Not All Errors Are Created Equal – Error Discrimination May Be Good & Necessary

B.H. Juang
Georgia Institute of Technology

Georgia Institute of Technology

CSIP Center for Signal & Image Processing

Errors That Make Sense?

```
O   AT   N.   E.   C.   THE   NEED   FOR
   2.317 3.138 3.135 2.784  1.275  3.675  2.027
     INTERNATIONAL MANAGERS WILL KEEP RISING
       3.259          3.797      2.481  3.689  3.925
```

1. AT *ANY* < *del* > *SEE* THE NEED FOR INTERNATIONAL MANAGERS WILL KEEP RISING
2. AT N. E. C. < *del* > *NEEDS* FOR INTERNATIONAL *MANAGER'S* WILL KEEP RISING

- Two decoded sentences have identical (conventional) error rate;
- If error cost is defined as mis-information, E1=27.25%, E2=25.24%; i.e., 2nd sentence has less mis-information and is better

$$\mathrm{E} = \frac{\sum_{s=1}^{S}[-\ln P(w_s)] + \sum_{d=1}^{D}[-\ln P(w_d)] + \sum_{i=1}^{I}[-\ln P(w_i)]}{\sum_{n=1}^{N}[-\ln P(w_n)]}$$

Georgia Institute of Technology

Do Not Take Max A Posteriori For Granted

Bayes Decision Theory

$$R(C_i \mid X) = \sum_{j=1}^{M} e_{ij} P(C_j \mid X) \qquad e_{ij} = \begin{cases} 1, & i \neq j \\ 0, & i = j \end{cases} \qquad R(C_i \mid X) = \sum_{j,\, j \neq i}^{M} e_{ij} P(C_j \mid X) = 1 - P(C_i \mid X)$$

$$\mathcal{L} = E\{R(C(X) \mid X\} = \int R(C(X) \mid X)\, p(X)\, dX$$

MAP: $C(X) = \arg\min_i R(C_i \mid X) = \arg\min_i \{1 - P(C_i \mid X)\} = \arg\max_i P(C_i \mid X)$

What if

$$[e_{ij}] = \begin{bmatrix} 0 & 0.1 & 0.5 \\ 0.1 & 0 & 0.5 \\ 0.1 & 0.1 & 0 \end{bmatrix} \qquad [P(C_j \mid X)] = \begin{bmatrix} 0.35 \\ 0.4 \\ 0.25 \end{bmatrix} \qquad [R(C_i \mid X)] = \begin{bmatrix} 0.165 \\ 0.16 \\ 0.075 \end{bmatrix}$$

$$C_2 = C_{MAP} = \arg\max_{i=1,2,3} P(C_i \mid X)$$

$$C_3 = \arg\min_{i=1,2,3} R(C_i \mid X) = \arg\min_{i=1,2,3} \sum_{j=1}^{M} e_{ij} P(C_j \mid X)$$

Georgia Institute of Technology

How to Learn from Data with Non-U Error Cost?

- Conventional MAP is no longer valid; the usual distribution estimation approach has lots of pitfalls (even when error cost is uniform);
- Learning now involves not only the correct label but also what kind of error the given recognizer is going to make for each and every token;
- The minimum classification error (MCE) framework is the right candidate for this extension or generalization.

Georgia Institute of Technology

Learning w/ Non-Equal Error Significance

Empirical error cost:

$$L = \frac{1}{N}\sum_{X\in\Omega} e_{i_X j_X} = \frac{1}{N}\sum_{X\in\Omega}\sum_{i\in I_M}\sum_{j\in I_M} e_{ij}\,1[j_X = j]\,1\left[i = \arg\max_k g_k(X;\Lambda)\right]$$

Smooth embedding for parameter optimization

$$\sum_{i\in I_M} e_{ij}\,1\left[i = \arg\max_k g_k(X;\Lambda)\right] \approx \sum_{i\in I_M} e_{ij}\,\frac{g_i(X;\Lambda)}{G(X;\Lambda)}$$

$$G(X;\Lambda) = \left[\sum_{i\in I_M} g_i^{\eta}(X;\Lambda)\right]^{1/\eta}$$

Other forms of approximation are possible.

Then, optimize parameters to achieve

$$\min_{\Lambda}\tilde{L}(\Lambda) = \min_{\Lambda}\left\{\frac{1}{N}\sum_{X\in\Omega}\sum_{i\in I_M}\sum_{j\in I_M} e_{ij}\,1[j_X = j]\,\tilde{1}\left[i = \arg\max_k g_k(X;\Lambda)\right]\right\}$$

Terminology

Ground Truth – cost-free decision

Bayes (Minimum) Cost – expected cost when $i_X = j_X$

$$\mathcal{L}_B = E\left\{\sum_{j\in I_M} e_{j_X j} P(j\mid X)\right\}$$

Empirical Bayes Cost – also $i_X = j_X$

$$L_B = \frac{1}{N}\sum_{X\in\Omega}\sum_{j\in I_M} e_{j_X j} P(j\mid X)$$

Empirical Cost $L(\Lambda) = \frac{1}{N}\sum_{X\in\Omega}\sum_{j\in I_M}\sum_{i\in I_M} e_{ij}\,1(i_X = i)\,1(j_X = j)$

Smoothed Empirical Cost

$$\tilde{L}(\Lambda) = \frac{1}{N}\sum_{X\in\Omega}\sum_{j\in I_M}\sum_{i\in I_M} e_{ij}\,\tilde{1}(i_X = i)\,1(j_X = j)$$

Rise of Non-uniform Error Cost

- User specified cost matrix;
- Use confusion matrix as diagnostic tool and derive cost matrix to "manage" the confusion matrix (to drive undesirable errors away);
- Cross-layer modeling and decoding, e.g.
  - Word recognition using phone models – phone error sensitivity varies
  - Semantic classification (or speech understanding) based on word recognition – some words may carry more semantically relevant information than others; e.g., speech understanding using keyword spotting

Confusion Matrix (Baseline – ML)

	1	2	3	4	5	6	7	8	9	0	oh	DEL
One	*	0	0	0	0	0	0	0	0	0	**0**	0
Two	0	*	0	0	0	0	0	0	0	0	**4**	0
Three	0	3	*	0	0	0	0	0	0	0	**0**	0
Four	**1**	**0**	**0**	*	**0**	**0**	**0**	**0**	**0**	**0**	**5**	**0**
Five	0	0	0	0	*	0	0	0	0	0	**0**	0
Six	0	0	0	0	0	*	1	0	0	0	**0**	0
Seven	1	0	0	0	0	0	*	0	0	0	**1**	0
Eight	0	0	0	0	0	1	0	*	0	0	**0**	1
Nine	1	0	0	0	1	0	0	0	*	0	**1**	0
Zero	0	0	0	0	0	0	0	0	0	*	**0**	0
Oh	0	0	0	0	0	0	1	0	0	0	*	18
Ins	2	0	0	0	0	1	0	3	1	1	**92**	

Connected Digit ASR Results

Cost matrix:

$$e_{ij} = \begin{cases} 1, & i \neq j, i \neq 11, j \neq 4 \\ 0, & i = j \\ 10, & i = 11, j = 4 \end{cases}$$

$$[e_{ij}] = \begin{bmatrix} 0 & 1 & 1 & 10 & \cdots & 1 \\ 1 & 0 & 1 & 10 & \cdots & 1 \\ 1 & 1 & 0 & 10 & \cdots & 1 \\ \cdots & & \cdots & & \cdots & \\ 1 & 1 & 1 & 10 & \cdots & 1 \\ 10 & 10 & 10 & 10 & \cdots & 0 \end{bmatrix}$$

	Total Errors	"4" to others	Others to "oh"
Baseline	140	6	11
MCE (Uniform)	76	4	6
MCE (non-U)	67	1	4

Georgia Institute of Technology

Confusion Matrix (MCE)

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**0**	**oh**	DEL
One	*	0	0	0	0	0	0	0	0	0	**0**	0
Two	0	*	0	0	0	0	0	0	0	0	**4**	0
Three	0	3	*	0	0	0	0	0	0	0	**0**	0
Four	**2**	**0**	**0**	*	**0**	**0**	**0**	**0**	**0**	**0**	**2**	**0**
Five	0	0	0	0	*	0	0	0	0	0	**0**	0
Six	0	1	0	0	2	*	1	0	0	0	**0**	0
Seven	1	0	0	1	0	0	*	0	0	0	**0**	0
Eight	0	0	0	0	0	1	0	*	0	0	**0**	4
Nine	2	0	0	0	1	0	0	0	*	0	**0**	0
Zero	0	0	0	0	0	0	0	0	0	*	**0**	0
Oh	1	0	0	0	0	0	1	1	0	2	*	25
Ins	1	0	0	0	0	0	0	2	1	1	**18**	

Georgia Institute of Technology

Confusion Matrix (MCE – Non-U Cost)

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**0**	**oh**	DEL
One	*	0	0	0	0	0	0	0	0	0	**0**	0
Two	0	*	0	0	0	0	0	0	0	0	**4**	0
Three	0	3	*	0	0	0	0	0	0	0	**0**	0
Four	**1**	**0**	**0**	*	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Five	0	0	0	0	*	0	0	0	0	0	**0**	0
Six	0	0	0	0	0	*	1	0	0	0	**0**	0
Seven	1	0	0	0	0	0	*	0	0	0	**0**	0
Eight	0	0	0	0	0	1	0	*	0	0	**0**	4
Nine	1	0	0	0	1	0	0	0	*	0	**0**	0
Zero	0	0	0	0	0	0	0	0	0	*	**0**	0
Oh	0	0	0	0	0	0	2	4	1	0	*	23
Ins	1	0	0	0	0	1	0	2	1	1	**14**	

Georgia Institute of Technology

Take-home Message

- Recognition errors have varying significance;
- A new design dimension that allows incorporation of non-uniform error significance is now available.

Georgia Institute of Technology