On the Convergence Rate of Good-Turing Estimators

David McAllester
Robert Schapire
AT&T Labs-Research

Outline

- Large deviation theorems for the Good-Turing estimators
- Smoothing n-gram language models
- Selecting the unilexical subset of a bilexical grammar model
- Proof methods

The Missing Mass

Consider a distribution D over a countable set of "words" where word w has probability P_w.

Let S be a sample of m words drawn IID according to D.

We wish to estimate the "missing mass" M_0 defined as follows.

$$M_0 \equiv \sum_{w \notin S} P_w.$$

The Good-Turing Missing Mass Estimate

Let S_1 be the set of words that occur once in the sample.

The Good-Turing estimate of the missing mass is **the fraction of the sample consisting of words that only occur once**.

$$G_0 \equiv \frac{|S_1|}{m}$$

Theorem[Turing early 1940s, Good 1953]

$$E[|M_0 - G_0|] \leq \frac{1}{m}$$

.

Our Result

$$M_0 \leq G_0 + (2\sqrt{2} + \sqrt{3})\sqrt{\frac{\ln(\frac{3}{\delta})}{m}}$$

This should be compared to the Chernoff bound for the estimator $\hat{p}$ of a coin bias p.

$$p \leq \hat{p} + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}}$$

Higher Order Good-Turing Estimators

Let M_k be the probability that a fresh word is one that occurred k times in the sample.

Let G_k be the fraction of the sample due to words that occurred $k+1$ times in the sample.

Our Higher-Order Results

$$|\mathrm{E}\left[\mathrm{G_k}\right] - \mathrm{E}\left[\mathrm{M_k}\right]| \leq \mathrm{O}\left(\frac{\mathrm{k}}{\mathrm{m}}\right)$$

$$|G_k - \mathrm{E}\left[\mathrm{G_k}\right]| \leq \mathrm{O}\left(\mathrm{k}\sqrt{\frac{\ln\frac{1}{\delta}}{\mathrm{m}}}\right)$$

$$|M_k - \mathrm{E}\left[\mathrm{M_k}\right]| \leq \mathrm{O}\left(\left(\mathrm{k} + \ln\frac{\mathrm{m}}{\delta}\right)\sqrt{\frac{\ln\frac{1}{\delta}}{\mathrm{m}}}\right)$$

Interpolation Smoothing

Let S be a sample of m pairs $\langle s, w \rangle$ drawn IID from a fixed but unknown distribution D.

Consider an interpolated bigram model

$$P_{S,\vec{\lambda}}(w|u) = \lambda_u \frac{c(\langle u, w \rangle)}{c(u)} + (1 - \lambda_u) \frac{c(w)}{m}$$

Can we prove generalization bounds for some $\vec{\lambda}$?

Leave-One-Out Estimation

$$L\left(P_{S,\vec{\lambda}}\right) \equiv \mathrm{E}_{\langle u, w\rangle \sim \mathrm{D}}\left[\log \frac{1}{\mathrm{P}_{\mathrm{S},\vec{\lambda}}(\mathrm{w}|\mathrm{u})}\right]$$

$$\mathrm{LOO}\left(\mathrm{S},\ \vec{\lambda}\right) \equiv \frac{1}{\mathrm{m}} \sum_{\mathrm{i}=1}^{\mathrm{m}} \left(\log \frac{1}{\mathrm{P}_{\mathrm{S}\backslash \mathrm{i},\vec{\lambda}}(\mathrm{w}|\mathrm{u})}\right)$$

$$\mathrm{E}_{\mathrm{S}\sim\mathrm{D}^{\mathrm{m}}}\left[\mathrm{LOO}\left(\mathrm{S},\ \vec{\lambda}\right)\right] = \mathrm{E}_{\mathrm{S}\sim\mathrm{D}^{\mathrm{m}-1}}\left[\mathrm{L}\left(\mathrm{P}_{\mathrm{S},\ \vec{\lambda}}\right)\right]$$

Leave-One-Out Minimization

$$\vec{\lambda}^* \equiv \arg\min_{\vec{\lambda}} \quad \mathrm{LOO}\left(\mathrm{S},\ \vec{\lambda}\right)$$

Empirical Loss Minimization

$$\hat{L}(h) \equiv \frac{1}{m} \sum_{i=1}^{m} \left(\log \frac{1}{h(w_i|u_i)} \right)$$

$$h^* \equiv arg \min_{h \in H} \hat{L}(h, S)$$

$$\forall^\delta S \;\; L(h^*) \leq \left(\min_{h \in H} L(h) \right) + 2L_{\max} \sqrt{\frac{\ln |H| + \ln \frac{2}{\delta}}{2m}}$$

Our Result Again

$$\forall^\delta S \quad M_0 \leq G_0 + (2\sqrt{2} + \sqrt{3})\sqrt{\frac{\ln(\frac{3}{\delta})}{m}}$$

This should be compared to the Chernov confidence interval for $\hat{L}(h,\ S)$.

$$\forall^\delta S \ \ \left|L(h) - \hat{L}(h, S)\right| \leq L_{\max}\sqrt{\frac{\ln\frac{2}{\delta}}{2m}}$$

Bilexical Stochastic Grammars

$$\mathrm{VP}_{\mathrm{house}} \rightarrow \mathrm{Det}_{\mathrm{house}}\mathrm{N}_{\mathrm{house}}\mathrm{PP}^{*}_{\mathrm{house}}$$

$$\mathrm{PP}_{\mathrm{house}} \rightarrow \mathrm{PP}_{\mathrm{on}}$$

$$\mathrm{PP}_{\mathrm{house}} \rightarrow \mathrm{PP}$$

PAC-Bayes Generalization Bounds

$$L(G) \equiv \mathrm{E_s}\left[\min\left(\mathrm{L_{max}},\ -\log \mathrm{P_G(s)}\right)\right]$$

$$\hat{L}(G,\ S) \equiv \frac{1}{m} \sum_{i=1}^{m} \ \min\left(L_{\mathrm{max}},\ -\log(P_G(s_i))\right)$$

$$\forall^{\delta} S \ \ \forall G \ \ |L(G) - \hat{L}(G,\ S)| \leq \sqrt{\frac{(\ln 2)|G| + \frac{2}{\delta}}{2m}}$$

Bilexical Grammar Induction

Let G be a unilexical grammar structure with incomplete probability mass, e.g.,

$$\mathrm{ADJ}_{\mathrm{house}} \quad \rightarrow \quad \mathrm{ADJ}$$

$$\mathrm{ADJ} \quad \rightarrow \quad \mathrm{big}$$

Define G_S to be G plus bilexical productions derived from count ratios in S, e.g.,

$$\mathrm{ADJ}_{\mathrm{house}} \quad \rightarrow \quad \mathrm{big}$$

Leave-One-Out Minimization

$$\mathrm{LOO}\,(\mathrm{G},\ \mathrm{S}) \equiv \frac{1}{\mathrm{m}} \sum_{\mathrm{i}=1}^{\mathrm{m}} \min\left(\mathrm{L}_{\mathrm{max}},\ -\log \mathrm{P}_{\mathrm{G}(\mathrm{S}\backslash \mathrm{i})}(\mathrm{s}_{\mathrm{i}})\right)$$

We would like to prove:

$$\forall^{\delta} S\ \ \forall G\ \ |L(G_S) - \mathrm{LOO}\,(\mathrm{G},\ \mathrm{S})\,| \leq \mathrm{O}\left(\sqrt{\frac{|\mathrm{G}| + \ln \frac{1}{\delta}}{\mathrm{m}}}\right)$$

and optimize G

$$G^* = \arg\min_{\mathrm{G}}\ \ \mathrm{LOO}\,(\mathrm{G},\ \mathrm{S}) + \mathrm{O}\left(\sqrt{\frac{|\mathrm{G}| + \ln \frac{1}{\delta}}{\mathrm{m}}}\right)$$

Difficulties

One must work with sums of “negatively correlated” or “weakly correlated” variables.

One must work with sums of variables with widely different variances and with large upper bounds.

A Useful Result

If $\sigma^2(\beta) \leq \tilde{\sigma}^2$ for $0 \leq \beta \leq \beta(x)$ then

$$P(X \leq x) \leq \mathrm{EXP}\left(-\frac{(\mathrm{X} - \overline{\mathrm{X}})^2}{2\tilde{\sigma}^2}\right)$$

A More General Statement

$$P(X \leq x) \leq e^{-G(x)}$$

where

$$G(\overline{X}) = 0$$

$$G'(\overline{X}) = 0$$

$$G''(x) = \frac{1}{\sigma^2(\beta(x))}$$

Summary

- We have proved large deviation results for the Good-Turing leave-one-out estimators.
- We have proposed leave-one-out loss minimization for setting interpolation parameters in n-gram language models.
- We have proposed using leave-one-out estimation to select the unilexical component of bilexical stochastic grammar language model.
- We have conjectured that an MDL-like principle can be proved for leave-one-out estimators.
- We have stated some theorems that seem likely be useful tools in proving properties of leave-one-out estimation.