Hypothesis Evaluation

- Two definitions of error

. The true error of hypothesis h with respect to target function f and distribution D is the probability that h will misclassify an instance drawn at random according to D :

$$error_D(h) \equiv \Pr_{x \in D}[f(x) \neq h(x)]$$

. The sample error of h with respect to target function f and data sample S is proportion of examples h misclassifies

$$error_S(h) \equiv \frac{1}{n} \sum_{x \in S} I(f(x) \neq h(x))$$

where $I(f(x) \neq h(x))$ is 1 if $f(x) \neq h(x)$, and 0 otherwise.

- Problems of estimating error

. $error_S(h)$ is an estimator of $error_D(h)$.

. How well does $error_S(h)$ estimate $error_D(h)$?

. bias of $error_S(h)$ as an estimator of $error_D(h)$:

$$b_{error_D}(error_s) = E[error_s] - error_D$$

if $b_{error_D}(error_s) = 0$ for all $error_D$, we say $error_s$ is an unbiased estimator of $error_D$.

. The mean square error of $error_s$ is given as follows:

$$\begin{aligned} E[(error_s - error_D)^2] &= E[(error_s - E[error_s] + E[error_s] - error_D)^2] \\ &= E[(error_s - E[error_s])^2] + E[(E[error_s] - error_D)^2] + \\ &\quad 2E[(E[error_s] - error_D](error_s - E[error_s])] \\ &= E[(error_s - E[error_s])^2] + (E[error_s] - error_D)^2 \\ &= Var(error_s) + b^2_{error_D}(error_s) \end{aligned}$$

That is, the mean square error of $error_s$ is equivalent to the variance of $error_s$ plus the square of bias of $error_s$.

. Let $X_i \in \{0,1\}$ be a random variable which has the mean $error_D$, that is, $E[X_i] = error_D$. Here, we assume that X_is are independent and identically distributed.

Then, $error_s$ can be described by

$$error_S = \frac{1}{N}\sum_{i=1}^{N} X_i$$

where N represents the total number of trials.

In this case,

$$E[error_S] = E[\frac{1}{N}\sum_{i=1}^{N} X_i] = \frac{1}{N}\sum_{i=1}^{N} E[X_i] = error_D.$$

That is, $error_S$ is an unbiased estimator of $error_D$.

. example:

Hypothesis h misclassifies 50 of the 100 samples in S.

In this case,

$$error_S(h) = \frac{50}{100} = 0.50.$$

Then, what is $error_D(h)$?

. Given observed $error_S(h)$ what can we conclude about $error_D(h)$?

- Binomial probability distribution

. Let X be a binomial random variable with parameters (n, p). Then, X represents the number of successes in n trials and p represents the probability of success.

. example: tossing a coin.

Probability $\Pr(r)$ of r heads in n coin flips can be described by

$$\Pr(r) = \binom{n}{r} p^r (1-p)^{n-r} = \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r}$$

where $p = \Pr(head)$.

In this case, the mean value of X is

$$E[X] = \sum_{i=0}^{n} i \Pr(i) = np \quad \text{and}$$

the variance of X is

$$Var(X) = E[(X - E[X])^2] = np(1-p).$$

. $error_S(h)$ follows a binomial distribution, that is,

$$error_S(h) = \frac{X}{n},$$

$$E[error_S] = E[\frac{X}{n}] = \frac{1}{n} E[X] = p = error_D, \text{ and}$$

$$Var(error_S) = Var(\frac{X}{n}) = \frac{1}{n^2} Var(X) = \frac{p(1-p)}{n} = \frac{error_D(1-error_D)}{n}.$$

– Normal distribution approximates Binomial

. Let X_i be a random variable which has the value of 0 or 1 and $\Pr[X_i = 1] = p$.

Then, the random variable X having binomial distribution with parameters (n, p) can be described by

$$X = \sum_{i=1}^{n} X_i.$$

Here, the mean of X_i is

$$E[X_i] = 1 \cdot p + 0 \cdot (1-p) = p \quad \text{and}$$

the variance of X_i is

$$Var(X_i) = E[X_i^2] - E^2[X_i] = p - p^2 = p(1-p).$$

. Central Limit Theorem:

Consider a set of independent, identically distributed (i. i. d.) random variables $X_1, X_2, \cdots, X_n$ all governed by an arbitrary probability distribution with mean μ and finite variance σ^2. Let us define a new random vector

$$X = \sum_{i=1}^{n} X_i.$$

Then, as n goes to infinity, the distribution governing X approaches a normal (or Gaussian) distribution, with mean $n\mu$ and variance $n\sigma^2$. That is,

$$X \sim N(n\mu, n\sigma^2).$$

cf. In the case of Bernoulli trial, $X \dot{\sim} N(n\mu, n\sigma^2)$ when $n \geqq 30$. That is, X has an approximately Normal distribution with mean $n\mu$ and variance $n\sigma^2$. Here, the sample error of h can be described by

$$error_S(h) = \frac{X}{n} \dot{\sim} N(\mu, \frac{\sigma^2}{n})$$

where

$$\mu = error_D(h) \quad \text{and}$$

$$\frac{\sigma^2}{n} = \frac{error_D(1-error_D)}{n} \approx \frac{error_S(1-error_S)}{n}.$$

- Normal distribution

. The probability density function is given by

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}.$$

. The mean value of X: $E[X] = \mu$.

. The variance of X: $Var(X) = \sigma^2$

. The standard deviation of X: $\sigma_X = \sigma$.

- Calculating confidence intervals

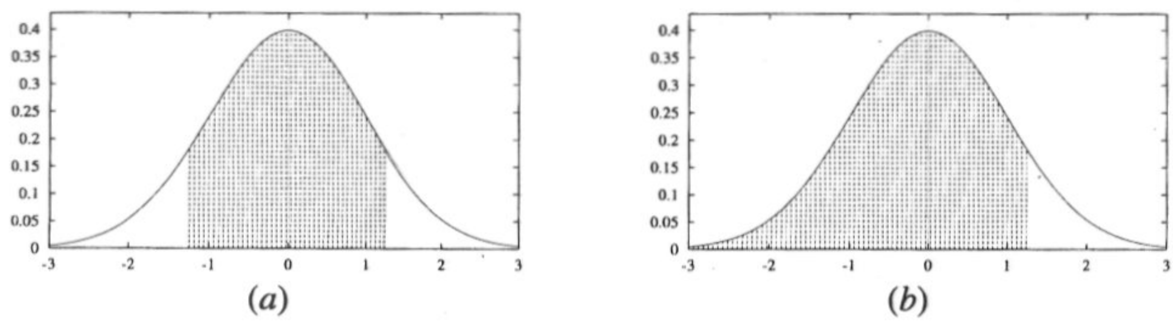

FIGURE 5.1
A Normal distribution with mean 0, standard deviation 1. (*a*) With 80% confidence, the value of the random variable will lie in the two-sided interval [−1.28, 1.28]. Note $z_{.80} = 1.28$. With 10% confidence it will lie to the right of this interval, and with 10% confidence it will lie to the left. (*b*) With 90% confidence, it will lie in the one-sided interval [−∞, 1.28].

. $N\%$ of area (probability) lies in $\mu \pm z_N \sigma$.

Values of z_N for two-sided $N\%$ confidence intervals:

$N\%$	50%	68%	80%	90%	95%	98%	99%
z_N	0.67	1.00	1.28	1.64	1.96	2.33	2.58

eg. 95% of area lies in $\mu \pm 1.96\sigma$.

Let $\hat{\mu}$ is an estimator of μ and

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

where X_is are i. i. d. random variables having mean $\mu = p$ and variance $\sigma^2 = p(1-p)$. Then,

$$\hat{\mu} \dot{\sim} N(\mu, \frac{\sigma^2}{n}).$$

Let us make a unit (or standard) normal distribution of $\hat{\mu}$:

$$\frac{\hat{\mu} - \mu}{\sigma / \sqrt{n}} \dot{\sim} N(0, 1).$$

This implies that

$$-1.96 < \frac{\hat{\mu} - \mu}{\sigma/\sqrt{n}} < 1.96$$ with the probability of 0.95.

Due to the symmetry of normal distribution,

$$-1.96 < \frac{\mu - \hat{\mu}}{\sigma/\sqrt{n}} < 1.96.$$

Therefore, we get

$$\hat{\mu} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \hat{\mu} + 1.96\frac{\sigma}{\sqrt{n}}$$

where $\sigma = \sqrt{p(1-p)}$.

-> True mean μ lies in $\hat{\mu} \pm 1.96\frac{\sigma}{\sqrt{n}}$ with the probability of 0.95.

In general, if $\hat{\mu} \sim N(\mu, \sigma^2)$,

the $N\%$ confidence interval of $\hat{\mu}$: $\hat{\mu} \pm z_N \sigma$

-> With $N\%$ probability, μ lies in interval $\hat{\mu} \pm z_N \sigma$.

The sample error is given by

$$error_S(h) = \frac{X}{n} \dot{\sim} N(\mu, \frac{\sigma^2}{n})$$

where

$\mu = error_D(h)$ and

$$\frac{\sigma^2}{n} = \frac{error_D(1 - error_D)}{n} \approx \frac{error_S(1 - error_S)}{n}.$$

With approximately 95% probability, $error_D(h)$ lies in interval

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

example.
Hypothesis h misclassifies 50 of the 100 samples in S.

In this case,

$$error_S(h) = \frac{50}{100} = 0.50 \quad \text{and}$$

$$Var(error_S(h)) = \frac{0.5 \cdot 0.5}{100}.$$

Then, with approximately 95% probability, $error_D(h)$ lies in interval

$$0.50 \pm 1.96\sqrt{\frac{0.50 \cdot 0.50}{100}} = 0.50 \pm 0.098.$$

That is, the 95% confidence interval of $error_S(h)$ is

$$0.50 \pm 0.098.$$