentence should be rewritten since applies actually

Answer: Thank you for your comment. Corrected

90. does not related to a proper noun word.

Answer: Thank you for your comment. Corrected

91. We get new transformed matrix Z – a is missing.

Answer: Thank you for your comment. Corrected

92. Step 1. Objects are texts.described

Answer: Thank you for your comment. Corrected

93. if matrix some loss-function ρ(·, ·) is used?

Answer: Thank you for your comment. Corrected

94. All results are statistically significantly different, which is

Answer: Thank you for your comment. Corrected

95. proved by Mann-Whitney U test. – All the results…

Answer: Thank you for your comment. Corrected

96. Figure 14. Coverage or diversity. – Coverage and diversity

Answer: Thank you for your comment. Corrected

97. There are three main reasons to train embeddings – are those embeddings supervised machine learning techniques?

Answer: Thank you for your comment. No, my approach is unsupervised since it does not use the target label of objects

98. the interception of columns C and rows R – the term interception was not introduced

Answer: Thank you for your comment. Corrected

99. Each column of C contains the coefficients of the representation of a row in U via the vectors from S. – where this U has been introduced?

Answer: Thank you for your comment. Corrected the notation

100. These features are called categorical, nominal or factor. – simply factors?

Answer: Thank you for your comment. Corrected the wording

101. It means that many widely used and very powerful real-valued techniques, e.g., Random

Forests [14], cannot be efficiently applied to these tasks. – It depends on input datasets and basic decision tree inducers within the ensemble.

Answer: Thank you for your comment. Yes, you are right. But I focus on cases of large number of categorical values — it can not be handled by standard decision trees

102. In equation (30) usually alpha is out the brackets in the nominator, and alpha multiplied by the number of feature values in the denominator.

Answer: Thank you for your comment. Yes, you are right, but I have written this in such way just to simplify the formula

103. Let D be a multiset of all word-context pairs observed in the corpus. – Do you really use a multiset here?

Answer: Thank you for your comment. Yes, since we need to know the number of seeing of each pair of word within the corpus.

104. The obtained factor P in Rn×r – Usually factors are columns or rows of the resulting product matrices.

Answer: Thank you for your comment. I have corrected the wording.

105. The most standard way to feature values co-occurrence frequencies, – may be cooccurrence frequencies of feature values

Answer: Thank you for your comment. I have corrected the wording.

106. in eq. (41) and (45) the universal quantifier is used after the variables it relates.

Answer: Thank you for your comment. I have corrected the notation

107. In eq. (45), should the equality sign be by definition sign?

Answer: Thank you for your comment. Yes it could be done so, but I decision not to use this notation just not to overcomplicate the thesis

108. Similarly, for eq. (47).

Answer: Same

109. we have also experimented – Capital We.

Answer: Thank you for your comment. Corrected

110. i not in k and add it to the seed set: – it is hard to track where an index or vector is used without proper notation.

Answer: Thank you for your comment. Corrected the notation

111. we use the 5-fold cross validation with respect to the set of users in all experiments. – why THE 5-fold?

Answer: Thank you for your comment. It was not motivated by any experiments — I just tried this option and it gave good results. I believ that there is a room for additional improvements