Unlabeled Data and Other Marginals

Mark Hasegawa-Johnson, Jui-Ting Huang, and Xiaodan Zhuang

University of Illinois

ACL/ICML/ISCA Symposium, June 27, 2011

Outline

1. Fundamentals
2. Semi-Supervised Learning
3. MMI+NCE
4. Pronunciation Modeling
5. Conclusions

Outline

1. Fundamentals
2. Semi-Supervised Learning
3. MMI+NCE
4. Pronunciation Modeling
5. Conclusions

Hoeffding's Inequality

$z_1, \ldots, z_n$ i.i.d., $P(z_i \in [0, R]) = 1$, then

$$P\left(|E[z] - \langle z \rangle|\right) \geq \epsilon) \leq 2e^{-\frac{2\epsilon^2 n}{R^2}}$$

for $\langle z \rangle \equiv \frac{1}{n} \sum z_i$ and $E[z] \equiv \int zp(z)dz$

Probably Approximately Correct (PAC) Learning

- **Hypothesis Space:** $h : \mathcal{X} \rightarrow \mathcal{Y}$ has cardinality $N(\mathcal{H})$
- **Loss Function:** $f(h(x_i), y_i) \in [0, R]$ w/probability one
- **Confidence:**
  $\delta \equiv P\left(\max_{h \in \mathcal{H}} |E[f(h(x), y)] - \langle f(h(x), y) \rangle| \geq \epsilon\right)$
- **The Basic PAC Bound:**

$$\epsilon \leq R\sqrt{\frac{\ln 2N(\mathcal{H}) - \ln \delta}{2n}}$$

Continuous Hypothesis Spaces: Covering Number

$N(\mathcal{H})$ =size of the ϵ-covering set for empirical and stochastic averages of $f(\mathcal{H})$, i.e., the smallest possible discrete set $\{h_1, \ldots, h_{N(\mathcal{H})}\}$ such that

$$\max_h \left(\min_{1 \leq j \leq N(\mathcal{H})} |E[f(h_j(x), y)] - E[f(h(x), y)]| \right) \leq \epsilon$$

$$\max_h \left(\min_{1 \leq j \leq N(\mathcal{H})} |\langle f(h_j(x), y) \rangle - \langle f(h(x), y) \rangle| \right) \leq \epsilon$$

Continuous Hypothesis Spaces: Revised PAC Bound

$$\delta \equiv P\left(\max_{h \in \mathcal{H}} |E[f(h(x), y)] - \langle f(h(x), y) \rangle| \geq 3\epsilon \right)$$

$$\epsilon \leq R \sqrt{\frac{\ln 2N(\mathcal{H}) - \ln \delta}{2n}}$$

Outline

1 Fundamentals

2 Semi-Supervised Learning

3 MMI+NCE

4 Pronunciation Modeling

5 Conclusions

Kernel Estimators of Conditional Risk

Define $f_X(h(\xi), y)$ to be the kernel projection of $h(\xi)$ onto x,

$$f_x(h(\xi), y) \equiv f(h(\xi), y) K(x, \xi)$$

for some symmetric positive-definite kernel, $K(x, \xi) \in [0, 1]$.

Conditional Covering Number

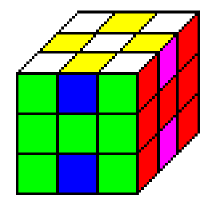
Define $N(\mathcal{H}|x)$ to be size of a set h_j which is big enough to explain all of the losses incurred only by the data points that are "near" x, where the word "near" is defined by the kernel. Specifically,

$$\max_h \left(\min_{1 \le j \le N(\mathcal{H}|x)} \left| E_{\xi,y}[f_x(h_j(\xi), y)] - E_{\xi,y}[f_x(h(\xi), y)] \right| \right) \le \epsilon$$

$$\max_h \left(\min_{1 \le j \le N(\mathcal{H}|x)} \left| \langle f_x(h_j(\xi), y) \rangle - \langle f_x(h(\xi), y) \rangle \right| \right) \le \epsilon$$

Usually, $N(\mathcal{H}|x) \ll N(\mathcal{H})$.

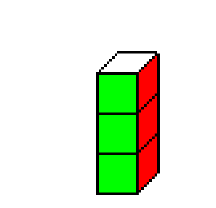
Covering Number: Example



- $\mathcal{X} = [0,1]^2, \quad \mathcal{Y} = [0,1]$
- $f(h(x), y) = h(x) - y$

$$N(\mathcal{H}) \sim \left(\frac{1}{\epsilon}\right)^3$$

Conditional Covering Number



Let's use a 2ϵ-width rectangular kernel:

$$f_x(h(\xi), y) = \begin{cases} h(\xi) - y & |x - \xi| < \epsilon \\ 0 & \text{else} \end{cases}$$

so

$$N(\mathcal{H}|x) \sim \left(\frac{1}{\epsilon}\right)$$

Confidence of the Conditional Risk Estimate

$$\delta(x) \equiv P\left(\max_{h \in \mathcal{H}(x)} \left| E[f_x(h(\xi), y)] - \langle f_x(h(\xi), y) \rangle \right| \geq 3\epsilon \right)$$

A Semi-Supervised PAC Bound

Suppose (1) $p(x)$ is known, e.g., because we have lots and lots of unlabeled data, (2) we don't really care about $\delta(x)$, but only about

$$\ln \delta \equiv E_x \left[\ln \delta(x) \right]$$

If we're willing to redefine "confidence" in this way, then it is possible to bound ϵ much more tightly in the semi-supervised case than in the supervised case, for two reasons.

- **Range:** $\langle f_x(h, y) \rangle \equiv \langle f(h, y) K(x, \xi) \rangle$ tends to be much smaller than $\langle f(h, y) \rangle$. We compensate by rescaling R.
- **VC Dimension:** $\ln N(\mathcal{H}|x)$ is less than $N(\mathcal{H})$. The reduced VC dimension creates a better bound.

PAC Bound for Semi-Supervised Learning

$$\epsilon \leq \bar{R}\sqrt{\frac{E_x[\ln 2N(\mathcal{H}|x)] - \ln \delta}{2n}}$$

- **Range:** $f_x(h,y)$ has a much smaller range than $f(h,y)$. the root-harmonic-mean-squared radius, $\bar{R} \ll R$, compensates for the difference in range.

$$\bar{R} = R\left(E_x\left[\left(\frac{1}{n}\sum_i K^2(x,x_i)\right)^{-1}\right]\right)^{-1/2}$$

- **VC Dimension:** In addition to the much smaller range, $f_x(h,y)$ also typically has a much smaller covering number than $f(h,y)$. The VC dimension, $E_x[\ln N(\mathcal{H}|x)]$, may therefore be much smaller than the VC dimension, $\ln N(\mathcal{H})$, that can be achieved without the unlabeled data.

Outline

1. Fundamentals
2. Semi-Supervised Learning
3. MMI+NCE
4. Pronunciation Modeling
5. Conclusions

Maximum Mutual Information (MMI)

MMI is defined by the hypothesis and loss function

$$\vec{h}(x) = \begin{bmatrix} \ln \hat{p}(Y=1|x) \\ \vdots \\ \ln \hat{p}(Y=c|x) \end{bmatrix}, \quad f(\vec{h},y) = \vec{h}^T \vec{\delta}_y = -\ln \hat{p}(Y=y|x)$$

MMI training chooses $\vec{h} \in \mathcal{H}$ to minimize

$$\langle f(\vec{h},y) \rangle \equiv -\frac{1}{n} \sum_{i=1}^{n} \ln \hat{p}(Y=y_i|x_i)$$

PAC bound on the resulting risk is

$$E[f(\vec{h},y)] \le \langle f(\vec{h},y) \rangle + R\sqrt{\frac{\ln 2N(\mathcal{H}) - \ln \delta}{2n}}$$

Covering Number for the MMI Loss

$f(\vec{h}, y) = -\ln \hat{p}(Y = y|x)$ has infinite covering number. Finite covering number is possible for an exponentiated average:

$$\max_h \left(\min_{1 \leq j \leq N(\mathcal{H}|x)} \left| e^{\langle f_x(h_j(\xi), y) \rangle} - e^{\langle f_x(h(\xi), y) \rangle} \right| \right) \leq \epsilon$$

For example, suppose we choose some arbitrary entropy threshold E_{max}, and limit the hypothesis space to:

$$\mathcal{H}(x) = \left\{ h : -\sum_{y \in \mathcal{Y}} \hat{p}(y|x) \ln \hat{p}(y|x) \leq E_{max} \right\}$$

then the covering number is

$$N(\mathcal{H}|x) \sim e^{E_{max}}$$

Semi-Supervised MMI

Estimate the VC dimension using unlabeled data, $\mathcal{D}_U = \{x_{n+1}, \ldots, x_{n+u}\}$:

$$E_x[\ln N(\mathcal{H}|x)] \approx -\frac{1}{u} \sum_{i=n+1}^{n+u} \sum_{y \in \mathcal{Y}} \hat{p}(x_i, y_i) \ln \hat{p}(y_i|x_i)$$

Choose $h(x)$ as

$$h = \arg\min -\frac{1}{n} \sum_{i=1}^{n} \ln \hat{p}(y_i|x_i), \quad \text{s.t. } E_x[\ln N(\mathcal{H}|x)] \leq E_{max}$$

whose corresponding Lagrangian is

$$\mathcal{F}(\vec{h}) = -\frac{1}{n} \sum_{i=1}^{n} \ln \hat{p}(y_i|x_i) - \frac{\alpha}{u} \sum_{i=n+1}^{n+u} \sum_{y \in \mathcal{Y}} \hat{p}(x_i, y) \ln \hat{p}(y|x_i)$$

Discriminative Training Criteria

Supervised: Maximum Mutual Information Minimum probability of error = maximum probability of the correct class = maximum mutual information (MMI) between observations and labels

$$\mathcal{F}_{MMI}^{(\mathcal{D}_L)}(\vec{h}) = \frac{1}{n}\sum_{i=1}^{n} \ln \hat{p}(y_i|x_i)$$

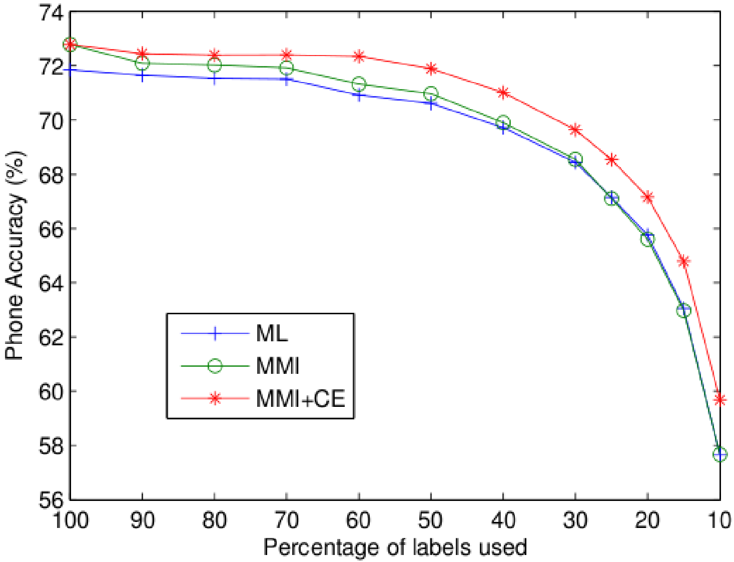
Unsupervised: Negative Conditional Entropy Encourage the model to have the greatest possible certainty about its labeling decisions

$$\mathcal{F}_{NCE}^{(\mathcal{D}_U)}(\hat{h}) = \frac{1}{u}\sum_{i=n+1}^{l+u}\sum_{y} \hat{p}(x_i, y) \ln \hat{p}(y|x_i)$$

Experiments: Phone Classification

- On TIMIT corpus
  - Training: 462 speakers, 3696 utterances, 140225 segments
  - Development: 50 speakers, 400 utterances, 15057 segments
  - Test: 118 speakers, 944 utterances, segments, 35697 segments
- 48 phone classes
- To create a semi-supervised setting: Labels of s% of the training set are kept ((100-s)% are unlabeled)
- Segmental features [Halberstadt '98]: a fixed length vector is calculated from the frame-based spectral features (12PLP coefficients plus energy)
  - Divide the frames for each phone segment into three regions with 3-4-3 proportion
  - Plus the 30 ms regions beyond the start and end time of the segment
  - Log duration
- Each phone is modeled by a GMM with two full-covariance Gaussian components

Results: Phone Recognition Accuracy



Outline

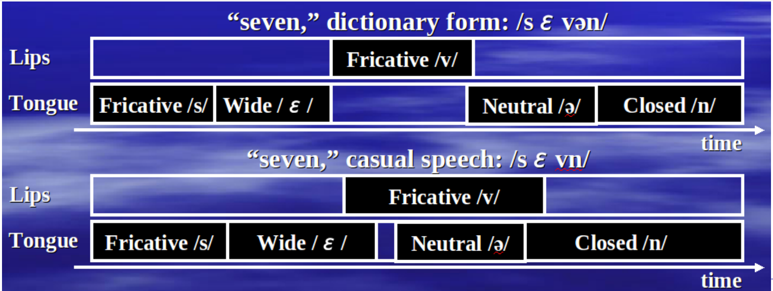
1. Fundamentals
2. Semi-Supervised Learning
3. MMI+NCE
4. Pronunciation Modeling
5. Conclusions

Pronunciation Modeling

What does it mean for similar tokens to have similar labels?

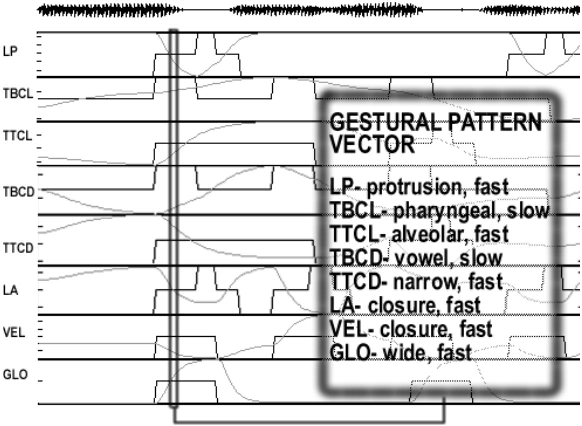
d(phone string 1, phone string 2) =
alignment-edit-distance(corresponding gestural scores)

- Gesture deletions, insertions, substitutions impossible (infinite distance)
- Gesture edge swaps (temporal re-alignment) possible with finite cost per swap

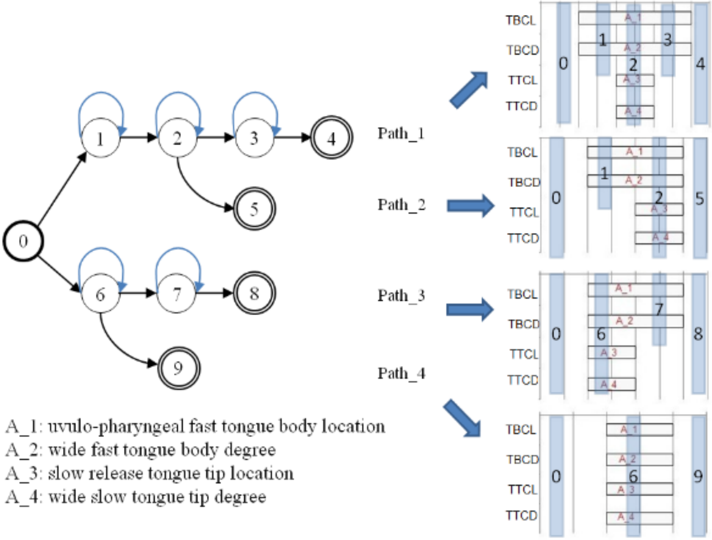


A Mapping Between Gestures and Phones

- Each phone corresponds to a canonical "gestural pattern vector" (GPV)
- There are more GPVs than phones; most GPVs correspond to non-English phones, allophones, or pseudo-phones



Proximity of Gestural Scores: “The”



A_1: uvulo-pharyngeal fast tongue body location
A_2: wide fast tongue body degree
A_3: slow release tongue tip location
A_4: wide slow tongue tip degree

Experimental Test: Recognition of Synthetic Speech

- Isolated word recognition: $\hat{w} = \arg\max p(O|Q)p(Q|w)$
- $O = [\vec{o}_1, \ldots, \vec{o}_T]$=Articulograph observations
- $Q = [q_1, \ldots, q_T]$=GPV sequence
- Observation PDF $p(O|Q)$ = ANN-GMM-HMM, trained on 277 words, tested on 139 words
- Pronunciation model $p(Q|w)$
  - Initialized using dictionary
  - Expanded to include up to N_Q alternate pronunciations with similar gestural scores, N_Q fixed in advance
  - No learning yet!! Similar gestural scores are assumed, *a priori*, to be members of the same class (same word)
  - (Future work: learning goes here?)

Accuracy, Synthetic Speech

Recognizer	Word Recognition Accuracy
GPV Bigram (models local GPV sequence statistics, not global)	85%
GPV-FST, $N_Q = 1$ pronunciation/word	88%
GPV-FST, $N_Q = 50$ pronunciations/word	90%
GPV-FST, $N_Q = 200$ pronunciations/word	90.7%

Outline

1. Fundamentals
2. Semi-Supervised Learning
3. MMI+NCE
4. Pronunciation Modeling
5. Conclusions

Conclusions

Conditional Learning: The hypothesis space for a given x is much smaller than the global hypothesis space ($N(\mathcal{H}|x) \ll N(\mathcal{H})$).

Semi-Supervised Learning: The expected log risk, over x, is bounded by the expected log covering number, $E_x[\ln N(\mathcal{H}|x)]$. Prior knowledge of $p(x)$ allows us to calculate and explicitly minimize this number, rather than the looser bound, $\ln N(\mathcal{H})$.

MMI+NCE: For the MMI loss function, the log covering number is the conditional class entropy. MMI+NCE therefore reduces phone classification error.

Pronunciation Modeling: Articulatory phonology specifies a similarity metric over phone sequences—a kind of label-sequence marginal, $p(y)$. Preliminary results suggest it may help ASR.