Machine Learning Set 01

2 python questions, 3 statistics & probability, 3 Feature engineering, 4 machine learning, 1 regularization, 2 loss and metrics, 2 scenario based questions

Scenario based question

- You are given a dataset with missing values that spread along 1 standard deviation from the median. What percentage of data would remain unaffected?
- You build a random forest model with 1000 trees. Training error is 0.0, while the validation error is 29.6. Explain what went wrong and how could we improve the result

Python & SQL

- Why is Python a dynamically typed programming language? (easy)
- What is the difference between left join and right join? (medium)

Statistics & Probability

- Correlation vs covariance
- Given two fair dice, what is the probability that the two sum to 8?
- What is selection bias? (easy) *(hint: https://youtu.be/FC3nHzbqMVE)*

Feature Engineering

- How do you handle missing values? (easy)
- Explain normalization vs standardization (medium)
- What is the difference between one hot encoding and label encoding (easy)

Machine learning

- What is the difference between Supervised and Unsupervised learning and provide applications for each (easy) *(hint: https://youtu.be/uK68Iz7toNQ)*
- What is the difference between hard margin SVM and soft margin SVM? (medium)
- Explain bagging and boosting (hard)
- What is pruning in decision trees. How is it done? (hard)
  (hint: https://youtu.be/wsH55R5dJCY)

Regularization

- Explain lasso vs ridge regression (medium)

Loss and metrics

- Explain precision, recall, specificity and sensitivity (easy)
  (hint: https://youtu.be/AgHXr2CDjNo)
- What is the difference type 1 and type 2 error (medium)