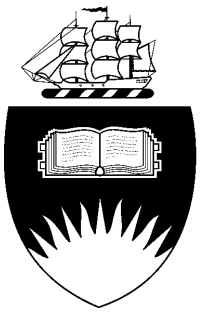

FLINDERS UNIVERSITY
ADELAIDE • AUSTRALIA

ALTS2003: Validation & Evaluation L3 Evaluation

	+ R	– R			+ R	– R	
+ P	C_{11}	C_{10}	$\|P_1\|$	+ P	E_{11}	E_{10}	$\|P_1\|$
– P	C_{01}	C_{00}	$\|P_0\|$	– P	E_{01}	E_{00}	$\|P_0\|$
	$\|R_1\|$	$\|R_0\|$	N		$\|R_1\|$	$\|R_0\|$	N

Common Evaluation Measures based on IR model

- Recall $\rho = C_{11}/|R_1| = p_P(1|1)$ (probability correct classes predicted)
- Precision $\pi = C_{11}/|P_1| = p_R(1|1)$ (probability predicted classes correct)
- Inverse Recall $\rho_0 = C_{00}/|R_0| = p_P(0|0)$ (prob. incorrect class rejected)
- Inv. Precision $\pi_0 = C_{00}/|P_0| = p_R(0|0)$ (prob. rejected class incorrect)
- Fallout $\varphi = \rho_0 = C_{10}/|R_0| = p_R(0|1)$ (prob. incorrect class predicted)

- M-Measure $M = \alpha\pi + (1-\alpha)\pi$ (Arithmetic mean of recall & precision)
- G-Measure $G = (\pi^{\alpha} \cdot \rho^{(1-\alpha)})$ (Geometric mean of recall & precision)
- F-Measure $F = 1 / [\alpha/\pi + (1-\alpha)/\rho]$ (Harmonic mean of recall&precision)
- E-Measure E = 1-F (Effectiveness measure – lower is better)
- D-Measure $D = 1 - 1 / [1/\pi + 1/\rho - 1]$ (? – lower is better)

- Interpolated K-precision: estimated for specific levels of $\rho = K\%$
- Absolute K-precision: determined at specific levels of $C_{11} = K$
- R-precision: determined at $|P_1| = |R_1|$ (correct number predicted)

- Precision-Recall plots (based on Interpolated or Absolute K-precision)
- Recall-Fallout plots (chance level difference is expected to be zero)

powers@acm.org

Common Evaluation Measures from Statistics

Rand Accuracy

- In many applications we are interested in both + and − labeling accuracy
- We can weight Prec and Inv Prec by the respective number of predictions
- We can weight Rec and Inv Rec by the respective number of documents
- In both cases we get the Rand Accuracy: $(C_{11} + C_{00}) / N$

Tanimoto or Jaccard Index

- In other cases we are more interested in + than − labels
- We thus want to exclude correctly labeled −ve cases from consideration
- And we get the Tanimoto or Jaccard Index $C_{11} / (N - C_{00})$

Error Rate

- The error rate $(C_{10} + C_{01}) / N$ is the complement of Rand Accuracy

Receiver Operating Characterstics (ROC)

- Tradeoff of standard statistics is common in Medicine via ROC analysis
- Precision-like indicators are used referenced to predictions made
- True positive rate $C_{11}/ |P_1| = \pi$ is identical with Precision
- False positive rate $C_{10}/ |P_1|$ is often confused with Fallout
- ROC usually plots *true positive* rate against *false positive* rate
- The graph is drawn by varying a threshold or *sensitivity*
- The ROC curve is commonly used to choose the *operating point*

- The *area under* the ROC *curve* (AUC) is used to choose classifiers
- It can be shown to be a function of the Wilcoxon statistic (*significance*)
- The best *operating point* need not belong to the highest AUC classifier

- Usually the assumption is equal cost of false positives and false negatives
- A common implicit but in general false assumption is $|R_1|/N=|R_0|/N=0.5$
- Under these two assumption chance performance defines a 45° ROC line
- Equal cost (*isocost*) lines run parallel to the chance ROC line
- The best operating point lies on the *isocost* line nearest (TPR,FPR)=(1,0)

- ROC curves are always monotonic but not always convex
- It is possible to operate a classifier on the convex hull of the ROC
- Random and functional combinations of ROC classifiers can be better
- ROC provides a useful but not always optimal basis for classifier fusion

 powers@acm.org

Cost: Profit and Loss

- Significance measures how *unusual* a pattern is, not how *useful*
- Recall & Precision & ROC AUC do not place a cost on *errors*

- Base level due to chance for recall is the probability of **+ P**, $|P_1|/N$
  e.g. if I always guess **+ P** I will return all **+ R** cases: $\rho = 100\%$
- Base level due to chance for precision is probability of **+ R**, $|R_1|/N$
  e.g. if all cases are **+ R** all **+ P** guesses will be correct: $\pi = 100\%$

- Techniques based on financial problems do take into account costs

A Trading Edge

- In buying and selling commodities we measure success in dollars
- In an efficient market successful speculation should be impossible
- What goes up must come down and what A gains B loses
- A successful long term trader hopes for unbounded growth
- A successful short term trader hopes for an untapped edge
- Exploitation of an inefficiency/edge should lead to its elimination

- A NN/ML model generally assumes equilikely rises and falls
- Prices are determined by the market or by a market maker
- Every trade results in a profit or a loss – with a 50:50 *a priori* chance
- Costs of trading and bid/offer spread are a cost against profits & losses
- There is no real reason to expect past patterns to continue in the future

A Fair Bookmaker

- A bookmaker sets odds based on previous performance/expectations
- A fair bookmaker's odds accurately reflect the chances of winning
- Fair bookmakers don't exist – odds are discounted to make a living
- Fair bookmakers can't exist – past performance ≠ future performance

- Fair odds for our contingency table are $|R_0| : |R_1|$ for betting on **+ P**
- If you bet $|R_0|$ on **+ P** you stand to win $|R_1|$ on **+ R** or lose $|R_0|$ on **– R**

- Fair odds for our contingency table are $|R_1| : |R_0|$ for betting on **– P**
- If you bet $|R_1|$ on **– P** you stand to win $|R_0|$ (**– R**) or lose $|R_1|$ (**+ R**)

- An optimal strategy will bet in accordance with the odds
- An optimal evaluation will cost in accordance with the odds
- Costing based on the odds is thus reasonable when costs are unknown

 powers@acm.org

The Bookmaker Algorithm

Binomial/binary case

- The amount at risk (the *ante*) is the sum of what you can win or lose
  viz. Your bet is at risk but so is the bookmaker's complementary bet
  Note: This is like a pool or *ante* in a poker game – all put in their bets

- Your probability of winning on + P is $p_R(1)=|R_1|/N$
- Your probability of losing on + P is $p_R(0)=|R_0|/N$

- The + P odds of $|R_0|$: $|R_1|$ can be scaled to $A:B = 1/p_R(1) : 1/p_R(0)$
  so your expected win is $\$B \cdot p_R(1) = \$1$ if you bet $A=1/p_R(1)$ on + P
  and your expected loss is $-\$A \cdot p_R(0) = -\$1$
- This illustrates that the odds are fair and the game is zero-sum
- It also shows that we expect a Bookmaker score of $0 if we guess

- Unlike significance tests N is irrelevant:
  we want a score of 100% if we predict perfectly for all N trials
  and 0% if we are just guessing over a sufficiently large N trials
  and in fact we also achieve −100% if we predict wrongly for N trials
- Hence we normalize our contingency matrix by N to use probabilities
  viz. $p_{PR}(i,j) = C_{ij} / N$ with margins $p_P(i)=|P_i|/N$ and $p_R(j)=|R_j|/N$
- Now our Bookmaker formula reflects us betting P_i with $p_P(i)$,
  winning $w(i|i)=1/p_R(i)$ or losing $w(j|i)= -1/p_R(j)$, $i \neq j$ with $p_{PR}(i,j)$
  viz. $\text{Gain} = \Sigma_{ij}\, p_P(i) \cdot p_{PR}(i,j) \cdot w(j|i)$

Multinomial case

- When we bet on a horse we lose a fixed amount if *any* other horse wins
  viz. win $w(i|i)=1/p_R(i)$ or lose $w(j|i)= -1/(1-p_R(i))$, $i \neq j$ with $p_{PR}(i,j)$
  and again $\text{Gain} = \Sigma_{ij}\, p_P(i) \cdot p_{PR}(i,j) \cdot w(j|i)$

- Note that in the multinomial case we can't directly interpret a loss
  as the amount of the loss depends on which alternative we choose

Alternate formulation

- The above gain formula sets independent of your predictions
- An equivalent formula adds a dependency and makes it an average
  viz. $\text{Gain} = \Sigma_{ij}\, p_{PR}(i,j) \cdot w^*(j|i)$
  where $w^*(j|i) = p_P(i) \cdot w(j|i)$

 powers@acm.org

Informedness

- The Bookmaker formula measures '*informedness'*
- If we guess G% of the time then it returns Gain = G%
- This is equivalent to using the perfect decision matrix G% of the time and the random contingency matrix $p_{PR}(i,j) = E_{ij} / N$ at other times

- The Bookmaker formula is unique in having this property
- This follows from considerations of linearity

Information

- Informedness should not be confused with Information
- $Log_2(Gain)$ in fact reflects the conditional information of prediction *viz.* information about the actual *correct* class being predicted

- Finding how much information is in a pattern is akin to significance
- Whether the information is useful is a different question
- We are specifically interested in *correct* information!

- The formula has the same form as the standard conditional entropy, an information theoretic measure based on *precision*
  viz. $Info = \Sigma_{ij}\, p_{PR}(i,j) \cdot h(j|i)$
  where $h(j|i) = \log_2 p_R(j|i)$

- This tells us only about the information contained in the pattern
- It doesn't tell us whether it is correct or not

- Inverting + P and – P in a perfect contingency matrix makes it wrong but it doesn't affect the result of the Info computation

EXCEL spreadsheets are available to illustrate Bookmaker and various other significance and accuracy measures

A Matlab/Octave script is available to compute Bookmaker and various other accuracy measures

 powers@acm.org

Examples

Trading Edge – Day of Week, Day of Month
AVSR
EEG

References

B F Buxton, W B Langdon and S J Barrett, Data Fusion by Intelligent Classifier Combination, Measurement and Control, vol 34, No. 8, October 2001, p229-234 [http://www.cs.ucl.ac.uk/staff/W.Langdon/mc/]

W B Langdon, Receiver Operating Characteristics (ROC) [http://www.cs.ucl.ac.uk/staff/W.Langdon/roc/]

[Entw98a] Jim Entwisle and David M. W. **Powers**, "*The Present Use of Statistics in the Evaluation of NLP Parsers*", pp215-224, **NeMLaP3/CoNLL98 Joint Conference**, Sydney, January 1998. (http://www.cs.flinders.edu.au/Research/AI/papers/199801a-CoNLL-USE.pdf)

[Huan01a] Huang, J. and D. M. W. **Powers** (2001). *Large-scale Experiments on Correction of Confused Words,* pp77-82, **Australian Computer Science Conference, Bond University, Queensland AUS** (http://www.cs.flinders.edu.au/Research/AI/papers/200101-ACS-CCW.pdf)

[Lewi01a] Lewis, T. W. and D. M. W. **Powers** *Lip Feature Extraction using Red Exclusion*. In P. Eades and J. Jin (eds), **CRPIT: Visualisation 2000, vol 2**.

[Lewi03a] Lewis, T. W. and D. M. W. **Powers** (2003). *Audio-Visual Speech Recognition using Red Exclusion and Neural Networks.* **Journal of Research and Practice in Information Technology 35#1:**41-64. (http://www.cs.flinders.edu.au/Research/AI/papers/200301-JRPIT-AVSRNN.pdf)

[Li03a] Yan Li, David **Powers** and Kenneth Pope (2003). *A new approach to blind signal deconvolution using recurrent neural networks.* **International Journal of Knowledge-Based Intelligent Engineering Systems 7**#2:62-69.

[Powe89a] David M. W. **Powers** and Christopher Turk, **Machine Learning of Natural Language**, Research Monograph, Springer-Verlag, 1989, ISBN 3-540-19557-2/0-387-19557-2

[Powe91c] David M. W. **Powers** and Larry Reeker, eds., **Proceedings of the AAAI Spring Symposium on Machine Learning of Natural Language and Ontology**, Document D-91-09 (205pp), DFKI, Univ. Kaiserslautern FRG.

[Powe98c] David M. W. **Powers**, "*Reconciliation of Unsupervised Clustering, Segmentation and Cohesion*", pp307-310, **NeMLaP3/CoNLL98 Paradigms and Grounding in Language Learning Workshop**, Adelaide, January 1998. (http://www.cs.flinders.edu.au/Research/AI/papers/199801e-PaGiLL-UCSC.pdf)

[Powe03a] David M. W. **Powers** (2003). *Recall and Precision versus the Bookmaker.* **International Conference on Cognitive Science**, University of New South Wales, July 2003. (pp529-534 http://www.cs.flinders.edu.au/Research/AI/papers/200302-ICCS-Bookmaker.pdf + poster + matlab/octave code + binary and ternary excel spreadsheets)

http://www.cs.flinders.edu.au/People/David_Powers/

 powers@acm.org