

# Intro. to machine learning (CSI 5325)

## Lecture 10: evaluating hypotheses

Greg Hamerly

Spring 2008

Some content from Tom Mitchell.

- 1 Quick probability lesson
- 2 Sample error, true error
- 3 Estimators
- 4 Confidence intervals for observed hypothesis error
- 5 Binomial distribution, Normal distribution, Central Limit Theorem

# Random variables

A random variable is a name for a stochastic (aka random) process.

A random variable has an associated probability distribution.

An outcome of a random variable is the observation of its value.

Examples of random variables:

- whether a coin lands head up or down when tossed
- number of heads in 100 coin tosses
- number of students who show up late to data structures
- the examples in a randomly generated training set

# Random variable transformations

We can change random variables to get new ones:

- sum up 5 random variables:  $R = A + B + C + D + E$
- divide a random variable by 8:  $S = F/8$
- etc.

The properties of the new random variables  $(R, S)$  can be derived from the random variables used to make them  $(A, B, C, D, E, F)$ .

## Expected values

The expected value of a random variable  $X$ , written  $E[X]$ , is its 'average' value.

For discrete random variable  $X$  with possible values  $x_1, x_2, \dots, x_n$ :

$$E[X] = \sum_{i=1}^n x_i P(X = x_i)$$

For continuous random variable  $Y$ :

$$E[Y] = \int_{-\infty}^{\infty} y P(Y = y) dy$$

# Expectation is linear

If  $A$  and  $B$  are random variables and  $R = A + B$ , then

$$E[R] = E[A + B] = E[A] + E[B]$$

## Mean, variance, standard deviation

The mean of a random variable  $X$  is its expected value:  $\mu = E[X]$

The variance of a random variable  $X$  is the expected value of the squared difference from the mean:

$$\begin{aligned}\sigma^2 = \text{Var}(X) &= E[(X - E[X])^2] \\ &= \sum_{i=1}^n (x_i - E[X])^2 P(X = x_i) \quad (\text{discrete}) \\ &= \int_{-\infty}^{\infty} (x - E[X])^2 P(X = x) dx \quad (\text{continuous})\end{aligned}$$

The standard deviation  $\sigma$  is the square-root of the variance.

## Two Definitions of Error

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

$$\text{error}_{\mathcal{D}}(h) \equiv \Pr_{x \in \mathcal{D}} [f(x) \neq h(x)]$$

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

$$\text{error}_S(h) \equiv \frac{1}{n} \sum_{x \in S} \delta(f(x) \neq h(x))$$

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

How well does  $\text{error}_S(h)$  estimate  $\text{error}_{\mathcal{D}}(h)$ ?

# Problems Estimating Error

- 1 *Bias*: If  $S$  is training set,  $error_S(h)$  is optimistically biased

$$bias \equiv E[error_S(h)] - error_{\mathcal{D}}(h)$$

For unbiased estimate,  $h$  and  $S$  must be chosen independently.

- 2 *Variance*: Even with unbiased  $S$ ,  $error_S(h)$  may still vary from  $error_{\mathcal{D}}(h)$ , due to randomness in choosing  $S$ .

# Example

Hypothesis  $h$  misclassifies 12 of the 40 examples in  $S$

$$error_S(h) = \frac{12}{40} = .30$$

What is  $error_{\mathcal{D}}(h)$ ?

# Estimators

Experiment:

- 1 choose sample  $S$  of size  $n$  according to distribution  $\mathcal{D}$
- 2 measure  $error_S(h)$

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

$error_S(h)$  is an unbiased *estimator* for  $error_{\mathcal{D}}(h)$

Given observed  $error_S(h)$  what can we conclude about  $error_{\mathcal{D}}(h)$ ?

# Confidence Intervals

If

- $S$  contains  $n$  examples, drawn independently of  $h$  and each other
- $n \geq 30$

Then

- With approximately 95% probability,  $error_{\mathcal{D}}(h)$  lies in interval

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

# Confidence Intervals

If

- $S$  contains  $n$  examples, drawn independently of  $h$  and each other
- $n \geq 30$

Then

- With approximately  $N\%$  probability,  $error_{\mathcal{D}}(h)$  lies in interval

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

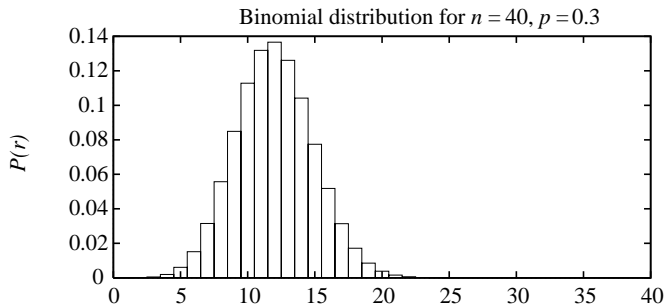
where

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

## $error_S(h)$ is a Random Variable

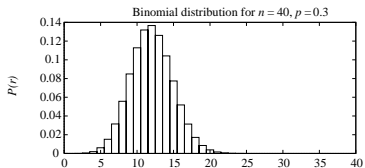
Rerun the experiment with different randomly drawn  $S$  (of size  $n$ )

Probability of observing  $r$  misclassified examples:



$$P(r) = \frac{n!}{r!(n-r)!} \text{error}_{\mathcal{D}}(h)^r (1 - \text{error}_{\mathcal{D}}(h))^{n-r}$$

# Binomial probability distribution



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

Probability  $P(r)$  of  $r$  heads in  $n$  coin flips, if  $p = \Pr(\text{heads})$

- Expected, or mean value of  $X$ ,  $E[X]$ , is

$$E[X] \equiv \sum_{i=0}^n iP(i) = np$$

- Variance of  $X$  is

$$\text{Var}(X) \equiv E[(X - E[X])^2] = np(1-p)$$

- Standard deviation of  $X$ ,  $\sigma_X$ , is

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

# Normal Distribution Approximates Binomial

$error_S(h)$  follows a *Binomial* distribution, with

- mean  $\mu_{error_S(h)} = error_{\mathcal{D}}(h)$
- standard deviation  $\sigma_{error_S(h)}$

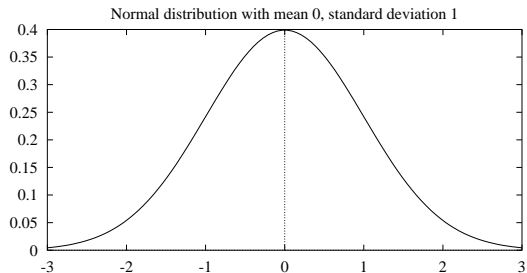
$$\sigma_{error_S(h)} = \sqrt{\frac{error_{\mathcal{D}}(h)(1 - error_{\mathcal{D}}(h))}{n}}$$

Approximate this by a *Normal* distribution with

- mean  $\mu_{error_S(h)} = error_{\mathcal{D}}(h)$
- standard deviation  $\sigma_{error_S(h)}$

$$\sigma_{error_S(h)} \approx \sqrt{\frac{error_S(h)(1 - error_S(h))}{n}}$$

# Normal (aka Gaussian) probability distribution

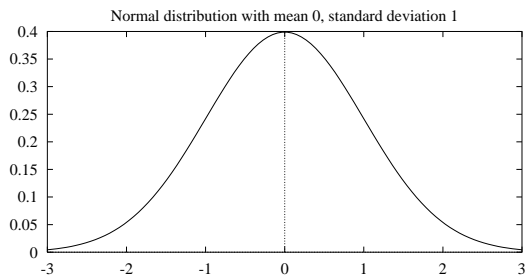


$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

The probability that  $X$  will fall into the interval  $(a, b)$  is given by

$$\int_a^b p(x) dx$$

# Normal (aka Gaussian) probability distribution



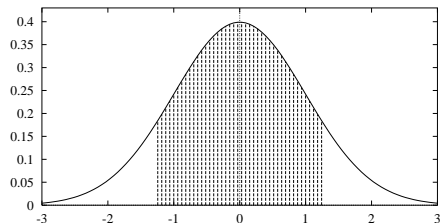
$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Expected, or mean value of  $X$ :  $E[X] = \mu$

Variance of  $X$ :  $\text{Var}(X) = \sigma^2$

Standard deviation of  $X$ :  $\sigma_X = \sigma$

# Normal (aka Gaussian) probability distribution



80% of area (probability) lies in  $\mu \pm 1.28\sigma$

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

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

# Notes on probability distribution equations

Where do the various parts come from?

Binomial:

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

Gaussian:

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

Gaussian and Binomial are part of a class of useful distributions called the exponential family.

# Notation

Using Gaussian as an example,

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

is also commonly written

$$p(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

This second notation makes it explicit that sometimes we are using data to estimate  $\mu$  and  $\sigma$  (but more on that later).

# Confidence Intervals, More Correctly

If

- $S$  contains  $n$  examples, drawn independently of  $h$  and each other
- $n \geq 30$

Then

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

$$error_{\mathcal{D}}(h) \pm 1.96 \sqrt{\frac{error_{\mathcal{D}}(h)(1 - error_{\mathcal{D}}(h))}{n}}$$

equivalently,  $error_{\mathcal{D}}(h)$  lies in interval

$$error_S(h) \pm 1.96 \sqrt{\frac{error_{\mathcal{D}}(h)(1 - error_{\mathcal{D}}(h))}{n}}$$

which is approximately

# Central Limit Theorem

Consider a set of independent, identically distributed (i.i.d.) random variables  $Y_1 \dots Y_n$ , all governed by an arbitrary probability distribution with mean  $\mu$  and finite variance  $\sigma^2$ .

Define the sample mean,

$$\bar{Y} \equiv \frac{1}{n} \sum_{i=1}^n Y_i$$

**Central Limit Theorem:** As  $n \rightarrow \infty$ , the distribution governing  $\bar{Y}$  approaches a Normal distribution, with mean  $\mu$  and variance  $\frac{\sigma^2}{n}$ .

The Central Limit Theorem states that the sum of a large number of independent, identically distributed random variables follows a distribution that is approximately Normal.

# Calculating Confidence Intervals

- 1 Pick parameter  $p$  to estimate
  - $error_{\mathcal{D}}(h)$
- 2 Choose an estimator
  - $error_S(h)$
- 3 Determine probability distribution that governs estimator
  - $error_S(h)$  governed by Binomial distribution, approximated by Normal when  $n \geq 30$
- 4 Find interval  $(L, U)$  such that N% of probability mass falls in the interval
  - Use table of  $z_N$  values