Learning to Recognize Speech from a Small Number of Labeled Examples

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

PAC Learning

- **Hypothesis Space:** \( h : X \rightarrow Y \) has cardinality \( N(H) \).
- **Loss Function:** \( \ell(h(x), y) \in [0, R] \) w/ probability one
- **Hoeffding’s Inequality:** if \( D = \{z_1, \ldots, z_n\} \) i.i.d., \( z_i \in [0, R] \), then
  \[
P_D(\left|E[h(z)] - \langle h(z) \rangle \right| \geq \epsilon) \leq 2e^{-\frac{2n\epsilon^2}{R^2}}
\]
  for \( \epsilon \geq 0 \) and \( E[h(z)] \equiv \int z \rho(z)dz \)
- **Union Bound:**
  \[
P_H \max_{h \in H} \left|E[h(x), y] - \langle h(x), y \rangle \right| \geq \epsilon
  \]
  \[
  \leq N(H)2e^{-\frac{2n\epsilon^2}{R^2}}
  \]
- **The Basic PAC Bound:** with probability at least \( 1 - \delta \),
  \[
  \epsilon \leq R\sqrt{\frac{\ln 2N(H) - \ln \delta}{2n}}
  \]

**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 \( R/\epsilon \) distinct values for each value of \( y \), so
  \[
  N(H) \leq \frac{R}{\epsilon} N(Y)
  \]
  and with probability \( 1 - \delta \), \( |E_h(\ell) - \langle \ell \rangle| \) is bounded by
  \[
  \epsilon \leq R\sqrt{\frac{\ln 2N(H(x)) - \ln \delta}{2n(x)}}
  \]

**Semi-Supervised PAC Bound**

Suppose (1) \( \rho(x) \) 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
  \[
  \epsilon^2 \equiv E_x[\epsilon(x)^2]
  \]

Rather than minimizing a PAC bound on the worst-case risk, we minimize the expected squared PAC bound.
\[
\epsilon \leq R\sqrt{E_x[\ln 2N(H(x)) - \ln \delta]} \]

Since \( E_x[\ln N(H(x))] \ll \ln N(H) \), 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
\[
\tilde{h}(x) = \begin{cases}
\ln \hat{p}(Y = 1|x) \\
\ln \hat{p}(Y = c|x)
\end{cases}
\]
\[
\ell(h, y) = R^T \delta_y - \ln \hat{p}(Y = y|x)
\]

MMI training chooses \( \tilde{h} \in H \) to minimize
\[
\langle \ell(\tilde{h}, y) \rangle \equiv -\frac{1}{n} \sum_{i=1}^{n} \ln \hat{p}(Y = y|x_i)
\]

**PAC bound on the resulting risk** is
\[
E[\langle \ell(\tilde{h}, y) \rangle] \leq \langle \ell(\tilde{h}, y) \rangle + R\sqrt{\frac{\ln 2N(H) - \ln \delta}{2n}}
\]

**Reduced Hypothesis Space**

Suppose we choose \( H \) to include only the multinomial vectors whose entropy is less than \( \eta \), for some arbitrary upper bound \( \eta \):
\[
H = \left\{ \tilde{h} : \sum_{i=1}^{c} h_i = 1, -\sum_{i=1}^{c} h_i \ln h_i \leq \eta \right\}
\]

By Jensen’s inequality,
\[
H \subset \left\{ \tilde{h} : \sum_{i=1}^{c} h_i = 1, \sum_{i=1}^{c} h_i^2 \geq e^{-\eta} \right\}
\]

The upper bound on \( H \) is a \((c - 1)\)-simplex minus a \((c - 1)\)-ball:
\[
N(H(x)) \leq V(H) \left( \sqrt{c-1} \right)^{c-1}
\]

which is on the order of
\[
\ln N(H(x)) = O(\eta)
\]

**Semi-Supervised: MMI+NCE**

- Given a set of labeled data \( D_L \) and a set of unlabeled data \( D_U \):
  - Regularize MMI using the entropy \( \eta \), multiplied by Lagrange multiplier \( \lambda \).
  - Find the parameter set \( \theta \) that maximizes
  \[
  J(\theta) = F_M^{\left( D_L \right)}(\theta) - \lambda \eta
  \]
  \[
  = \frac{1}{L} \sum_{l=1}^{L} \ln p_{e}(y|x_l)
  \]
  \[
  + \lambda \sum_{i=1}^{n} \sum_{y=1}^{Y} p(y|x_i) \ln p_{e}(y|x_i)
  \]

Experiments

- **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.