Homework 1, CS 6241 Spring 2019

Instructor: Austin R. Benson
Due Thursday March 7, 2019 at 11:59pm ET on CMS

Assignment

The goal of this assignment is to give you a better understanding of a numerical method for data science and also to provide some basic preparation for the final project.

This homework is designed to be open-ended, and the assignment has two parts. First, you will implement a numerical method related to the material covered in the first 10 lectures.[1] Second, you will use your implementation to analyze a "real-world" dataset of your choosing.[2]

Numerical method

In the first 10 lectures, we will have covered numerical methods for:

1. linear least squares—regression, regularization, sparse;
2. latent factor models and linear dimensionality reduction;
3. randomized numerical linear algebra;
4. tensors and low-rank decompositions; and
5. non-linear dimensionality reduction.

Pick a numerical method from these topics and implement the method. The method could be one we specifically covered in lecture (e.g., the CUR factorization) or a related method that you find on your own. Include a brief description of the numerical method and why it is useful.

Your implementation should not just call a library function of that method. Instead, you should implement the basic computations (for this, you can use libraries). Many of the Jupyter notebooks from class follow this pattern.[3] Figure 1 shows an example based on the alternating least squares method for nonnnegative matrix factorization that we saw in class.

Data analysis

Next, pick a "real-world" (i.e., not synthetic) dataset on which to apply your numerical method. Provide a brief description of the dataset and explain what the numerical method reveals about the data. Provide a couple of qualitative insights.

[1] See the course content at `http://www.cs.cornell.edu/courses/cs6241/2019sp/`.

[2] See the course web site for some data repositories.

```
,
using SparseArrays, DelimitedFiles, Gurobi, JuMP

function ALS_NMF(A::Array{Float64,2}, k::Int64)
    n = size(A)[1]

    function update_R(L::Array{Float64,2})
        model = Model(solver=GurobiSolver(Presolve=0))
        @variable(model, R[1:k,1:n])
        @objective(model, Min, sum((A - L * R).^2))
        @constraint(model, R .>= 0)
        solve(model)
        return getvalue(R)
    end

    function update_L(R::Array{Float64,2})
        model = Model(solver=GurobiSolver(Presolve=0))
        @variable(model, L[1:n, 1:k])
        @objective(model, Min, sum((A - L * R).^2))
        @constraint(model, L .>= 0)
        @constraint(model, sum(L, dims=2) .== 1)
        solve(model)
        return getvalue(L)
    end

    L, R = rand(n, k), rand(k, n)
    for _ = 1:100  # alternating loop
        L = update_L(R)
        R = update_R(L)
    end
    return L, R
end

function main()
    data = readdlm("zacharys-KC.txt", ' ', Int)
    n = 34  # 34 nodes in graph
    A = sparse(data[:,1], data[:,2], 1, n, n)
    A = convert(Matrix{Float64}, max.(A, A'))
    L, R = ALS_NMF(A, 2)
end
```

Figure 1: Code for alternating least squares method for nonnegative matrix factorization that we used in class.

[3] `https://github.com/arbenson/cs6241_2019sp`

Preparation & submission guidelines

Typesetting. All homeworks should be prepared with a proper typesetting tool (namely, LaTeX). Handwritten homeworks will not be accepted.

Code. Part of the assignment involves implementing a numerical method. You need to include your code in your submission, and you can easily do so using the `listings` package. This is how Figure 1 was created. You do not need to include code for the qualitative data analysis component.

Collaboration. You are encouraged to discuss and collaborate on the homework to the extent of exchanging, formulating, and discussing ideas as a group. However, you have to write your own homework submission completely on your own and also understand what you are writing. You must also list your collaborators on your homework.

Academic Integrity. I expect you to maintain academic integrity in the course. For example, follow the collaboration guidelines mentioned above and do not copy someone else's software implementation. Failure to maintain academic integrity will be penalized severely. Plagiarism is a form of academic misconduct, so make sure to provide proper citations. Cornell has a number of guidelines on plagiarism.[4]

[4] `https://plagiarism.arts.cornell.edu/tutorial/index.cfm`

Submission. Your homework should be submitted as a single PDF that includes the following:

1. Your name, along with names of any collaborators (if applicable).
2. A brief description of the numerical method that you are using.
3. Your code (as outlined above).
4. A brief description of data that you used.
5. Qualitative analysis of the dataset obtained via the numerical method.

Submit your PDF on CMS.[5]

[5] `https://cmsx.cs.cornell.edu`