Algorithms for Graphical Models (AGM)

Iterative Proportional Fitting

$Date: 2006/12/04 16:24:30 $

AGM-15

In this lecture

- Likelihood
- Maximising a likelihood
- Iterative Proportional Fitting (IPF)
- Decomposition for IPF

Acknowledgements

This lecture draws on material from Steffen Lauritzen and Sam Roweis.

The problem

- You (somehow) know the conditional independence relations between a set of variables, and
- they are best expressed by a *hierarchical model*.
- You have some data for these variables. . .
- . . . and want to choose factors (actual numbers) to define a probability distribution which 'fits' this data.

Fitting parameters to hierarchical models

Recall: a *hierarchical model* is a set of probability distributions with the same associated (reduced) hypergraph; and thus the same conditional independence assumptions.

Given (1) A (reduced) hypergraph representing a hierarchical model and (2) some data

Find The probability distribution in the hierarchical model which maximise the *likelihood* of the data

The probability of single data point

Let θ be a vector of parameters defining a particular probability distribution for our model. θ is just the set of all numbers in the factors defining a distribution. We allow the possibility that θ only defines a distribution up to normalisation (i.e. there's the nasty Z_θ).

Let $\mathbf{x}$ be a joint instantiation of the variables.

$$P(\mathbf{x}|\theta) = Z_\theta^{-1} \prod_{h \in \mathcal{H}} f_h(\mathbf{x}_h|\theta_h)$$

θ_h is just the sub-vector of θ relevant to factor f_h, the factor whose variables are the hyperedge h. So for some particular θ, $f_h(\mathbf{x}_h|\theta_h)$ is just the number for instantiation $\mathbf{x}_h$ in factor f_h.

The likelihood of a dataset

- Assume that our data is a collection of joint instantiations $\mathcal{D} = \mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n$. (These superscripts are indices, not powers!)

- We assume that each of these data points was sampled independently from the 'true' distribution in our hierarchical model: they are *independent and identically distributed* (iid). Thus:

$$P(\mathcal{D}|\theta) = \prod_{i=1}^{n} P(\mathbf{x}^i|\theta)$$

This probability is called the *likelihood of the data*. We view it as a function of θ.

AGM-15

Maximum likelihood estimation

Just find

$$\arg\max_{\theta} P(\mathcal{D}|\theta)$$

For each $h \in \mathcal{H}$, let $\mathbf{x}_h$ be a variable ranging over joint instantiations of the variables in h: these are the rows of a corresponding factor. Let $n(\mathbf{x}_h)$ be the number of times instantiation $\mathbf{x}_h$ appears in the data, and let n be the number of datapoints.

Maximising the *log-likelihood* is easier. We have:

$$\log P(\mathcal{D}|\theta) = \sum_{h \in \mathcal{H}} \sum_{\mathbf{x}_h} n(\mathbf{x}_h) \log f(\mathbf{x}_h|\theta_h) - n \log Z_\theta^{-1}$$

Finding the maximum likelihood θ

$$\frac{\partial P(\mathcal{D}|\theta)}{\partial f_h(\mathbf{x}_h|\theta_h)} = \frac{n(\mathbf{x}_h)}{f_h(\mathbf{x}_h|\theta_h)} - n\frac{P(\mathbf{x}_h|\theta)}{f_h(\mathbf{x}_h|\theta_h)}$$

This (partial) derivative hits zero when

$$P(\mathbf{x}_h|\theta) = \frac{n(\mathbf{x}_h)}{n}$$

We can prove that this provides the global maximum, so model marginals equal observed marginals at the maximum likelihood θ.

Iterative proportional fitting

- We need θ such that $P(\mathbf{x}_h|\theta) = \frac{n(\mathbf{x}_h)}{n}$. How to find it?

- With iterative proportional fitting, we repeatedly *fit* each factor in turn.

- This corresponds to climbing the likelihood surface one coordinate at a time.

Iterative proportional fitting (ctd)

Here's how to update factor f_h (I've dropped the θs to cut down on clutter):

$$f_h^{t+1}(\mathbf{x}_h) = f_h^t(\mathbf{x}_h)\frac{n(\mathbf{x}_h)/n}{P^t(\mathbf{x}_h)}$$

Clearly, the hard work is computing the marginal $P^t(\mathbf{x}_h)$.
Note that there's no change once $n(\mathbf{x}_h)/n = P(\mathbf{x}_h)$

Towards faster IPF

Join $\mathcal{H}_1 \vee \mathcal{H}_2 = \mathsf{red}(\mathcal{H}_1 \cup \mathcal{H}_2)$

Meet $\mathcal{H}_1 \wedge \mathcal{H}_2 = \mathsf{red}(\{h_1 \cap h_2 : h_1 \in \mathcal{H}_1, h_2 \in \mathcal{H}_2\})$

If $\mathcal{H}_1 = \{\{A, B, C\}, \{A, F\}, \{F, G\}, \{B, G\}\}$ and $\mathcal{H}_2 = \{\{B, C, D\}, \{D, E\}\}$ then

$$\begin{aligned} \mathcal{H}_1 \vee \mathcal{H}_2 &= \{\{A, B, C\}, \{A, F\}, \{F, G\}, \{B, G\}, \{B, C, D\}, \{D, E\}\} \\ \mathcal{H}_1 \wedge \mathcal{H}_2 &= \{\{B, C\}\} \end{aligned}$$

$\mathcal{H}_1 \vee \mathcal{H}_2$ is a *direct join* of $\mathcal{H}_1$ and $\mathcal{H}_2$ because $\mathcal{H}_1 \wedge \mathcal{H}_2$ has only one hyperedge.

AGM-15

Decomposable hypergraphs

- A hypergraph is *simple* if has only one hyperedge.

- A hypergraph is decomposable if either it is simple or it is the direct join of two smaller decomposable hypergraphs.

Back to IPF

If the hypergraph $\mathcal{H}$ for our hierarchical model is the direct join of two smaller hypergraphs $\mathcal{H}_1$ and $\mathcal{H}_2$, then we can do IPF for $\mathcal{H}_1$ and $\mathcal{H}_2$ *separately* and combine the results.

Let H_1 be the variables in $\mathcal{H}_1$ and H_2 be the variables in $\mathcal{H}_2$:

$$\hat{P}_{\mathcal{H}}(\mathbf{x}) = \frac{\hat{P}_{\mathcal{H}_1}(\mathbf{x}_{H_1})\hat{P}_{\mathcal{H}_2}(\mathbf{x}_{H_2})}{n(\mathbf{x}_{H_1 \cap H_2})/n}$$

Fitting decomposable models

- If the hypergraph is decomposable, we can extend the 'direct join' approach to show that we don't need IPF at all.

- Recall from earlier that a hypergraph is decomposable iff it is the clique hypergraph of some triangulated graph, and that these cliques ($\mathcal{C}$) can be arranged into a join forest with separators ($\mathcal{S}$). We have:

$$\hat{P}_{\mathcal{H}}(\mathbf{x}) = \frac{\prod_{c \in \mathcal{C}} n(\mathbf{x}_c)}{\prod_{s \in \mathcal{S}} n(\mathbf{x}_s)^{\nu(s)}}$$

$\nu(s)$ is how often the separator s appears in the join forest.