

050/520/600.666 Information Extraction from Speech and Text

Homework # 8

Due April 28, 2005.

Likelihood Based Clustering of Data

Consider the problem of modeling independent samples generated by S different sources using *possibly distinct* Gaussian densities $\mathcal{N}(\mathbf{y}; \mathbf{m}_s, \mathbf{U}_s)$ with parameters $\langle \mathbf{m}_s, \mathbf{U}_s \rangle$, $s = 1, \dots, S$. Let

$$\begin{aligned} \mathbf{Y}^1 &= \{\mathbf{y}_1^1, \mathbf{y}_2^1, \dots, \mathbf{y}_{N_1}^1\}, \\ \mathbf{Y}^2 &= \{\mathbf{y}_1^2, \mathbf{y}_2^2, \dots, \mathbf{y}_{N_2}^2\}, \\ &\vdots \\ \mathbf{Y}^S &= \{\mathbf{y}_1^S, \mathbf{y}_2^S, \dots, \mathbf{y}_{N_S}^S\}, \end{aligned} \quad \mathbf{y}_n^s \in \mathbb{R}^d \quad \forall s = 1, \dots, S, n = 1, \dots, N_s,$$

denote the observed data, and let the sample sum and sample sum-of-squares be denoted by

$$\mathbf{s}_s = \sum_{n=1}^{N_s} \mathbf{y}_n^s \quad \text{and} \quad \mathbf{Q}_s = \sum_{n=1}^{N_s} \mathbf{y}_n^s \mathbf{y}_n^{sT} \quad \text{respectively,} \quad s = 1, \dots, S.$$

This problem will explore the question of *clustering* these sources via the observed data.

1. If each source s is modeled by a *different* Gaussian density, what parameter values $\langle \hat{\mathbf{m}}_s, \hat{\mathbf{U}}_s \rangle$ maximize the total likelihood of the data, $\prod_{s=1}^S \prod_{n=1}^{N_s} \mathcal{N}(\mathbf{y}_n^s; \mathbf{m}_s, \mathbf{U}_s)$? Express your answer(s) in terms of the statistics \mathbf{s}_s and \mathbf{Q}_s .
2. Compute the *value* of this maximum total likelihood in terms of the $\hat{\mathbf{U}}_s$'s.
3. If two sources i and j are assumed to *share* a Gaussian density, what *tied* parameter values $\langle \hat{\mathbf{m}}_{\{i,j\}}, \hat{\mathbf{U}}_{\{i,j\}} \rangle$ maximize the total likelihood of $\mathbf{Y}^i \cup \mathbf{Y}^j$, $\prod_{s=i,j} \prod_{n=1}^{N_s} \mathcal{N}(\mathbf{y}_n^s; \mathbf{m}_{\{i,j\}}, \mathbf{U}_{\{i,j\}})$? Express your answer(s) in terms of $\mathbf{s}_i, \mathbf{s}_j, \mathbf{Q}_i$ and \mathbf{Q}_j .
4. Use your answer from part 3 to describe a procedure for choosing two sources, say, i^* and j^* , such that *tying* together their Gaussian densities results in a higher total likelihood of the data than tying together any other pair of sources.

$$(i^*, j^*) = \arg \max_{i,j \in \{1, \dots, S\}, i \neq j} \left[\prod_{s \neq i,j} \prod_{n=1}^{N_s} \mathcal{N}(\mathbf{y}_n^s; \hat{\mathbf{m}}_s, \hat{\mathbf{U}}_s) \times \prod_{s=i,j} \prod_{n=1}^{N_s} \mathcal{N}(\mathbf{y}_n^s; \hat{\mathbf{m}}_{\{i,j\}}, \hat{\mathbf{U}}_{\{i,j\}}) \right]$$

Hint: you may want to write the likelihood in the square-brackets above as

$$\left[\prod_{s=1}^S \prod_{n=1}^{N_s} \mathcal{N}(\mathbf{y}_n^s; \hat{\mathbf{m}}_s, \hat{\mathbf{U}}_s) \times \frac{\prod_{s=i,j} \prod_{n=1}^{N_s} \mathcal{N}(\mathbf{y}_n^s; \hat{\mathbf{m}}_{\{i,j\}}, \hat{\mathbf{U}}_{\{i,j\}})}{\prod_{s=i,j} \prod_{n=1}^{N_s} \mathcal{N}(\mathbf{y}_n^s; \hat{\mathbf{m}}_s, \hat{\mathbf{U}}_s)} \right],$$

and work with the log-likelihood in order to simplify your computation.

5. Show that the maximum total likelihood in part 4 is *necessarily* less than or equal to the maximum total likelihood in part 2.
6. Extend your procedure from part 4 to come up with an “algorithm” for bottom-up clustering of the S data sources, and interpret each internal node of the tree-structured hierarchy in terms of *similarity of the sources under a Gaussian model*.

Finally, for any arbitrary two-way clustering of the S sources, say, $\Phi_0 = \{\mathbf{Y}^1, \dots, \mathbf{Y}^{S_1}\}$ and $\Phi_1 = \{\mathbf{Y}^{S_1+1}, \dots, \mathbf{Y}^S\}$, compute the maximum total data likelihood

$$\prod_{s=1}^{S_1} \prod_{n=1}^{N_s} \mathcal{N}(\mathbf{y}_n^s; \hat{\mathbf{m}}_{\Phi_0}, \hat{\mathbf{U}}_{\Phi_0}) \times \prod_{s=S_1+1}^S \prod_{n=1}^{N_s} \mathcal{N}(\mathbf{y}_n^s; \hat{\mathbf{m}}_{\Phi_1}, \hat{\mathbf{U}}_{\Phi_1})$$

attainable by a two-Gaussian model, and compare it with that attainable under a one-Gaussian model

$$\prod_{s=1}^S \prod_{n=1}^{N_s} \mathcal{N}(\mathbf{y}_n^s; \hat{\mathbf{m}}, \hat{\mathbf{U}}).$$

Express your answers in terms of $\hat{\mathbf{U}}_{\Phi_0}$, $\hat{\mathbf{U}}_{\Phi_1}$ and $\hat{\mathbf{U}}$.

Reading Assignment

Read the following paper and summarize it in your own words in no more than 3 pages.

S. Young and P. Woodland, “State Clustering in hidden Markov model-based Continuous Speech Recognition,” *Computer Speech and Language*, 8(4):369-383, 1994.