Minimum Bayes Risk

SFPLODD

7 March 2018

Some Things You Know

- How to decode by finding the single best global structure
  - Lots of ways to think about the algorithms
- How to find posterior marginals for “parts” (a.k.a. “cliques”), if we interpret scoring probabilistically

A Different View of Decoding

- **Cost** (sometimes called "loss"): a function that tells how bad every guess y is, given every correct answer y*:

  cost : Val(Y) × Val(Y) → [0, ∞)

- **Risk**: pretend Y* is random and distributed according to your model distribution; risk is the expectation of cost, for a given y:

  risk: Val(Y) → [0, ∞)

- **MBR decoding**: pick the y that minimizes risk.

$$\arg\min_{\boldsymbol{y}} \sum_{\boldsymbol{y}^* \in \mathcal{Y}} p(\boldsymbol{y}^* \mid \boldsymbol{x}) \times \mathrm{cost}(\boldsymbol{y}, \boldsymbol{y}^*)$$

Derivation

$$
\begin{aligned}
\min_{\boldsymbol{y}} \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{Y}^*)}[\mathrm{cost}(\boldsymbol{y},\boldsymbol{Y}^*)] &= \min_{\boldsymbol{y}} \sum_{\boldsymbol{y}^* \in \mathcal{Y}} p(\boldsymbol{x},\boldsymbol{y}^*) \times \mathrm{cost}(\boldsymbol{y},\boldsymbol{y}^*) \\
&= \min_{\boldsymbol{y}} \sum_{\boldsymbol{y}^* \in \mathcal{Y}} p(\boldsymbol{x}) \times p(\boldsymbol{y}^* \mid \boldsymbol{x}) \times \mathrm{cost}(\boldsymbol{y},\boldsymbol{y}^*) \\
&= p(\boldsymbol{x}) \times \min_{\boldsymbol{y}} \sum_{\boldsymbol{y}^* \in \mathcal{Y}} p(\boldsymbol{y}^* \mid \boldsymbol{x}) \times \mathrm{cost}(\boldsymbol{y},\boldsymbol{y}^*)
\end{aligned}
$$

Example 1: Posterior Decoding

- model: sequence labeling with bigram label factors
- cost(y, y*): number of tokens you mislabeled (sometimes called "Hamming" cost)
- risk(y): expected number of mislabeled tokens in y

$$\sum_{\boldsymbol{y}^*} p(\boldsymbol{y}^* \mid \boldsymbol{x}) \sum_{i=1}^{n} \mathbf{1}\{y_i \neq y_i^*\} = \mathbb{E}_{p(\boldsymbol{Y}^* \mid \boldsymbol{x})} \left[\sum_{i=1}^{n} \mathbf{1}\{y_i \neq Y_i^*\} \right]$$

$$= \sum_{i=1}^{n} \mathbb{E}_{p(\boldsymbol{Y}^* \mid \boldsymbol{x})} [\mathbf{1}\{y_i \neq Y_i^*\}]$$

$$= \sum_{i=1}^{n} \left(1 - \mathbb{E}_{p(\boldsymbol{Y}^* \mid \boldsymbol{x})} [\mathbf{1}\{y_i = Y_i^*\}]\right)$$

Example 2: 0-1 cost

- model: anything
- cost(y, y*): 0 if y = y*, 1 otherwise
- risk(y): $1 - p(y \mid x)$

Example 2: 0-1 cost

- model: anything
- cost(y, y*): 0 if y = y*, 1 otherwise
- risk(y): 1 – p(y | x)

$\boxed{\text{this is MAP}}$

Example 3: Maximum Expected Recall (Goodman, 1996)

- model: PCFG
- cost(y, y*) = number of labeled spans in y* that are not in y
- risk(y) = sum of
  (1 - posterior probability of a labeled span)

Example 4: Weighting Different BIO Errors

- model: BIO
- cost: different costs for recall, precision, and *boundary* errors:

correct:	B-B	B-I	B-O	I-B	I-I	I-O	O-B	O-O
B-B		split	prec.		split	prec.		prec.
B-I	merge		bound.	merge		bound.	bound.	bound.
B-O	recall	recall		recall	bound.		recall	
I-B		split	prec.		split	prec.		prec.
I-I	merge		bound.	merge		bound.	bound.	bound.
I-O	recall	recall		recall	bound.		recall	
O-B		prec.	prec.		bound.	prec.		prec.
O-O	recall			recall	recall		recall	

General MBR Algorithm

Assumption: cost factors locally into parts

1. Calculate posterior distribution for each part (generalized inside algorithm)
2. If parts don't overlap, pick local argmax for each part.
3. Otherwise, decode with a model that defines:

$$\bar{f}_{j,\boldsymbol{\pi}}(\boldsymbol{\pi}') = -\mathrm{localcost}(\boldsymbol{\pi}, \boldsymbol{\pi}')$$

$$\bar{w}_{j,\boldsymbol{\pi}} = p(\mathrm{part}\ j = \boldsymbol{\pi} \mid \boldsymbol{x})$$

Pop Quiz

Can you think of a cost function such that minimum Bayes risk decoding *can't* be done in polynomial time?