Evaluating Hypotheses

Overview

- Sample versus Generalisation Error
- Error Estimators and Confidence Intervals
- Methods of Evaluation
- Comparing Machine Learning Methods

Sample vs Generalisation Error

- Suppose we have
  - hypothesis *h*
  - target concept/function *c*
  - data distribution *D*
  - sample of data *S* (drawn from *D*)

- Sample error

$$\text{error}_S(h) \equiv 1/n \, . \, \Sigma_{x \in S} \, \delta(c(x) \neq h(x))$$

where $\delta(c(x) \neq h(x)) = 1$ if $c(x) \neq h(x)$; 0 otherwise

- Generalisation error

$$\text{error}_D(h) \equiv \Pr_{x \in D} [\, c(x) \neq h(x) \,]$$

- How well does $\text{error}_S(h)$ estimate $\text{error}_D(h)$?

Bias and Variance Over Sample Error

- Bias

$$E[\mathrm{error}_S(h)] - \mathrm{error}_D(h)$$

  - $\mathrm{error}_S(h)$ can be optimistically biased when S is used to train h
  - Use S independent to h for an unbiased estimate of $\mathrm{error}_D(h)$

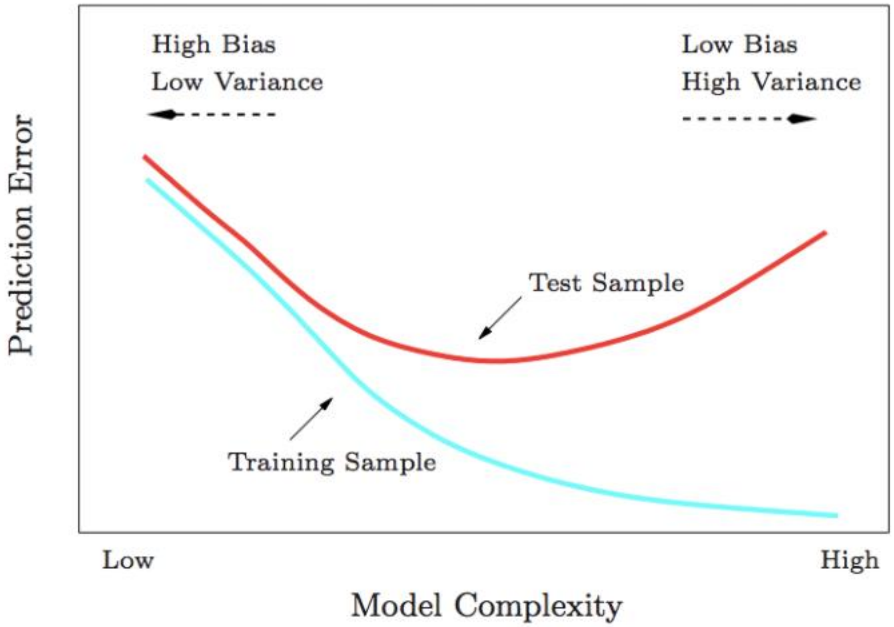
- Variance

$$E[\ (\ \mathrm{error}_S(h) - E[\mathrm{error}_S(h)]\ )^2\ ]$$

  - Even with unbiased S, $\mathrm{error}_S(h)$ may still VARY from $\mathrm{error}_D(h)$
  - Smaller $S \Rightarrow$ greater expected variance

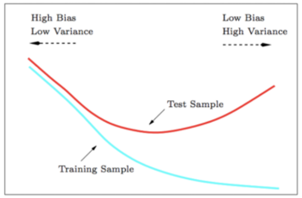
Bias and Variance Over Sample Error

- More complex models have
  - Lower error bias
    - More expressive hypothesis representations allow models to be overfit
  - Higher error variance
    - Allow small changes in data to cause greater changes in trained model



Bias and Variance Over Sample Error

- Example – decision trees
  - As decision trees grow larger
    - More constraints placed over instances
    - Fewer instances per leaf node
  - Evidence or support for a given rule (at a leaf node) is weaker
  - Small changes in training data will affect resultant tree more
  - A tree that fits a training set *S'* too perfectly, will not adequately generalise over *D* (i.e., overfitting)
    - Measuring error over *S*' is a poor indicator of *D*
    - It will likely give an overly optimistic estimate (as since the graph from the previous slide)



General Practice

- Use data independent to training for evaluation

- This corresponds to any part of the training process, including the use of wrapper-based methods
  - Feature selection
  - Algorithm hyperparameters

- There may be a need for validation as well as testing data
  - First validate models for selection
  - Then evaluate selected models

- Usually, all data is drawn from the same distribution
  - May wish to test a *h* on data from a different distribution
  - Ensure data is drawn from distribution being reported on

Estimating $error_S(h)$

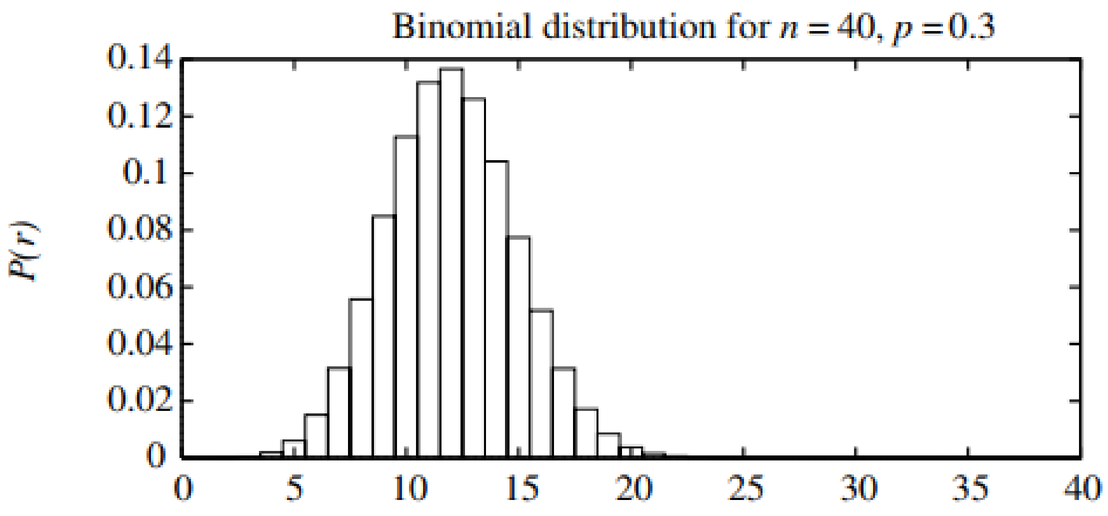
- Experiment
  1. Choose sample S (independent of h) according to D
     - where the size of S, $|S| = n$
  2. Measure $error_S(h)$

- $error_S(h)$ is a random variable
  - i.e., result of an experiment

- The success or failure of each $x \in S$, is a ***Bernoulli Trial***
  - The outcome $c(x) \neq h(x)$ is either True or False
  - Given h, the outcome for each x_i and x_j are independent, where $x_i, x_j \in S, i \neq j$

Random Variable $error_S(h)$

- Rerun the experiment with different randomly drawn *S* (of size *n*)
  - Probability of observing *r* misclassified examples is given by the ***Binomial Distribution***



$$P(r) = n! \,/\, (r!\,(n - r)!) \,.\, error_D(h)^{\,r} \,.\, (1 - error_D(h))^{\,n - r}$$

Binomial Distribution

- Probability r misclassifications by h over n instances is

$$\Pr[\, X = r \,]$$

- Since X follows a Binomial distribution, we have

$$\Pr[\, X = r \,] = P(r) = n! \,/\, (r!\,(n - r)!) \,.\, p^{\,r} \,.\, (1 - p)^{\,\mathrm{n - r}}$$

  – Expected, or mean value X (based on n trials $X_1, \ldots, X_n$)

$$E[X] \equiv E[X_1 + \ldots + X_n] = E[X_1] + \ldots + E[X_n] = p + \ldots + p = np$$

  – Variance of X is

$$\begin{aligned} \mathrm{Var}[X] \text{ or } \sigma^2_{\mathrm{X}} \equiv \mathrm{Var}[X_1 + \ldots + X_n] &= \mathrm{Var}[X_1] + \ldots + \mathrm{Var}\,[X_n] \\ &= p(1 - p) + \ldots + p(1 - p) = np(1 - p) \end{aligned}$$

  – Standard deviation of X is

$$\sigma_X \equiv (E[\,(X - E[X]\,)^2\,])^{0.5} = (np(1 - p))^{0.5}$$

Binomial Distribution for $error_S(h)$

- $error_S(h)$ follows a Binomial Distribution with
  - Expected, or mean $error_S(h)$ is

    $$E[error_S(h)] \equiv error_D(h) \approx p = r / n$$
  - $error_S(h)$ is an unbiased estimator of $error_D(h)$
    - Expected value of r is np (by Binomial Distribution)
    - Expected value of $r / n = p$ (since n is constant)
  - Variance in $error_S(h)$ comes from solely from variance in r
    - Variance of r is $np(1 - p)$
    - Variance of $error_S(h)$ is $np(1 - p) / n^2$ (try work this out)
  - Standard deviation of $error_S(h)$ is thus
    - Variance of Standard deviation of r divided by n
    - Or $((error_S(h))(1 - error_S(h)) / n)^{0.5}$

Central Limit Theorem

- Consider a set of independent, identically distributed random variables $Y_1, \ldots, Y_n$, all governed by an arbitrary probability distribution with mean μ and finite variance σ^2.

- With sample mean, $\bar{Y} = 1/n \,.\, \Sigma_{i = 0, \ldots, n} \; Y_i$

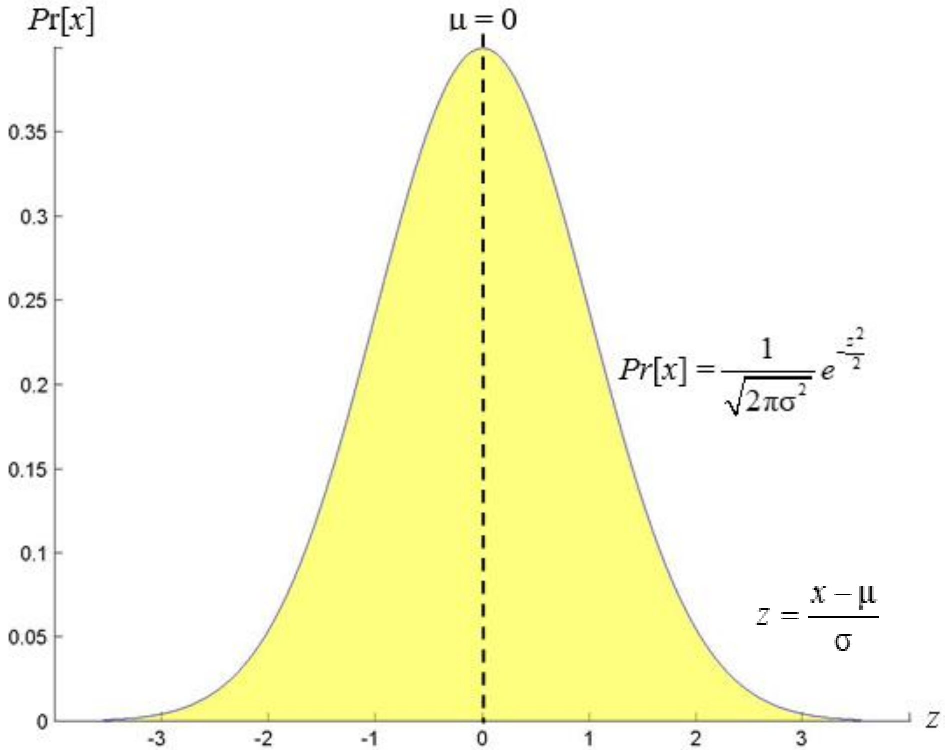
- **Central Limit Theorem**:

  As $n \rightarrow \infty$, the distribution governing $\bar{Y}$ approaches a **Normal distribution**, with mean μ and variance σ^2/n.

- Whenever we define an estimator that is a mean of some sample (e.g., $\mathrm{error}_S(h)$), the distribution governing this estimator can be approximated by a Normal distribution for sufficiently large n.

Normal Approximation

- A binomial distribution ***B***(*n*, *p*) may be approximated by the normal distribution ***N***(*np*, *np*(1 - *p*))
  - Loosely, this assumes that *n* is large enough and *p* is not too skewed towards the extremes (0 or 1)
  - By using this approximation, we may define a 1-tailed or 2-tailed confidence interval that encapsulates α% of the area under the normal curve



Example

- Suppose we observe 12 errors in a validation sample of 40 instances

  - $error_S(h) = r / n = 12 / 40 = 0.3$
  - $Var[r] = np(1 - p) = 40(0.3)(1 - 0.3) = 8.4$
  - standard deviation of $r = (8.4)^{0.5} \approx 2.9$
  - standard deviation of $error_S(h) = 2.9 / 40 = 0.07$
  - For a 95% confidence interval over $error_D(h)$:

    $$error_S(h) \pm 1.96((error_S(h))(1 - error_S(h)) / n)^{0.5}$$

  - For the example above, this gives the interval

    $$0.3 \pm 0.14$$

Problems with Normal Approximation

- Several arguments have been made ***against*** using the normal approximation
  - Bound may exceed [0,1]
  - Zero-width intervals at r = 0, 1; these falsely imply certainty
  - Observed inconsistencies with significance testing

- There are several more favourable alternatives
  - e.g., Wilson score

- You should review these independently

Questions?