Big Data Research Lab
Computer Science Department
Colorado State University

PREDICTING THE OUTCOMES OF A GIVEN STOCHASTIC SCENARIO

Apr 3, 2015

Stochastic discrete event simulation

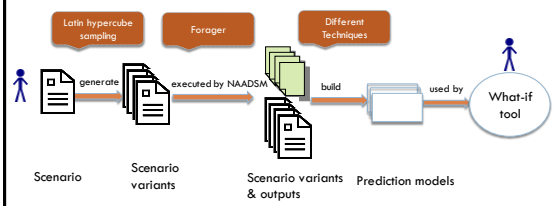
- Models all possible interactions between the entities of real-world system (scenario)
- Events can change with certain probability
- Several hours needed to execute one iteration
- Large number of iterations has to be executed to have results with high confidence

The problem

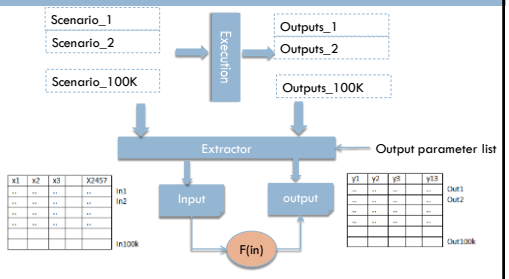
- When a planner creates a scenario, it needs to be executed multiple times
- The planner has to wait hours or even days to have the scenario's outcomes
- Performing what-if analysis is very expensive (sometimes impossible)
  - Preventing the appropriate preparations for critical situations

Building what-if tool that allows real time analysis

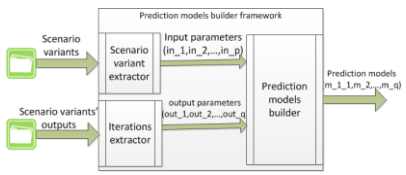
- To reach that point:



Estimating relationship between the input and outputs

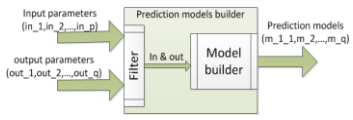


Configurable framework for building the prediction models



One prediction model for each output parameter

7



- Input dimensionality is reduced
  - Based on correlation to output parameter
  - Collinearity between the input parameters is removed

Python is used to build and test the prediction models

8

- Python is used as language to build and test the prediction models
- scikit-learn and pybrain machine learning packages are used to apply different approaches
- Applied learning techniques
  - Artificial neural networks
  - Linear regression
  - Ensemble methods
    - Random forest
    - Gradient boosting

Building and evaluating the prediction models

9

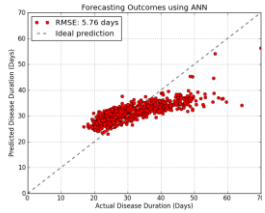
- The models are built using 90,000 observations and tested on 10,000 unseen observations
- Root mean squared error (rmse) is used to measure the prediction error

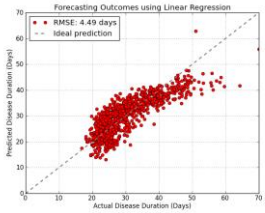
$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- K-fold cross-validation is used to assess the generalization

Our predictions vs ideal predictions

10





Applying ensemble methods

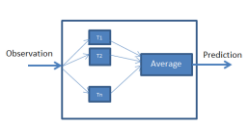
11

- An ensemble method is a process that builds multiple models by applying the learning process to a given problem
- The prediction of a given input is an aggregation of predictions of models in the ensemble
- The predictions obtained from the ensemble are usually better than any predictions obtained from individual predictor

Applying random forest to build the models

12

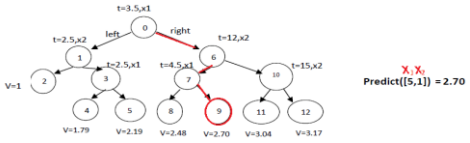
- Builds a collection of de-correlated decision trees (estimators)
- Each tree is independently built from sample and features randomly drawn with replacement from the training data



It uses decision tree as regression estimator

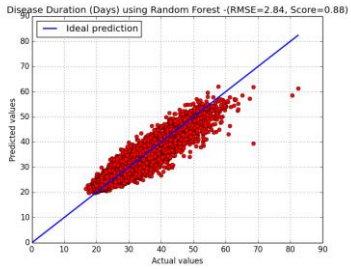
13

- In training process, a decision tree fits to the training data
- Leaf nodes hold the predicted values, while the other nodes split features and thresholds



Predicting disease duration (days) using random forest

14



- ANN (RMSE: 5.76)
- LR (RMSE: 4.49)

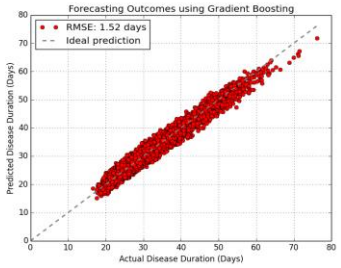
Applying gradient boosting

15

- It starts with quite weak model and improves it in stagewise manner
- In each stage the weights of poor predicted observations are increased
- In next stage the learning algorithm will focus more on poorly predicted data(high weighted data) than on good predicted data

Predicting disease duration (days) using gradient boosting

16



- ANN (RMSE: 5.76)
- LR (RMSE: 4.49)
- RF (RMSE: 2.84)

17 Questions?

What-If tool:
http://galileo.cs.colostate.edu/what-if-tool/