# CS60050 Machine Learning - Weekly Report

Nisarg Upadhyaya (19CS30031)

Week 5: 8-10 September

## 1 Topics covered

- Estimating Hypothesis Accuracy
- Confidence Interval
- Central Limit Theorem
- Comparing two hypotheses and learning schemes
- K-fold cross validation
- Bayesian Decision Theory and Learning
- Bayes Theorem and different learning scenarios
- Features of Bayesian Learning
- Concept learning under Bayesian framework
- Least mean squared error estimate as the ML hypothesis

## 2 Summary

### 2.1 Estimating Hypothesis Accuracy

Estimating hypothesis accuracy is straightforward when data is plentiful. However, when there is only limited data there are two problems:

1. Bias in estimate: Observing accuracy on training examples is not a good idea. Biased to give good accuracy from the sample it was trained.
2. Bias in variance: The smaller the set of test examples, the greater the expected variance.

Given a hypothesis $h$ over $n$ examples randomly drawn from a distribution $D$, what is the best estimate of accuracy of $h$ and the error in the estimate? If there are $r$ errors in $n$ samples then the sample error $E_S(h) = r/n$. If $f$ is the target function then true error $E_D(h) = Pr_{x \in D}\{f(x) \neq h(x)\}$. Given no other data a good estimate of $E_D(h)$ is given by the $N\%$ confidence interval. We can say approximately with $N\%$ probability that $E_D(h)$ lies between

$$E_S(h) \pm Z_N \cdot \sigma_{E_S(h)}$$

$$\sigma_{E_S(h)} = \sqrt{\frac{E_S(h)(1 - E_S(h))}{n}}$$

### 2.2 Central Limit Theorem

A general approach for finding confidence interval involves finding the probability distribution $D_Y$ of an estimator $Y$ for the parameter to be estimated. We then find thresholds $L$ and $U$ between which $N\%$ mass of $D_Y$ lies. The central limit theorem helps in the process because if the estimator $Y$ is the mean of some sample the distribution governing the estimator can be approximated by a Normal distribution for sufficiently large $n$. In practice it can be used when $n \geq 30$. If $Y_a = Average(Y_1, Y_2, ...Y_n)$ then as $n \to \infty$

$$Y_a \sim \mathcal{N}(\mu, \sigma/\sqrt{n})$$

### 2.3 Comparing two hypotheses and learning schemes

We want to estimate the difference $d$ between the true errors of two hypothesis $h1$ and $h2$. Suppose the observed $d$ when tested on two independent samples $S1$ and $S2$ is $\hat{d} = E_{S1}(h1) - E_{S2}(h2)$ and $\sigma_{\hat{d}} = \sqrt{\sigma^2_{E_{S1}(h1)} + \sigma^2_{E_{S2}(h2)}}$ then the difference with $N\%$ C.I. is $\hat{d} \pm Z_N \cdot \sigma_{\hat{d}}$. We compare learning schemes by an estimator $Y$ which measures the difference of performance measure of the two schemes on the same data set. For $k$ observations $Y_1, Y_2, ...Y_k$:

$$Y_a = Average(Y_1, Y_2, ...Y_k)$$

$$\sigma^2_Y = \frac{1}{k-1} \sum_{i=1}^{k} (Y_i - Y_a)^2$$

We have the $N\%$ C.I. as $Y_a \pm T_{N,(k-1)} \cdot \frac{\sigma_Y}{\sqrt{k}}$ where $T_{N,(k-1)} \to Z_N$ as $k \to \infty$.

### 2.4 K-fold cross validation

This involves partitioning data set $S$ in $k$ disjoint sets $S_1, S_2, ...S_k$. Use the set $S_i$ as test data set and the rest as training data set. Observe $Y_i, i = 1, 2, ...k$ and compute the $N\%$ C.I. as discussed before.

### 2.5 Bayesian Decision Theory and Learning

In Bayesian learning processes are modelled as random processes. Data $x$ is treated as an outcome of a random variable $X$ and we observe $P(X = x)$. Data with classes associated is observed as $P(X = x|C_i)$, where $C_i$ is the $i$-th class.

#### 2.5.1 Bayes Theorem and different learning scenarios

$$P(h|D) = \frac{P(h)P(D|h)}{P(D)}$$

where $P(h)$ is the *prior probability*, $P(D|h)$ is the *likelihood function*, $P(D)$ is the *unconditional probability* and $P(h|D)$ is the *posterior probability*. Some learning scenarios are:

1. Given a set of candidate hypotheses $H$ the learner is interested in finding the most probable $h \in H$ given the observed data $D$. In this case $h_{MAP}$ (maximum a posteriori hypothesis) $= argmax_{h \in H} P(D|h)P(h)$
2. In some cases we assume that all hypotheses in $H$ are equally probable. In this case $h_{ML}$ (maximum likelihood hypothesis) $= argmax_{h \in H} P(D|h)$

#### 2.5.2 Features of Bayesian Learning

- Flexible learning from each observable instance.
- Prior knowledge of hypothesis used.
- Accommodates hypotheses with probabilistic prediction.
- Provides a framework of optimal decision making.

#### 2.5.3 Concept learning under Bayesian framework

We take the prior probability considering a uniform distribution, i.e., $\frac{1}{|H|}$, the likelihood to be 1 if $h$ is in the version space otherwise 0 and $P(D) = \frac{|VS_{H,D}|}{|H|}$. Finally we get $P(h|D) = \frac{1}{|VS_{H,D}|}$ if $h$ is in $VS$ otherwise 0.

#### 2.5.4 Least mean squared error estimate as the ML hypothesis

If we assume that the different training values $d_i$ are generated by adding random noise to the true target value where this random noise is from a Normal distribution with zero mean and independent for each training value then we can argue that the maximum likelihood hypothesis is the one which minimizes the sum of squared errors between the observed training values $d_i$ and the prediction $h(x_i)$.

## 3 Challenging concepts

The determination of $N\%$ C.I. using the variance of the distribution was a bit confusing. However, visualising the same from the plot helped understand it clearly.

## 4 Interesting concepts

Central limit theorem because of the fact that without even knowing the distribution that governs individual samples being observed we know the distribution that governs the sample mean. This was quite surprising.

## 5 Concepts not understood

None

## 6 Ideas

If we have a lot of prior information available which is associated with the learning task at hand, Bayesian learning can be leveraged as it helps take into account prior information. We can use this in crucial medical diagnostics by taking into consideration the medical history of the patient. So we can find the probability that a patient is suffering from a condition $c$ based on his past medical conditions.

# 7 Quiz Feedback

Difficulty level: Moderate
Time given: Sufficient
The quiz questions did enhance my understanding and brought more clarity to the topics taught in class.