Machine Learning (CS60050) – Weekly Report

Kaushal Banthia (19CS10039)

Week 4: 8th – 10th September, 2021

Topics Covered:

- Accuracy of a Hypothesis and its Estimation
- Error of Hypothesis
- Probabilistic Analysis and Confidence Interval
- Central Limit Theorem
- Comparing 2 Hypotheses
- K-fold cross validation and comparison
- Random Process
- Bayesian Inferences and Learning Scenarios
- Features of Bayesian Learning
- Concept Learning under Bayesian framework
- Least mean squared error estimate as the ML hypothesis

Summary (Topic Wise):

- Accuracy of a Hypothesis and its Estimation

  - ➢ It is pretty straightforward in a large dataset, but somewhat challenging in a dataset, with limited training examples. The difficulties that might arise are:
    - Bias in the estimate
    - Variance in the estimate
  - ➢ This leads us to ask the question - Given a hypothesis h over n examples randomly drawn from a distribution D, the best estimate of accuracy of h?

- Error of Hypothesis

  - ➢ We define sample error $E_S(h) = \frac{1}{n}\sum_{x \epsilon S} e(f(x), h(x))$ and true error $E_D(h) = Pr_{x \epsilon D}\{f(x) \neq h(x)\}$, such that,
    - x is an instance (an element of D)
    - S is a data sample with size n
    - h is a hypothesis from X→ $\{0, 1\}$
    - f is a target function from X→ $\{0, 1\}$
    - e is the error function with $e(x, y) = 1, if\ x \sim = y$ & $e(x, y) = 0\ otherwise$
  - ➢ Given r errors in n samples, $E_S(h) = \frac{r}{n}$
  - ➢ If we have no other information at hand, then the most probable choice for $E_D(h)$ would be $E_S(h)$, because $E_D(h)$ lies between the confidence interval $E_S(h) \pm 1.96\sqrt{\frac{E_S(h)(1-E_S(h))}{n}}$ with approximately 95% probability.
  - ➢ This approximation works well when $n * E_S(h) * \left(1 - E_S(h)\right) \geq 5$
  - ➢ Smaller estimate value requires larger sample size.

- Probabilistic Analysis and Confidence Interval

  - Probability of error for a sample is given by $E_D(h) = p$
  - Probability of r error in n samples is given by $\binom{n}{r} p^r (1-p)^{n-r}$
  - $E(r) = n * p \text{ and } var(r) = n * p * (1-p) \text{ and } E\left(\frac{r}{n}\right) = p$
  - $var(p) \cong var\left(\frac{r}{n}\right) = \frac{n*p*(1-p)}{n^2} = \frac{p*(1-p)}{n}$
  - N% confidence interval is the interval containing the true value with N% probability. For large samples, binomial distribution approximates the normal distribution, such that $N\%\ C.I. = \ \mu \pm Z_N \sigma$
  - A general approach for deriving the Confidence Interval of an estimate
    - Let Y be the estimator of a parameter p.
    - Determine the probability distribution of D_Y of Y, by calculating its mean and variance.
    - Determine the N% C.I. by finding thresholds L and U such that N% mass of D_Y falls between L and U. Make use of the central limit theorem for estimating mean of a distribution.

- Central Limit Theorem

  - For $Y \sim D\ (\mu, \sigma)$, where D is any arbitrary probability distribution, $\mu \ = \ E(Y)$ and $\sigma^2 \ = \ E\left(\left(Y - E(Y)\right)^2\right)$
  - Consider n independent observations of Y: $Y_1, Y_2,\ \ldots\ , Y_n$
  - $Y_a = Average(Y_1, Y_2,\ \ldots\ , Y_n)$
  - We can approximate Y_a as $N\ (\mu, \frac{\sigma}{\sqrt{n}})$ (as $n \to \infty$) (since $\frac{Y_a - \mu}{\sigma/\sqrt{n}} \sim N\ (0,1)$)

- Comparing 2 Hypotheses

  - If $h_1 \text{ and } h_2 \text{ are } 2 \text{ competing hypotheses}$, d is the difference of their errors in the distribution D. $d = E_D(h_1) - E_D(h_2)$ and d' is the observed d, while they are tested on two independent samples S1 and S2 of sizes $n_1 and\ n_2 \geq 30$.
    $d' = E_{S1}(h_1) - E_{S2}(h_2)$. Also, $E(d') = d$, as both $E_{S1}(h_1) \text{ and } E_{S2}(h_2)$, follow normal distribution.
  - $\sigma_{d'}^2 = \frac{E_{S1}(h)*(1-E_{S1}(h))}{n_1} + \frac{E_{S2}(h)*(1-E_{S2}(h))}{n_2}$
  - Two different learning schemes can be compared by taking the difference of a performance measure of learning scheme 1 and learning scheme 2, on the same dataset, both train and test.
  - For k observations, $\sigma_Y = \frac{1}{k-1} \sum_{i=1}^{K} (Y_i - Y_a)^2$, where σ_Y is the unbiased estimate of the standard deviation of Y.

- K-fold cross validation and comparison

  - Partition the data set S into k disjoint sets, numbered from 1 to k. We can then use the i^{th} partition as a test data set, and the remaining as training sets and observe their respective Y_i
  - Compute the N% Confidence Interval. A value without such probabilistic interpretation is not statistically accepted.

- Random Process

  - Data generated by a process not completely known. Thus, the process is modeled as a random process, since it is convenient to handle the gap in knowledge.
  - Data x is treated as an outcome of a random variable, X, where P(X=x) is observable and it is possible to associate classes with data.
  - $P(X = x|C_i), C_i\ is\ the\ i^{th}\ class, i =\ 1,2, \dots, K$

- Bayesian Inferences and Learning Scenarios

  - Conditional Probability is written as $P(A|B)$. This means the probability of event A happening, given that event B has already happened.
  - $P(A \cap B) = P(A)P(B|A) = P(B)P(A|B)$
  - Bayes Theorem: $P(h|D) = \frac{P(h)P(D|h)}{P(D)}$ = Posterior Probability (the probability of the hypothesis, given the data.
  - Maximum a posteriori (MAP) Hypothesis: $h_{MAP} \equiv argmax_{h\epsilon H}P(h|D) = argmax_{h\epsilon H}\frac{P(h)P(D|h)}{P(D)} = argmax_{h\epsilon H}P(D|h)P(h), since\ P(D)\ is\ the\ same\ for\ any\ h$
  - Maximum Likelihood Hypothesis (ML) Hypothesis: $h_{ML} \equiv argmax_{h\epsilon H}P(D|h)$

- Features of Bayesian Learning

  - Flexible learning from each observable instance as there is either an increasing or a decreasing probability of a hypothesis being correct. Also, Prior knowledge of a is hypothesis used (Inductive bias).
  - It also accommodates hypotheses with probabilistic prediction. Each hypothesis in the version space of concept learning will have a weight while taking a decision.
  - It also provides a framework for optimal decision making, even when the computation is intractable!

- Concept Learning under Bayesian framework

  - Likelihood: $P(D|h) = 1\ if\ h\ \in VS_{H.D}$ & $0\ otherwise$
  - Prior: $P(h) = \frac{1}{|H|}$ (Prior could be taken with uniform distribution)
  - Marginal Probability of Data: $P(D) = \sum_{h\epsilon H}\left(P(D|h) * P(h)\right) = \frac{VS_{H,D}}{|H|}$
  - $P(h|D) = \frac{P(D|h)*P(h)}{P(D)} = \frac{1}{|VS_{H,D}|}, if\ h\ in\ VS, else\ 0$

- Least mean squared error estimate as the ML hypothesis

  - Target function: y = f(x) and Hypothesis: h
  - Mean squared error (MSE) = $\sum_{i=1}^{n}\left(y_i - h(x_i)\right)^2$
  - $y_i = h(x_i) + e_i\ for\ i = 1, 2, \dots, n\ and\ e_i \sim N(0, \sigma)$
  - Probability of $P(D|h) = \left(\frac{1}{\sqrt{2\pi\sigma}}\right)^n \prod_{i=1}^{n} e^{-\frac{1}{2}\left(\frac{y_i - h(x_i)}{\sigma}\right)^2}$
  - Log likelihood: $\log\left(P(D|h)\right) = n * \ln\left(\frac{1}{\sqrt{2\pi\sigma}}\right) - \sum_{i=1}^{n}\frac{1}{2}\left(\frac{y_i - h(x_i)}{\sigma}\right)^2$
  - $h_{ML} = argmin_{h\epsilon H}\sum_{i=1}^{n}\left(y_i - h(x_i)\right)^2$

Concepts Challenging to Comprehend: None yet.

Interesting and Exciting Concepts: K-fold cross validation and comparison

Concepts not understood: None yet.

A novel idea: The K-fold cross validation technique can be improved if we take the test data set as a combination of 2-3 random sets picked up from the set of disjoint sets, instead of just one set. It would also help, if this process is repeated many times (say 10-20 times) with variations in the choices of the training and the test data sets each time.

Difficulty level of the Quiz: Fair

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

Did the quiz questions enhance your understanding of the topics covered?: Yes, the questions given in the quiz did enhance my understanding, as they included both numerical and objective theory questions, thus covering all the aspects of all topics. The type of test was very good and informative.