Evaluating hypotheses

Jayanta Mukhopadhyay
Dept. of Computer Science and Engg.

Books

- Chapter 5 of “Machine learning” by Tom M. Mitchel

Estimating accuracy of a hypothesis

- Straightforward given a large dataset.
- Two key difficulties in limited data
  - Bias in the estimate
    - Biased by examples only.
      - May not be the same on unseen data.
    - More problematic in a rich hypothesis space.
    - Estimate using test data only.
  - Variance in the estimate
    - May vary with test data set.
    - Smaller the size, greater the expected variance.

Two key questions

- Given a hypothesis h over n examples randomly drawn from a distribution D, the best estimate of accuracy of h?
- Probable error in the estimate of accuracy?
  - Probable range of estimates?
    - So that true estimate lies within it with high probability of confidence (say 95%).

Error of hypothesis

- x: an instance,
  - an element of D
- S: a data sample
  - Size=n
- h: a hypothesis
  - h: X → {0,1}
- f: a target function
  - f: X → {0,1}
- e: the error function:
  - e(x,y)= 1, if x ~=y,
    = 0 otherwise

- Sample error

$$E_S(h) = \frac{1}{n}\sum_{x \epsilon S} e(f(x), h(x))$$

- True error

$$E_D(h) = Pr_{x \epsilon D}\{f(x) \neq h(x)\}$$

Given r errors in n samples:

$E_S(h)=r/n$

Statistical theory

- Given r errors in n samples ($n \geq 30$), $E_S(h)$?
  - $E_S(h)=r/n$
- Given no other information, most probable $E_D(h)$?
  - $E_D(h) = E_S(h)$
- With approximately 95% prob., $E_D(h)$ lies between

Confidence interval $\longrightarrow$

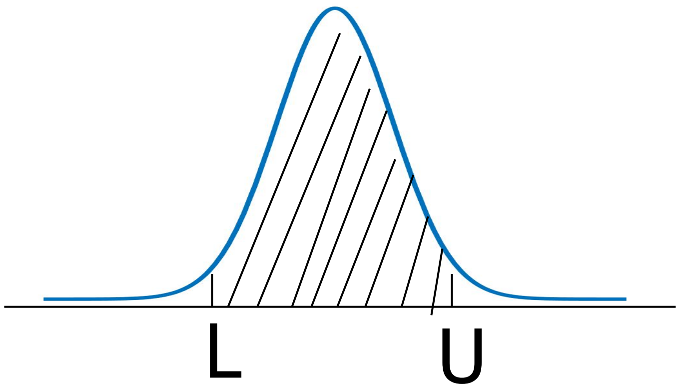
$$E_S(h) \pm 1.96 \sqrt{\frac{E_S(h)(1 - E_S(h))}{n}}$$

Above approximation works well for $n\, E_S(h)(1\text{-}\, E_S(h)) \geq 5$

minimum n ~ 30 when $E_S(h) = .213$.

Smaller estimate value requires larger sample size.

An example



- n=40, r=12
  - $E_S(h)=r/n = 0.3$
  - Confidence interval: $0.3 \pm 1.96 \text{ x } .07 = 0.3 \pm .14$
  - $[0.16, 0.44]$
    - With 95% probability, $E_D(h)$ lies within $[0.16, 0.44]$
    - With 97.5% probability $E_D(h)$ is less than 0.44.
  - From the properties of Binomial (~ Normal for large n) distribution.

Probabilistic analysis: Bias and variance of an estimate

- Prob. of error for a sample: $E_D(h) = p$
- Prob. of r errors in n samples: $\binom{n}{r} p^r (1-p)^{n-r}$

Unbiased estimate
E(estimate)=param

- $E(r)=np$, and $var(r)=np(1\text{-}p)$
- $E(r/n)=p$, and

Variance of estimate $E_S(h)$

- $var(p) \simeq var(r/n)=(np(1\text{-}p))/n^2=(p(1\text{-}p))/n$

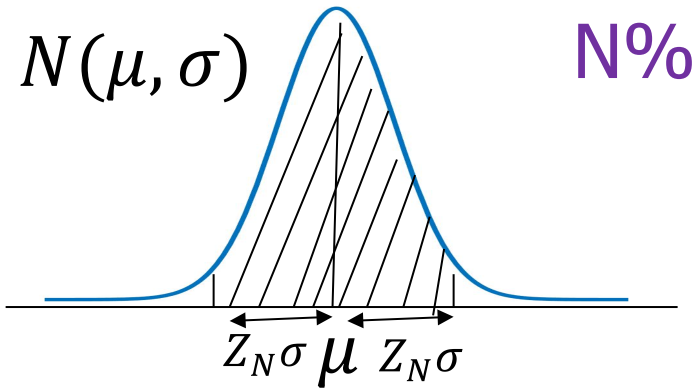
Bias of estimate: E(estimate) – true-parameter-value

Inductive bias: A set of assertions.

Bias of estimate: A numerical quantity

Confidence interval

- N% Confidence interval:
  - interval containing the true value with probability N%.
- For large sample, Binomial distribution approximates Normal Distribution.



N% C.I.= $\mu \pm Z_N \sigma$

N%	50%	68%	80%	95%	98%	99%
Z_N	0.67	1.0	1.28	1.96	2.33	2.58

A general approach for deriving C.I. of an estimate

- Let Y be the estimator of a parameter p.
- Determine the probability distribution D_Y of Y
  - 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.
  - **Use of Central Limit Theorem**
    - Estimating mean of a distribution
      - Estimation problem mapped to an estimation of a parameter following Normal distribution.

Central Limit Theorem

- $Y \sim D\,(\mu, \sigma)$ Any arbitrary probability distribution.
  - μ: $E(Y)$, and $\sigma^2 = E((Y\text{-}E(Y))^2)$
- n independent observation of Y
  - $Y_1, Y_2, \ldots Y_n$
- Y_a=Average $(Y_1, Y_2, \ldots Y_n)$
- $Y_a \sim N\,(\mu, \sigma/\sqrt{n})$ (as $n \rightarrow \infty$)
  - Normal distribution
  - $(Y_a - \mu) / (\sigma/\sqrt{n}) \sim N\,(0,1)$

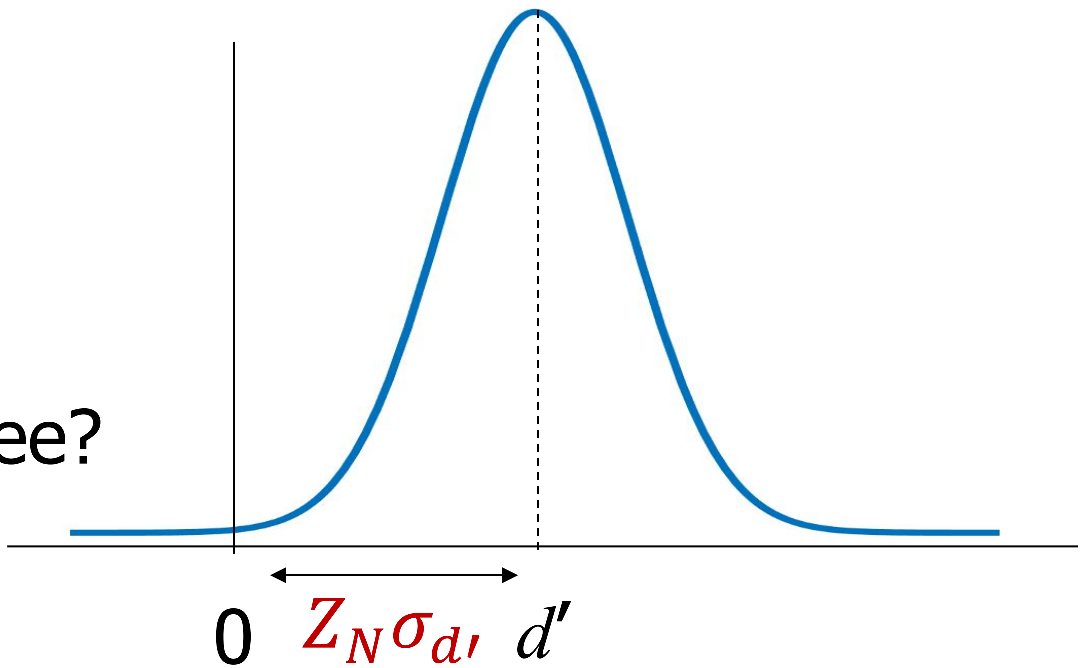
Comparing two hypotheses

- h_1, h_2: Two competing hypotheses.
- d: Difference of their errors in the distribution D.
  - $d = E_D(h_1) - E_D(h_2)$
- d': Observed d while they are tested on two independent samples $S1$ and $S2$ of sizes $n_1, n_2 \geq 30$.
  - $d' = E_{S1}(h_1) - E_{S2}(h_2)$
- $E(d') = d$ $\qquad \sigma_{d'}^2 = \frac{E_{S1}(h)(1 - E_{S1}(h))}{n_1} + \frac{E_{S2}(h)(1 - E_{S2}(h))}{n_2}$
  - as both $E_{S1}(h_1)$ and $E_{S2}(h_2)$ follow Normal Distr.
- $Var(d')$ is the sum of variances of $E_{S1}(h_1)$ and $E_{S2}(h_2)$.
  - N% CI = $d' \pm Z_N \sigma_{d'}$

$E_{S1}(h_1) > E_{S2}(h_2)$?

N%	50%	68%	80%	95%	98%	99%
Z_N	0.67	1.0	1.28	1.96	2.33	2.58

Can you apply it in pruning nodes of a decision tree?



d'>0 with (N+(100-N)/2)% confidence if the range lies in the +ve side.

Comparing two learning schemes

- Y= Diff. of a perf. measure of LS1 and LS2 on the same data set (both training and test data).
- Let there be k observations.
  - $Y_1, Y_2, .. Y_k$
  - $Y_a = Avg(Y_1, Y_2, .., Y_k)$
  - $\boldsymbol{\sigma}_Y$= Unbiased estimate of s.d. of Y
  - N% C.I.: $Y_a \pm T_{N,(k-1)} . \boldsymbol{\sigma}_Y / \sqrt{k}$

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

A constant from t-distribution of (k-1) d.f. for N% probability sum within the interval.

As $k \rightarrow \propto$, $T_{N,(k-1)} . \rightarrow Z_N$

K-fold cross validation and comparison

- Partition data set S in k disjoint sets, $S_1, S_2, .. S_K$.
- Use i th partition as a test data set and the rest as training set and observe Y_i, i=1,2,..k.
- Compute the N% confidence interval.
  - For any statistics, use similar technique to determine the confidence interval.
- A value without such probabilistic interpretation is not statistically accepted.

Summary

- Unbiased estimate of error as a fraction of test samples not satisfying target function, i.e. $E_S(h)=r/n$.
  - Compute also its variance as: $E_S(h).(1-E_S(h))/n$
  - N% Confidence interval defined using them.
- Evaluate competing hypothesis by using the probability distribution of the difference of errors.
- Central limit theorem used for estimating average of a statistics with a C.I.
- The same approach used in comparing two learning schemes by applying k-fold cross-validation.

Thank you!