ECE8813
Statistical Natural Language Processing

Lecture 7: Corpus-Based Work and Collocation

Chin-Hui Lee

School of Electrical and Computer Engineering

Georgia Institute of Technology

Atlanta, GA 30332, USA

chl@ece.gatech.edu

Corpora: New Tools for Language Research

- LDC (USA): http://www.ldc.upenn.edu
  - WSJ (your exercise so far)
  - Spanish Gigaword
  - English Gigaword
- ELRA (Europe): http://www.elra.info
- ICAME: http://icame.uib.no
- OTA: http://ota.ahds.ac.uk
- Brown Corpus: PoS Tagging
  - http://dingo.sbs.arizona.edu/~hammond/ling696f-sp03/browncorpus.txt
  - http://www.comp.leeds.ac.uk/amalgam/tagsets/brown.html
  - http://www.edict.com.hk/textanalyser/wordlists.htm

Machine Learning Toolkits

- Netlab : neural network and Gaussian process (matlab code)
  - http://www.ncrg.aston.ac.uk/netlab/over.php
- HTK and GMTK: speech modeling kits
  - http://htk.eng.cam.ac.uk/ (HTK)
  - **http://ssli.ee.washington.edu/~bilmes/gmtk/ (GMTK)**
  - **http://www.cs.ubc.ca/~murphyk/Software/BNT/bnt.html (Bayes Net Toolbox)**
- CMU AI Repository
  - http://www.cs.cmu.edu/afs/cs/project/ai-repository/ai/areas/learning/systems/0.html
- JMLR machine learning open source software
  - http://jmlr.csail.mit.edu/mloss/
- R: http://www.r-project.org/
  - A free alternative to S-Plus developed at Bell Labs
  - If you know C, you will be right at home with R
- Weka: data mining tool in Java
  - http://www.cs.waikato.ac.nz/ml/weka/

Roles of a Corpus

- Like survey data, it becomes a key to language research
- Raw data: plenty of them, copy right issues
  - Purposes for the data collection
  - Enough data to meet research and modeling requirements?
  - Balanced or biased samples? What is representative?
  - Side information: sources of the data (speakers and writers), means the data is collected, environments in which the data is collected, minimum dispute on the transcription of data
- Meta data: tagging information, data markup
  - Additional "ground truth" depending on needs, e.g. PoS tags
  - How are the tags assigned? Are them completely defined?
  - Who provides the tagging? Expert training required? Any consistency across tagging sessions? Any potential dispute?
  - How to minimize observation noise? Data Variability?

Properties of Text Data

- Programming environment and concerns
  - What is the best programming language? Perl, Python, C?
  - What is the best text editor: TextPad
  - Unix provides plenty of command line tools: grep, wc, awk
  - Other useful data structure: tree, heap, hash, table
  - Issues with programming efficiency: memory vs. time,
  - Problem with overflow: large vector sizes, model complexity
  - Problem with underflow: small probabilities, data transformation
- Count information: a basis for estimating probabilities
  - Unobserved events: plenty in bigrams, common in trigrams
  - Equivalent classes: make counting more general
  - Mismatches in training and testing conditions
    - Missing data or description in training but needed in testing
    - Garbage collection (filler) units to “fill in the blank”

Collocation of Linguistic Events

- Collocation: an expression consisting of two or more events (e.g. words) to mean something
  - Conventional and idiomatic, e.g. broad not bright daylight
  - Frequency (raw count) as a way to signifying collocation
- Table 5.1: Raw counts of some consequent works
- Table 5.2: Some useful tagging patterns
  - (A N), (N N), (A A N), (A N N), (N A N), (N N N), (N P N)
- Table 5.3: Justeson and Katz’s PoS filter
  - Searching for the longest sequences that fits one of the PoS patterns
  - Non-compositional phrases: “last year”, “last week”, “first time”
- Table 5.4: top 20 nouns after “strong” and “powerful”
  - New York Times and other text sources

Some Useful Statistics for Collocation

- Sample mean, variance and standard deviation (s.d.)

$$\text{sample mean}: \quad \overline{X} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\text{sample variance}: \quad S^2 = \frac{1}{N-1}\sum_{i=1}^{N} (x_i - \overline{X})^2$$

$$\text{sample mean}: \quad s = \sqrt{S^2}$$

- Table 5.5: finding collocations based on simple statistics
  - Mean distance between the words “New” and “York” is 0.43
  - Mean distance between the words “editorial” and “Atlanta” is 4.03

Sampling Distributions (I)

- For many applications, it is important to obtain the distribution of a sample statistic. We need to watch for assumptions about the random samples before we work out sample distributions.
  - realize what's known and unknown
- Example 1: Normalized Sample Mean
  - independent Gaussian samples with known variance

$$\boxed{\hat{\bar{X}} = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ is Gaussian with mean } \bar{X} \text{ and variance } \frac{\sigma^2}{n}}$$

$$\boxed{Z = \frac{\hat{\bar{X}} - \bar{X}}{\sigma/\sqrt{n}} \text{ is Gaussian with mean 0 and variance 1 (standardized r. v.)}}$$

  - note: *Z* can not be defined if we don't know the parameters

Sampling Distributions (II)

- Example 2: Normalized Sample Mean
  - independent Gaussian samples with unknown variance

$$T = \frac{\hat{\bar{X}} - \bar{X}}{\tilde{S}_2/\sqrt{n}} = \frac{\hat{\bar{X}} - \bar{X}}{S_2/\sqrt{n-1}} \text{ has a } \mathit{Student's} \text{ t-distribution with n-1 degrees of freedom}$$

- The pdf of *T* (assuming *v=n-1*) is of the form

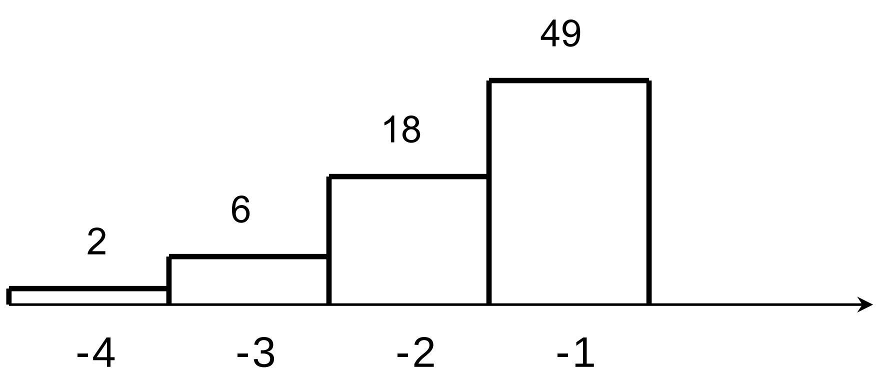
$$f_T(t) = \frac{\Gamma(\frac{v+1}{2})}{\sqrt{v\pi}\,\Gamma(\frac{v}{2})}(1+\frac{t^2}{v})^{-\frac{v+1}{2}} \quad \text{(Figure 4-2, } v=1,\ \Gamma(v) \text{ is the Gamma function)}$$

  - for large value of *v*, we have an approximate Gaussian

$$\Gamma(v+1) = v\Gamma(v),\ \Gamma(k+1) = k!\ (\text{integer } k),\ \Gamma(2)=\Gamma(1)=1,\ \Gamma(1/2)=\sqrt{\pi}$$

Some Useful Plots for Collocation

- Bar charts for position of words wrt another word
  - Figure 5.2a: “strong” vs. “opposition”: $\overline{X} = -1.15,\ \text{and}\ s = 0.67$
  - Figure 5.2b: “strong” vs. “support” (below): $\overline{X} = -1.45,\ \text{and}\ s = 1.07$
  - Figure 5.2c: “strong” vs. “for”: $\overline{X} = -1.12,\ \text{and}\ s = 2.15$
  - Variability indication and collocation discovery
  - Terminology extraction with collocation statistics



Statistical Hypothesis Testing (I)

- In essence, a hypothesis test partitions the entire observation space into two disjointed sets, *C* and *D*
- If an observation *X* lies in the region *C*, we reject *H0*; if *X* is in *D*, we accept *H0*. *C* is called the *critical region (rejection region),* often defined by critical values as discussed earlier
- *Type I error* (also called *false rejection error*) of a test:

$$\alpha = \mathrm{P}(E_1) = \mathrm{P}(X \in C \mid H_0) \Rightarrow \text{ level of significance}$$

  – Level of significance is the same as the size of critical region

- *Type II* error (also called *false alarm error*) of a test:

$$\beta = \mathrm{P}(E_2) = \mathrm{P}(X \in D \mid H_1) = 1 - \mathrm{P}(X \in C \mid H_1) = 1 - \gamma$$

Statistical Hypothesis Testing (II)

- In statistics, we normally need test a hypothesis based on some observation data. The problem is formulated as a test between two complementary hypotheses:
  - *H0: null hypothesis*
  - *H1: alternative hypothesis*
- Example: Given $X_1, X_2, \ldots, X_n$ as a random sample from a Gaussian distribution $N(\mu, \sigma^2)$, where variance σ^2 is known. We need to verify whether its mean is a given value. Thus we do hypothesis testing:
  - $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$

Center of Signal and Image Processing
Georgia Institute of Technology

Statistical Hypothesis Testing (III)

Neyman Pearson Lemma:

For a simple H_0 and simple H_1, if the distributions under both H_0 and H_1 are known, i.e., $f_0(X|\theta_0)$ and $f_1(X|\theta_1)$. Given any i.i.d. observation data $X=\{X_1,\ldots,X_T\}$, for any significance level α, the most powerful test is formulated as:

If $LR(X_1^T) = \dfrac{\prod_{t=1}^{T} f_0(X_t|\theta_0)}{\prod_{t=1}^{T} f_1(X_t|\theta_1)} > \tau$ **, accept *H_0*; otherwise reject *H_0*.**

The threshold τ is adjusted to make the significance of the test to be α. If the both pdf's have the same form, the only difference is parameters, The ratio is also called likelihood ratio (LR).

Center of Signal and Image Processing
Georgia Institute of Technology

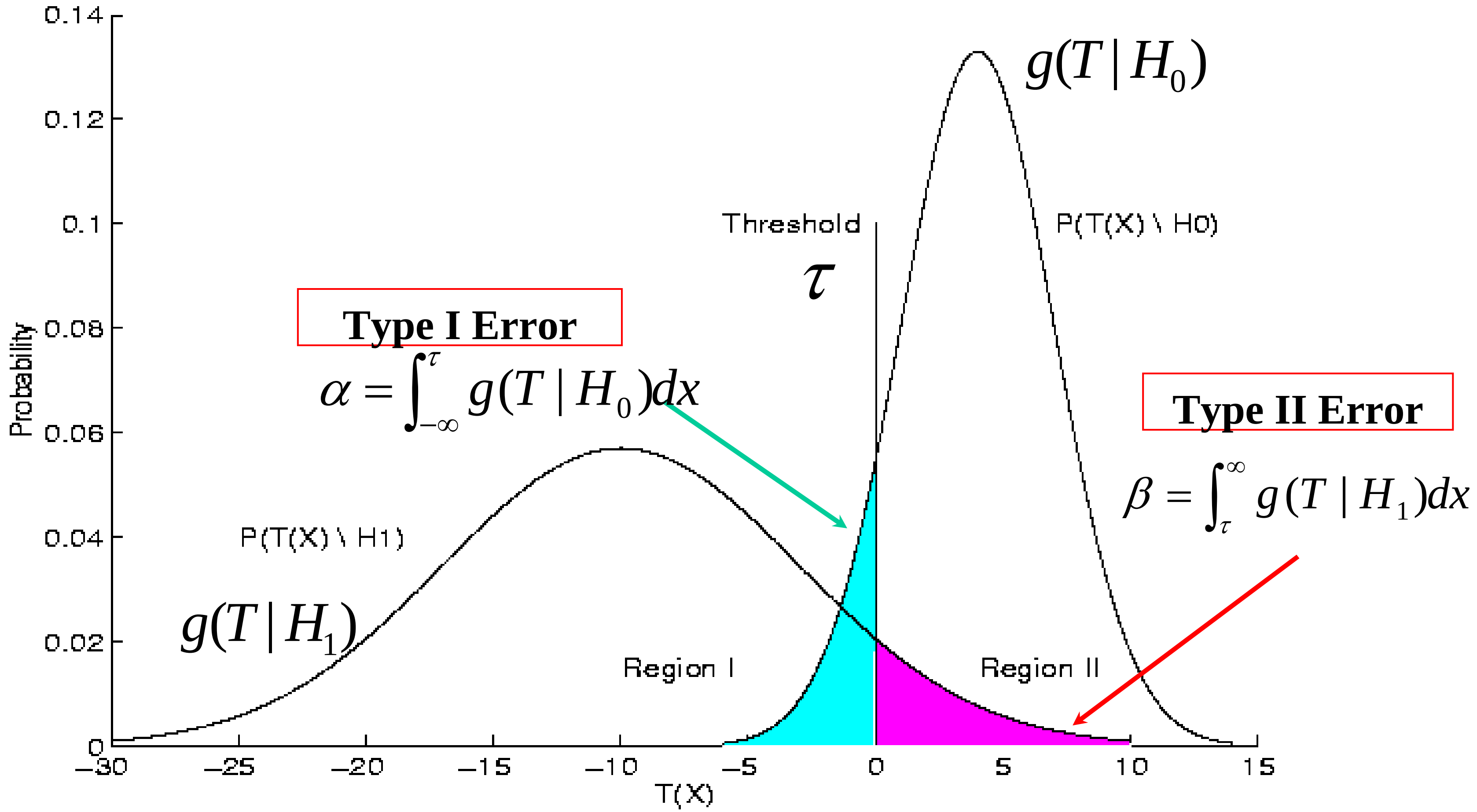
Statistical Hypothesis Testing (IV)

- The Neyman-Pearson Lemma provides a way to construct most powerful tests for simple hypotheses when the class distributions is known except for the parameter values
- How about if the hypothesis is composite
- Likelihood Ratio Test (LRT): assume the distributions are known except some parameters,

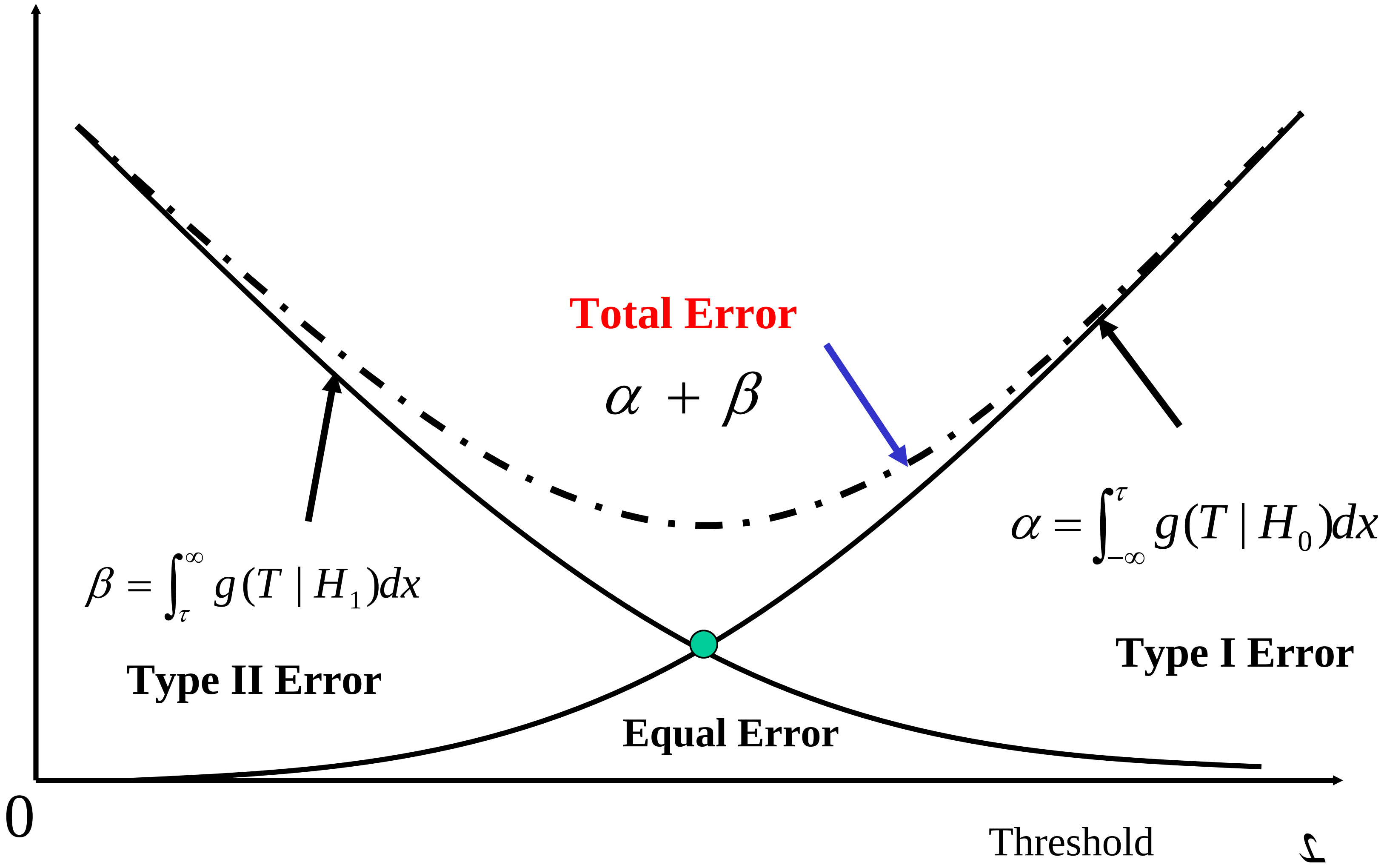
If $$T = \frac{\max_{\theta \in H_0} f_{H_0}(X \mid \theta)}{\max_{\theta \in H_1 \cup H_0} f_{H_1}(X \mid \theta)} > \tau$$ **, accept *H0*; otherwise reject *H0*.**

  - LRT is not uniformly most powerful
  - Distribution of *T* is complicated
  - Widely used for many practical applications

Distributions of Test Statistic *T*

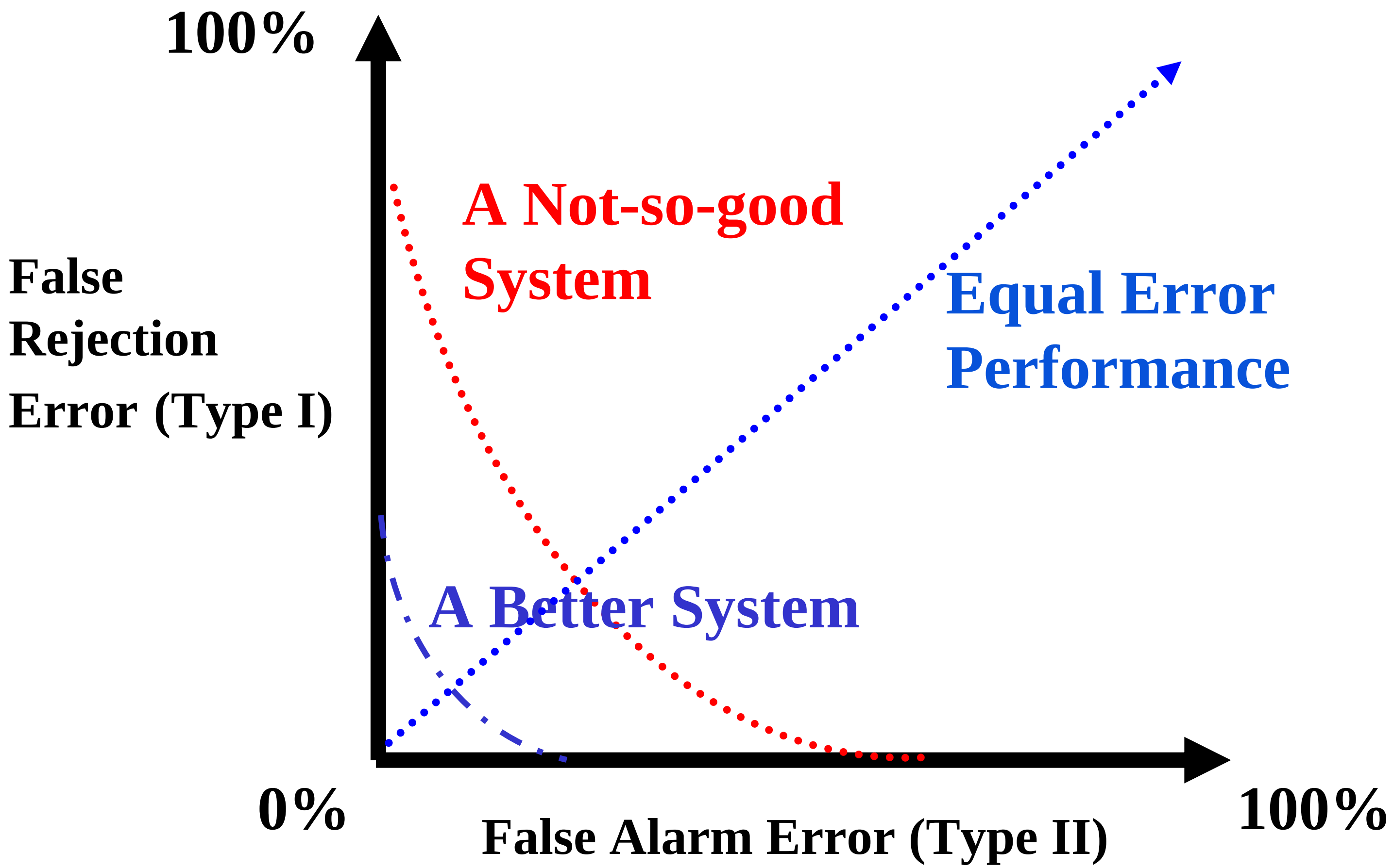


Evaluating Hypothesis Testing (I)



Center of Signal and Image Processing
Georgia Institute of Technology

Evaluating Hypothesis Testing (II): ROC (Receiver Operating Characteristic) Curve



Bernoulli Trials and Applications

- Binary Events:

$$P(A) = P(\text{"}success\text{"}) = p,\ P(\bar{A}) = P(\text{"}failure\text{"}) = q = 1 - p$$

- How about *k* successes in *n* independent trials?
  - How many such possibilities: *binomial coefficient*

$${}_nC_k = \binom{n}{k} = \frac{1}{k!}[n*(n-1)\cdots*(n-k+1)] = \frac{n!}{k!(n-k)!}$$

$$p_n(k) = P(k \text{ successes in } n \text{ trials}) = \binom{n}{k} p^k q^{(n-k)}$$

The *t* Test for Collocation Discovery

- Definition: *t*-statistic (testing against known mean)

$$t = \frac{\bar{X} - \mu}{\sqrt{S^2 / N}}$$

- An example: test of independence $P([w_1, w_2]) = P(w_1) * P(w_2)$
  - P("new")=15828/*N*, P("company")=4675/*N*, *N*=14307668
  - H0: *p*=P("new company")= P("new")*P("company")=0.0000003615, a binomial distribution with mean=*p*, and variance=*p(1-p)*
  - Sample mean= 8/*N*
  - *t*=0.999932, nor larger than the critical value of 2.576 at a significance level of 99.5%
  - Cannot reject the null hypothesis

The *t* Test for Difference Discovery

- Definition: *t*-statistic (assuming the known difference is 0)

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - \mu}{\sqrt{\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}}}$$

- Examples: Table 5.7 and text on Page 168
  - Intrinsic (e.g. strong) vs. extrinsic (e.g. powerful) properties

$$t \approx \frac{C([v^1, w]) - C([v^2, w])}{\sqrt{C([v^1, w]) + C([v^2, w])}}$$

Confidence Intervals

- Sample mean : a point estimate related to sample size
  - How about an interval estimate? How to choose *n*?
- *q*-percent confidence interval: *e.g. quartile, median*
  - Example: sample mean for Gaussian samples, known variance
  - For the sample mean: $\boxed{[\bar{X} - k\sigma/\sqrt{n}, \bar{X} + k\sigma/\sqrt{n}]}$

$$\boxed{P(\bar{X} - k\sigma/\sqrt{n} < \hat{\bar{X}} < \bar{X} + k\sigma/\sqrt{n}) = q/100}$$

- Confidence interval for other statistics can also be established if the distribution of the point estimate of interest can be evaluated (e.g. *t*-distribution).

One-Sided Test: An Example

- Testing of known Gaussian mean (known variance)

$$\text{Test statistic } z = [\bar{x} - \bar{X}]/[\sigma/\sqrt{n}] = [290 - 300]/[40/\sqrt{100}] = -2.5$$

$$\text{Accept } \bar{X} = 300 \text{ if } z > z_c \text{ with confidence } C(z_c) = \int_{z_c}^{\infty} f(z)dz = 1 - \Phi(z_c) \text{ or significance } \alpha = 1 - C(z_c)$$

$$\text{If } C(z_c) = 0.99 \Rightarrow z_c = -2.33, \text{ we reject the hypothesis } \bar{X} = 300 \text{ with 99\% confidence}$$

$$\text{and if } C(z_c) = 0.995 \Rightarrow z_c = -2.575, \text{ we accept the hypothesis } \bar{X} = 300 \text{ with 99.5\% confidence}$$

- Higher confidence level implies large acceptance region
  - a higher level of significance α implies a more severe test
- *T*-test: for smaller sample sizes (known variance)

$$\text{Test statistic } t = [\bar{x} - \bar{X}]/[\tilde{s}_1/\sqrt{n}] = [290 - 300]/[40/\sqrt{9}] = -0.75$$

$$\text{If } C(t_c) = 0.99 \Rightarrow t_c(8) = -2.896, \text{ we accept the hypothesis } \bar{X} = 300 \text{ with 99\% confidence}$$

Two-Sided Test: An Example

- Testing of known Gaussian mean (known variance)

$$\text{Test statistic } z = [\bar{x} - \bar{X}]/[\sigma/\sqrt{n}] = [10.3 - 10]/[1.2/\sqrt{100}] = 2.5$$

$$\text{Accept } \bar{X} = 10 \text{ if } -z_c < z < z_c \text{ with confidence } C(z_c) = \int_{-z_c}^{z_c} f(z)dz = 1 - 2\Phi(z_c) \text{ or significance } S(z_c) = 1 - C(z_c)$$

- *T*-test: for smaller sample sizes (known variance)

$$\text{If } C(z_c) = 0.95 \Rightarrow z_c = 1.96 \text{ (Table 4-1), we reject the hypothesis } \bar{X} = 10 \text{ with 95\% confidence}$$

$$\text{Test statistic } t = [\bar{x} - \bar{X}]/[\tilde{s}_1/\sqrt{n}] = [10.3 - 10]/[1.2/\sqrt{9}] = 0.75$$

  – small sample test is not as severe as a large sample one

$$\text{If } C(t_c) = 0.95 \Rightarrow t_c(8) = 2.306 \text{ (Table 4-2), we accept the hypothesis } \bar{X} = 10 \text{ with 95\% confidence}$$

- Critical Value: z_c and t_c are critical values of the tests

One- and Two-Sided Tests: Summary

One-sided (one-tailed) Test

$$H_0: \bar{X} = \mu_0 \text{ vs. } H_1: \bar{X} = \mu_1 > \mu_0$$

- *Large-sample test statistic:*

  $$z \approx (\bar{x} - \mu_0)/(S_2/\sqrt{n})$$

- *Small-sample test statistic:*

  $$t = (\bar{x} - \mu_0)/(S_2/\sqrt{n})$$

- Region of Rejection

$$z > z_\alpha \ (z < -z_\alpha) \text{ and } t > t_\alpha \ (t < -t_\alpha)$$

$$P(z > z_\alpha) = \alpha \text{ or } P(t > t_\alpha) = \alpha$$

Two-sided (two-tailed) Test

$$H_0: \bar{X} = \mu_0 \text{ vs. } H_1: \bar{X} = \mu_1 \neq \mu_0$$

- *Large-sample test statistic:*

  $$z \approx (\bar{x} - \mu_0)/(S_2/\sqrt{n})$$

- *Small-sample test statistic:*

  $$t = (\bar{x} - \mu_0)/(S_2/\sqrt{n})$$

- Region of Rejection

$$z > z_{\alpha/2} \text{ or } z < -z_{\alpha/2}$$

$$\text{and } t > t_{\alpha/2} \text{ or } t < -t_{\alpha/2}$$

$$P(z > z_{\alpha/2}) = \alpha/2 \text{ or } P(t > t_{\alpha/2}) = \alpha/2$$

Chi-Square Distributions

- Chi-Square: sum of square iid N(0,1) random variables

$$X^2 = Y_1^2 + Y_2^2 + \cdots + Y_n^2 \text{ with } Y_1, \ldots, Y_n \text{ iid N(0,1) r.v.}$$

X^2 is said to be Chi-square with n degree of freedom: $\chi^2(n)$

$$f_{\chi^2}(u) = \frac{u^{\frac{n}{2}-1}}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}\exp[-\frac{u}{2}],\ u \geq 0$$

Show that $\overline{U} = n$ and $\mathrm{Var}(U) = 2n$

- Implications: with proper normalization
  - Power random variable *W* is $\chi^2(1)$
  - Squared Rayleigh random variable R^2 is $\chi^2(2)$
  - Squared Maxwell random variable V^2 is $\chi^2(3)$

Pearson's Chi-Square Test

- Definition: *X-square* statistic (testing of variances)

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} : \begin{array}{l} O_{ij} : \text{observed count} \\ E_{ij} : \text{expected count} \end{array}$$

$$X^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})}{(O_{11}+O_{12})(O_{11}+O_{21})(O_{12}+O_{22})(O_{21}+O_{22})}$$

- An example: Table 5.8 (2x2 table)
  - H0: P("new company")= P("new")*P("company")
  - Show the above as indicated in Exercise 5.9
  - A *X*-square value of 1.55 is too small compared to the critical value of 3.841 at a significance level of 95% (chi-square distribution with one degree of freedom for a 2x2 table)
  - Cannot reject the null hypothesis the two words are independent

Likelihood (Probability) Ratio Test

- Definition: *LR*-statistic or log LR (*LLR*-statistic)

$$PR = \frac{P(H_0)}{P(H_1)} \text{ or } LLR = \log \frac{L(H_0)}{L(H_1)}$$

- An example:
  - H0: $p = p_1 = p_2$
  - H1: $p_1 \neq p_2$

$$p_1 = P(w_2 \mid w_1), p_2 = P(w_2 \mid \overline{w}_1)$$

$$p = \frac{c_1}{N}, p_1 = \frac{c_{12}}{c_1}, p_2 = \frac{c_2 - c_{12}}{N - c_1}$$

- An example: binomial distribution for H0 and H1

$$B(r; n, p) = \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r} \quad \text{where} \quad 0 \le r \le n$$

Log Likelihood Ratio Test

- Definition: *LLR*-statistic, asymptotically chi-square

$$
\begin{aligned}
LLR &= \log \lambda = \log \frac{L(H_0)}{L(H_1)} \\
&= \log L(c_{12}, c_1, p) + \log L(c_2 - c_{12}, N - c_1, p) \\
&- \log L(c_{12}, c_1, p_1) - \log L(c_2 - c_{12}, N - c_1, p_2) \\
L(k, n, r) &= r^k (1 - r)^{n-k}
\end{aligned}
$$

- Table 5.12: “computers” is more likely to follow “powerful” than other words 1.3*10**18
- Table 5.13: relative frequency ratio
  - Comparing general text with subject-specific text corpora

Center of Signal and Image Processing
Georgia Institute of Technology
CSIP

Correlation between Two Sets of Data

- Linear correlation coefficient (Pearson's *r*)

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2} * \sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \quad \text{with } \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \; \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

- Pearson's *r* approaches Gaussian for large *n*
  - significance of the value of *r*: small *r* is often meaningless unless the sample size *n* is large, and *f(x, y)* is known
  - large *r* implies a tighter coupling between *X* and *Y*

Curve Fitting

- Consider fitting *y=r(x)* to a set of pairs of random samples: $\{(x_1, y_1),(x_2, y_2),\ldots,(x_n, y_n)\}$
  - we will have curve fitting errors: d(i)
  - *r(.)* is a regression function $r(x) = \sum_{k=0}^{p} a_k x^k$
  - goodness of fit: minimizing least squared errors $D = \sum_{t=1}^{n} d_i^2$

- Polynomial fitting (MATLAB example):
- Linear fitting: *y=a+bx*
- Spline fitting
  - local and global optimization
  - various optimization criteria

Linear Regression

- Least Squares: Minimizing Sum of Squared Error

$$D = \sum_{t=1}^{n} d_i^2 = \sum_{t=1}^{n} [y_i - (a + bx_i)]^2 = \text{minimum}$$

- We obtain the following matrix normal equation

$$\boxed{\frac{\partial D}{\partial a} = 0 \Rightarrow \sum_{t=1}^{n} y_i = an + b\sum_{t=1}^{n} x_i, \quad \frac{\partial D}{\partial b} = 0 \Rightarrow \sum_{t=1}^{n} x_i y_i = a\sum_{t=1}^{n} x_i + b\sum_{t=1}^{n} x_i^2}$$

- Solving for intercept *a* and slope *b* : y=polyfit(y,x,n)

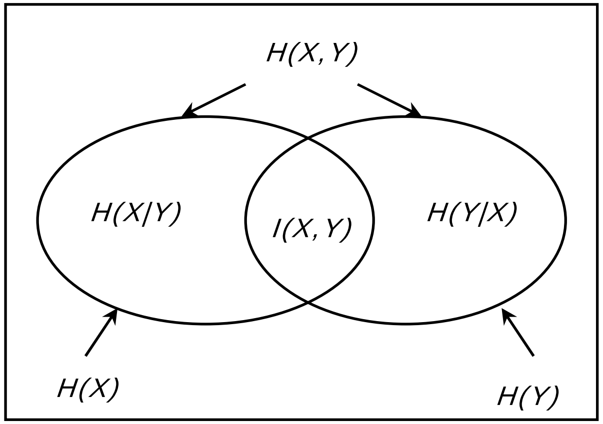
$$\boxed{\hat{b} = \frac{n\sum_{t=1}^{n} x_i y_i - (\sum_{t=1}^{n} x_i)(\sum_{t=1}^{n} y_i)}{n\sum_{t=1}^{n} x_i^2 - (\sum_{t=1}^{n} x_i)^2}, \hat{a} = \frac{(\sum_{t=1}^{n} y_i)(\sum_{t=1}^{n} x_i^2) - (\sum_{t=1}^{n} x_i)(\sum_{t=1}^{n} x_i y_i)}{n\sum_{t=1}^{n} x_i^2 - (\sum_{t=1}^{n} x_i)^2} = \frac{\sum_{t=1}^{n} y_i - \hat{b}\sum_{t=1}^{n} x_i}{n} = \hat{\bar{Y}} - \hat{b}\hat{\bar{X}}}$$

- Extend to more than one regressor (econometrics)

Mutual Information

Definition :

$$I(X,Y) = H(X) - H(X|Y)$$
$$= H(Y) - H(Y|X)$$
$$= H(X) + H(Y) - H(X,Y)$$



$$I(X,Y) = \sum_{x\in X} p(x)\log_2 \frac{1}{p(x)} + \sum_{y\in Y} p(y)\log_2 \frac{1}{p(y)} - \sum_{x\in X}\sum_{y\in Y} p(x,y)\log_2 \frac{1}{p(x,y)}$$

$$I(X,Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log_2 \frac{p(x,y)}{p(x)p(y)} \qquad i(x,y) = \log_2 \frac{p(x,y)}{p(x)p(y)}$$

Note: Eqs. (5.11)-(5.13) are point-wise mutual information *i*(.)

Point-wise Mutual Information

- Point-wise MI: the amount of information provided by the occurrence of the event represented by "*y*" about the occurrence of the event represented by "*x*"

$$i(x, y) = \log \frac{P(x \mid y)}{P(x)}$$

- *i*("*Ayatollah*", "*Ruhollah*") = 18.38 bits (Table 5.14)
- Table 5.15: collocation of "strength" and "power"
  - Larger corpus gives better estimate of mutual information
  - Many word pair only occurs once even in large corpora

Other Topics of Interest

- We did not have time to cover the following:
  1. Comparing two samples means (mean difference): for sampling distributions, confidence interval and hypothesis testing
  2. Multiple Regression (macroeconomics)
  3. Autoregression: Time Series (econometrics)
  4. Parameter Estimation
  5. Decision Theory
- Basic skills learned here can be applied to
  – The above and many other problems

Summary

- Today's Class
  - Corpus-based work and collocation
    - Some useful statistics for collocation evaluations
    - Statistical hypothesis testing: a useful tool
  - Lab1 due on Jan. 27
- Next Classes
  - *N*-gram estimation (Jan. 29 and Feb. 4)
- Reading Assignments
  - Manning and Schutze, Chapters 3, 4, 5 & 6
  - Reading M&S is critical because of the examples cited