Checkpoint #4 - High-Level summary, discussion between the team and the instructor
Team: Zixi Chen, Marcel Torne Villasevil, Zihe Zhang
Instructor: Weiwei Pan
Topic: Risk score learning for COVID-19 contact tracing apps

Inference

- Evaluate whether the model is able to learn the true parameters generated by the same type of model using SGD. What about variance? Sensitivity to initialization? Any relationship between infection rate and variance? Convergence issue - local optima or global optima? Is the solution really good?
- What happens if the model is not exactly equal? (case presented in the paper)
- The model only sees the bluetooth signal as an approximation of the (discretized) distance used to generate the data.
- What's the uncertainty of this model? We might bootstrap and observe the variances for each datapoint.

Other questions

- In the paper, they supposed that the bluetooth signal was perfectly accurate. What happens if we add some noise? How off will the model be with respect to this noise? If the bluetooth attenuation does worse and worse as an alternative measure of distance, when does the model stop working?
- Does bagging size affect the accuracy of the model? Why? What does this imply? What aspect is causing that?
- In the paper, the authors compared their model to the Swiss model; however, this was an unfair comparison since the Swiss model was not using one of the inputs. What would happen to this model if we remove this same input? Will the accuracy become the same as or even worse than the Swiss one? If so, what conclusions can we extract about the importance of this input?
- What are the implications of using simulated data? How bad can it be if deployed in the real world? We can kind of answer this question when analyzing the propagated error in the bluetooth signal for example. We can also point out that we can't infer the prior infection rate and discuss how this might affect model prediction.

Tips

- We don't have to replicate everything as we are answering our own questions.
- Check out why the model is running slow. Which part can be modified to boost the speed? Could it be that we have many for-loops in calculating the objective function, vectorization might boost the speed? We can also start with small and simple datasets.
- Ways to check whether or not inference on the model works:
    - Generate data from the actual model, do inference and compare parameter estimates to true parameters
- A realistic probability of exposure can be really low, and using it to generate a pool of exposure events can be highly inefficient since most of the generated cases are negative.
    - Weiwei's suggestion: the rate does not need to be realistic, just turn up the probability