Exercise 6: Particle Swarm Optimization

Due date: *November 27th, 2025 (push to GitHub repo before the lecture)*
General Requests: *Same as before.*
Specific Requests: *Please at least try to consider GitFlow and commit/issue conventions for your repo again!*

6.1 Particle Swarm Optimization (PSO)

We've regarded benchmarks, found optimum parameters for objective functions, investigated high-dimensional regression, and solved path-finding/scheduling problems now. This time, we'll look into a very common real-world problem: feature selection. Using Particle Swarm Optimization (PSO)[1, Chapter 6.7][2, Chapters 11 and 19] this time, we'll try to find a good subset of features for a classification model.

- **data**: we'll use the Covertype Dataset including 581.012 samples in 54 features about terrain and soil with labels of the corresponding 'cover classes' (i.e., types of trees growing there). the goal is to find a **subset** of the features that maximizes classification performance.
- decide on an encoding for candidate solutions [1]
- choose a classifier and performance metric you want (LDA/SVC/Random Forest.. - whatever strikes your fancy)
- **important:** there's a bunch of kaggle codes on this dataset out there (for example this one or this one), so while you absolutely **should** consider things like spread of the data, class counts, etc. like the data scientists you are, you can fetch some inspiration in those scripts to save you some time for this assignment. we all know the model choice is important, but in this case, just wing it a little
- implement a simple PSO as discussed in the lecture, i.e., include particles, velocities, previous / neighbors' (optionally also global) best, inertia weight, cognitive / social terms, etc.
- **important**: if you don't actively **penalize** selecting too many features, PSO will just select all 54[2]. To avoid that, add a penalty term to your cost or fitness function to make less features more attractive. possible approaches (also used here, for example) include linear combinations of target metrics and the penalizing term:

$$\text{fitness}(\vec{x}) = \alpha\,\text{acc}(\vec{x}) + (1-\alpha)\left(1 - \frac{n}{N}\right), \text{ or} \tag{1}$$

$$\text{cost}(\vec{x}) = \alpha\,(1 - \text{acc}(\vec{x})) + (1-\alpha)\frac{n}{N}, \tag{2}$$

  with $\alpha \in (0,1)$ a parameter setting the relative importance of your target metric versus the penalty, $n \leq N$ the number of selected features, and $N = 54$ the total number of features

6.2 Analysis and Reporting

Add a streamlit page on PSO to the documentation you created for the last exercises.
Your final documentation webpage should contain the following sections:

1. Introduction - Overview of the algorithm: basics, strengths/weaknesses, complexity, differences to other algorithms we talked about, etc.
2. Methods - Describe your implementation: pros/cons, relevant parameters/functions, critical design choices, etc.
3. Results - Display your results (results on validation/test sets, number of selected optimum features, type of selected optimum features, etc.); allow different PSO parameter settings
4. Discussion - Analyze your findings: expected versus unexpected results, solution quality, efficiency, limitations, possible improvements, etc.

The documentation doesn't need to be absurdly long, but it should give a comprehensive overview of the topic that you can use later on to study for the exam.

[1] an obvious choice might be a binary vector of length 54 indicating selection (1) and abandonment (0) of an individual feature
[2] That's a known problem for a bunch of other applications as well. When for example trying to find the coefficients for autoregressive models, it's the same thing for model order selection: if you don't explicitly penalize large order, you'll end up using all available time points for your model.

References

[1] A. Eiben and J. Smith, Introduction to Evolutionary Computing. Springer, 2003.

[2] D. Simon, Evolutionary Optimization Algorithms. Wiley, 2013. [Online]. Available: https://research-1ebsco-1com-1195qzf320241.perm.fh-joanneum.at/c/kofjhs/search/details/4sh2uyq6wn?db=nlebk&db=nlabk