Information Extraction from Speech and Text

Homework # 11
Due May 5, 2011.

Review Chapter 9 from *Statistical Methods for Speech Recognition* by Frederick Jelinek.

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

\[
\mathbf{Y}^1 = \{y^1_1, y^1_2, \ldots, y^1_{N_1}\},
\mathbf{Y}^2 = \{y^2_1, y^2_2, \ldots, y^2_{N_2}\},
\vdots
\vdots
\vdots
\mathbf{Y}^S = \{y^S_1, y^S_2, \ldots, y^S_{N_S}\},
\]

where $y^s_n \in \mathbb{R}^d \quad \forall \quad s = 1, \ldots, S,$ $n = 1, \ldots, N_s$.

This problem will explore the question of clustering these sources via the observed data. Let $s_s = \sum_{n=1}^{N_s} y^s_n$ and $Q_s = \sum_{n=1}^{N_s} y^s_n y^s_n^T$ respectively, $s = 1, \ldots, S$.

(a) 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 $\prod_{s=1}^{S} \prod_{n=1}^{N_s} \mathcal{N}(y^s_n; \mathbf{m}_s, \mathbf{U}_s)$ of the data? Express your answer(s) in terms of the statistics $s_s$ and $Q_s$.

(b) Compute the value of this maximum total likelihood in terms of the $\hat{\mathbf{U}}_s$’s.

(c) 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 likelihood $\prod_{s=i,j} \prod_{n=1}^{N_s} \mathcal{N}(y^s_n; \mathbf{m}_{\{i,j\}}, \mathbf{U}_{\{i,j\}})$ of $\mathbf{Y}^i \cup \mathbf{Y}^j$? Express your answer(s) in terms of $s_i$, $s_j$, $Q_i$ and $Q_j$.

(d) Use your answer from part 1c to describe a procedure for choosing two sources, say, $i^*$ and $j^*$, such that tying together their densities results in a higher total likelihood of the data than tying together any other pair of sources.

\[
(i^*, j^*) = \arg \max_{i,j \in \{1, \ldots, S\}, i \neq j} \left[ \prod_{s=1}^{S} \prod_{n=1}^{N_s} \mathcal{N}(y^s_n; \hat{\mathbf{m}}_s, \hat{\mathbf{U}}_s) \times \prod_{s=i,j} \prod_{n=1}^{N_s} \mathcal{N}(y^s_n; \hat{\mathbf{m}}_{\{i,j\}}, \hat{\mathbf{U}}_{\{i,j\}}) \right]
\]
Hint: you may want to write the likelihood in the square-brackets above as
\[
\left[ \prod_{s=1}^{S} \prod_{n=1}^{N_s} \mathcal{N}(y_n^s; \hat{m}_s, \hat{U}_s) \times \frac{\prod_{s=1,j} \prod_{n=1}^{N_s} \mathcal{N}(y_n^s; \hat{m}_{i,j}, \hat{U}_{i,j})}{\prod_{s=i,j} \prod_{n=1}^{N_s} \mathcal{N}(y_n^s; \hat{m}_s, \hat{U}_s)} \right],
\]

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

(e) Show that the maximum total likelihood in part 1d is necessarily less than or equal to the maximum total likelihood in part 1b.

(f) Extend your procedure from part 1d 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, \( \Phi_0 = \{ Y^1, \ldots, Y^{S_1} \} \) and \( \Phi_1 = \{ Y^{S_1+1}, \ldots, Y^S \} \), compute the maximum total data likelihood
\[
\prod_{s=1}^{S_1} \prod_{n=1}^{N_s} \mathcal{N}(y_n^s; \hat{m}_{\Phi_0}, \hat{U}_{\Phi_0}) \times \prod_{s=S_1+1}^{S} \prod_{n=1}^{N_s} \mathcal{N}(y_n^s; \hat{m}_{\Phi_1}, \hat{U}_{\Phi_1})
\]

attainable by a two-Gaussian model, and compare it with that attainable under a one-Gaussian model
\[
\prod_{s=1}^{S} \prod_{n=1}^{N_s} \mathcal{N}(y_n^s; \hat{m}, \hat{U}).
\]

Express you answers in terms of \( \hat{U}_{\Phi_0}, \hat{U}_{\Phi_1} \) and \( \hat{U} \).

2. Read the following paper and summarize it in your own words in no more than 2 pages.


Continue working on Project #3. If you encounter difficulties, please consult the TAs well in advance of the deadline, which is Thursday, May 10, 2011.