Weekly report of lessons

Name: Suryam Arnav Kalra
Roll No: 19CS30050
The week: 6th September 2021 to 10th September 2021

The topics covered:

- Expectation and Variance
- Error of hypothesis
- Modeling as a Bernoulli Distribution
- Bias and Variance of an estimate
- Confidence Interval
- General approach for deriving CI
- Central Limit Theorem
- Comparing two hypothesis
- Comparing two learning schemes
- K-fold cross validation
- Random process
- Bayes' theorem
- Learning Scenarios
- Features of Bayesian learning
- Least mean squared error estimate

Summary topic wise:

- Expectation and Variance: $E(X) = \sum xP(x)$ and $Var(X) = E(X^2) - E(X)^2$
- Error of hypothesis:
  - Sample Error : $E_S(h) = \frac{\sum e(f(x),h(x))}{n}$
  - True Error: $E_D(h) = P\{f(x) \neq h(x)\}$
- Modeling as a Bernoulli Distribution: Let $P(X = 1) = p$ then $E(X) = p$ and $Var(X) = p(1-p)$
- Bias and Variance of an estimate: If there are r errors in n symbols then $E(r) = np$ and $Var(r) = np(1-p)$ where $p = E_D(h)$
- Confidence Interval: N% confidence interval is the interval containing the true value with probability N%.
- General approach for deriving the CI : First, determine the probability distribution D_Y of Y and then find the thresholds L and U such that N% mass of D_Y falls between L and U.
- Central Limit Theorem: Let $Y \sim D(\mu, \sigma)$ and we have n independent observation of Y, then $Y_a \sim N(\mu, \sigma/\sqrt{n})$ where $Y_a = Average(Y_1, Y_2, \ldots, Y_n)$
- Comparing two hypothesis: Two competing hypothesis can be compared based on the difference between their errors in the distribution D as : $d = E_D(h_1) - E_D(h_2)$
- Comparing two learning schemes: We can compare two learning schemes based on the difference between their performance measure on the same data set(both training and test data). Let there be K observations $Y_1, Y_2, \ldots, Y_k$ then we can find Y_a using the CLT as stated above.

- K-fold cross validation: Partition the data set into K disjoint sets. Use one set as test data and the other K-1 sets as training data and repeat the process K times.
- Random process: When the data is produced through a process which is not completely known or is random is called a random process.
- Bayes' Theorem: $P(h|D) = \frac{P(h)P(D|h)}{P(D)}$ where $P(h|D)$ = Posterior probability, $P(h)$ = Prior probability, $P(D|h)$ = Likelihood function and $P(D)$ = Unconditional probability
- Learning Scenarios: There are two learning scenarios as given below:
    - MAP : $h_{MAP} = argmax\ P(D|h)P(h)$
    - ML : $h_{ML} = argmax\ P(D|h)$
- Features of Bayesian learning: It allows for flexible learning from each instance and also used the prior knowledge of hypothesis. It also accommodates hypothesis with probabilistic distribution.
- Least Mean Squared Error Estimate: $h_{ML} = argmin\ \sum(y(i) - h(x_i))^2$

Concepts challenging to comprehend:

Confidence Interval along with comparison of hypothesis and learning schemes are a little bit challenging to comprehend.

Interesting and exciting concepts:

Bayesian Theory and Methods for cross validation are quite interesting and exciting to learn.

Concepts not understood:

After going through the book and the video lectures the concepts are clearly understood.

Any novel idea of yours out of the lessons:

Using Confidence Interval, we can devise a method for pruning of regression/decision trees. It can be a post pruning method in which we will remove the node if the likelihood of the outcome is within the confidence interval that can be calculated using the methods learned in hypothesis evaluation.

Difficulty level of the Quiz:

The level was slightly on the tougher side.

Was the time given to you for solving the quiz appropriate? If not, why?:

Yes, the time given to solve the quiz was appropriate

Did the quiz questions enhance your understanding of the topics covered:

Yes, some questions such as depth of the decision tree and VC dimension analysis helped me further understand these topics.