Learning to Recognize Speech from a Small Number of Labeled Examples

Mark Hasegawa-Johnson, Jui-Ting Huang, Rehab Duwairi, Eiman Mustafawi, Roxana Girju, and Elabbas Benmamoun

PAC Learning

- **Hypothesis Space:** $h : \mathcal{X} \rightarrow \mathcal{Y}$ has cardinality $N(\mathcal{H})$
- **Loss Function:** $\ell(h(x_i), y_i) \in [0, R]$ w/probability one
- **Hoeffding's Inequality:** if $\mathcal{D} = \{z_1, \ldots, z_n\}$ i.i.d., $z_i \in [0, R]$, then

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

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

- **Union Bound:**

$$P_{\mathcal{D}} \max_{h \in \mathcal{H}} \left(|E[\ell(h(x), y)] - \langle \ell(h(x), y) \rangle|\right) \geq \epsilon)$$

$$\leq N(\mathcal{H}) 2e^{-\frac{2\epsilon^2 n}{R^2}}$$

- **The Basic PAC Bound:** with probability at least $1 - \delta$,

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

Conditional PAC bound

Bizarre proposal: suppose we know x, but not y. Then how many distinct hypotheses are there?

- Well, a lot less than there would be if x were unknown!
- $\ell(h(x), y) \in [0, R]$ has R/ϵ distinct values for each value of y, so

$$N(\mathcal{H}|x) \leq \left(\frac{R}{\epsilon}\right) N(\mathcal{Y})$$

and with probability $1 - \delta$, $|E_x(\ell) - \langle \ell \rangle|$ is bounded by

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

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 $\epsilon(x)$, but only about

$$\bar{\epsilon}^2 \equiv E_x\left[\epsilon(x)^2\right]$$

Rather than minimizing a PAC bound on the worst-case risk, we minimize the expected squared PAC bound.

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

Since $E_x[\ln N(\mathcal{H}|x)] \ll \ln N(\mathcal{H})$, the semi-supervised classifier generalizes well from training to test data.

MMI Learning

Maximum mutual information (MMI) learning is defined by the hypothesis and loss function

$$\vec{h}(x) = \begin{bmatrix} \ln \hat{p}(Y = 1|x) \\ \vdots \\ \ln \hat{p}(Y = c|x) \end{bmatrix}$$

$$\ell(\vec{h}, y) = \vec{h}^T \vec{\delta}_y = -\ln \hat{p}(Y = y|x)$$

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

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

PAC bound on the resulting risk is

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

Reduced Hypothesis Space

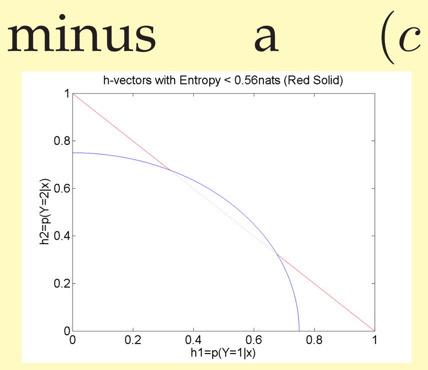
Suppose we choose $\mathcal{H}$ to include only the multinomial vectors whose entropy is less than η, for some arbitrary upper bound η:

$$\mathcal{H} = \left\{\vec{h} : \sum_{i=1}^{c} h_i = 1, \quad -\sum_{i=1}^{c} h_i \ln h_i \leq \eta\right\}$$

By Jensen's inequality,

$$\mathcal{H} \subset \left\{\vec{h} : \sum_{i=1}^{c} h_i = 1, \quad \sum_{i=1}^{c} h_i^2 \geq e^{-\eta}\right\}$$

The upper bound on $\mathcal{H}$ is a $(c-1)$-simplex minus a $(c-1)$-ball:



Covering number is the volume, divided into hypercubes each of whose side is $2\epsilon/\sqrt{c-1}$ (the rectangular grid that makes sure each point is within ϵ of a codevector):

$$N(\mathcal{H}|x) \leq V(\mathcal{H})\left(\frac{\sqrt{c-1}}{2\epsilon}\right)^{c-1}$$

which is on the order of

$$\ln N(\mathcal{H}|x) = \mathcal{O}\{\eta\}$$

Semi-Supervised: MMI+NCE

- Given a set of labeled data $\mathcal{D}_L$ and a set of unlabeled data $\mathcal{D}_U$:
- Regularize MMI using the entropy η, multiplied by Lagrange multiplier λ.
- Find the parameter set θ that maximizes

$$\mathcal{J}(\theta) = \mathcal{F}_{MMI}^{(\mathcal{D}_L)}(\theta) - \lambda\eta$$

$$= \frac{1}{l}\sum_{i=1}^{l} \ln p_\theta(y_i|x_i)$$

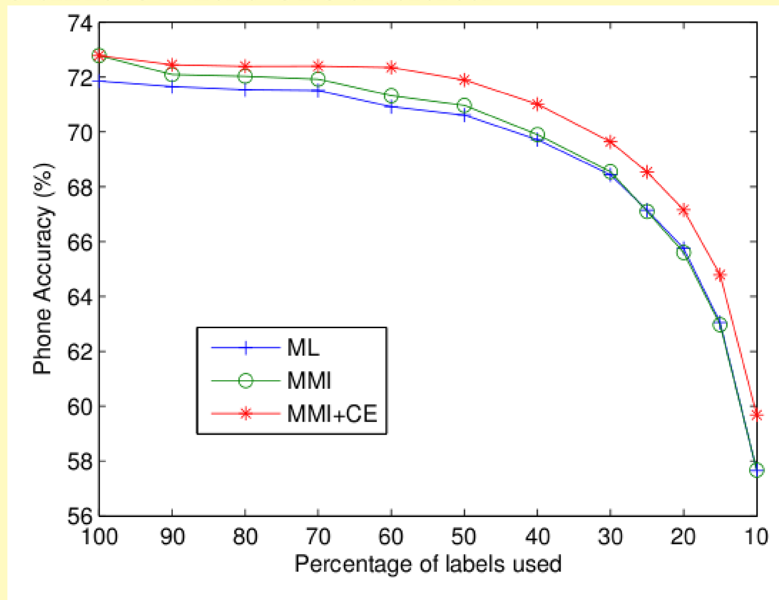
$$+\lambda\frac{1}{u}\sum_{i=l+1}^{l+u}\sum_{y} p_\theta(y|x_i) \ln p_\theta(y|x_i)$$

Experiments

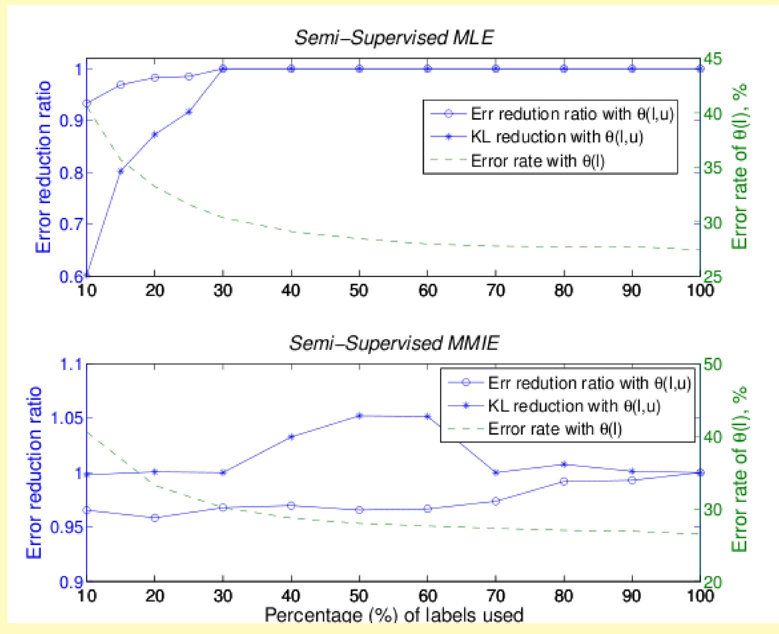
- **Data:** TIMIT phonetically labeled corpus of American English. Training: 462 speakers (140225 segments). Development: 50 speakers (15057 segments). Test: 118 speakers (35697 segments).
- **Labels:** 48 phone classes
- **Semi-Supervised:** Labels of s% of the training set are kept ((100-s)% are unlabeled)
- **Acoustic Features:** fixed length vector is calculated by averaging spectral features from each third of the segment (PLP+energy), and concatenating the three vectors together with segment duration.
- **Classifier:** Each phone is modeled by a GMM with two full-covariance Gaussian components

Results

Phone Recognition Accuracy as a function of the **amount of labeled data:**



- **ML training** can use unlabeled data to improve its estimate of the data distribution.
- **Discriminative training** can use unlabeled data to reduce the classifier error rate, even with very little unlabeled data.



Conclusions

- The **semi-supervised PAC bound** uses unlabeled data to re-weight the labeled data.
- **MMI+NCE** uses unlabeled data in order to learn a better classification boundary.
- Work in progress: Transfer learning, in order to use data from one Arabic dialect to help with speech recognition in another dialect.