LTAT.02.004 MACHINE LEARNING II

Missing roadmap to Expectation-maximisation algorithm

Sven Laur
University of Tartu

Hard clustering versus soft clustering

	Hard clustering	Soft clustering
Maximisation goal	$\mathrm{p}[\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \mid \boldsymbol{z}, \boldsymbol{\Theta}]$	$\mathrm{p}[\boldsymbol{\Theta}] \cdot \sum_{\boldsymbol{z}} \mathrm{p}[\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{z} \mid \boldsymbol{\Theta}]$
Optimisation method	Two-step maximisation algorithm	
Tactical objective	$F(\boldsymbol{z}, \boldsymbol{\Theta})$	$F(q, \boldsymbol{\Theta})$
Mixture proportions	Ignored by design	Core of the model
Cluster labels	Search goal	Integrated out

Desired properties of the tactical objective

Property I. Let $q_{\boldsymbol{\Theta}}(\cdot)$ be the optimal probability distribution for label vectors $\boldsymbol{z}$ for fixed model parameters $\boldsymbol{\Theta}$. Then the tactical objective coincides with the actual objective:

$$F(q_{\boldsymbol{\Theta}}, \boldsymbol{\Theta}) = \log \mathrm{p}[\boldsymbol{\Theta}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \ .$$

Property II. For fixed model parameters $\boldsymbol{\Theta}$ the optimal probability distribution can be found as the posterior probability of label vectors:

$$q_{\boldsymbol{\Theta}}(\boldsymbol{z}) = \mathrm{p}[\boldsymbol{z}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{\Theta}] \ .$$

Rationale

- ▷ The first property is essential for obtaining a local maxima.
- ▷ The second property is needed to justify the practical algorithm.

The derivation of the tactical objective

Let $q(\boldsymbol{z})$ be an arbitrary probability distribution over label vectors $\boldsymbol{z}$. Then tautology together with Jensen's inequality assures

$$\begin{aligned}\log \mathrm{p}[\boldsymbol{\Theta}|\boldsymbol{x}_1,\ldots \boldsymbol{x}_n] &= \log\left(\sum_{\boldsymbol{z}} q(\boldsymbol{z})\cdot\frac{\mathrm{p}[\boldsymbol{\Theta},\boldsymbol{z}|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n]}{q(\boldsymbol{z})}\right)\\ &\geq \sum_{\boldsymbol{z}} q(\boldsymbol{z})\cdot\log\left(\frac{\mathrm{p}[\boldsymbol{\Theta},\boldsymbol{z}|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n]}{q(\boldsymbol{z})}\right) = F(q,\boldsymbol{\Theta})\end{aligned}$$

For the probability assignment $q_{\boldsymbol{\Theta}}(\boldsymbol{z}) = \mathrm{p}[\boldsymbol{z}|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n,\boldsymbol{\Theta}]$ we get

$$\begin{aligned}F(q_{\boldsymbol{\Theta}},\boldsymbol{\Theta}) &= \sum_{\boldsymbol{z}} q_{\boldsymbol{\Theta}}(\boldsymbol{z})\cdot\log\left(\frac{\mathrm{p}[\boldsymbol{\Theta},\boldsymbol{z}|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n]}{\mathrm{p}[\boldsymbol{z}|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n,\boldsymbol{\Theta}]}\right)\\ &= \sum_{\boldsymbol{z}} q_{\boldsymbol{\Theta}}(\boldsymbol{z})\cdot\log\left(\mathrm{p}[\boldsymbol{\Theta}|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n]\right) = \log \mathrm{p}[\boldsymbol{\Theta}|\boldsymbol{x}_1,\ldots \boldsymbol{x}_n]\ .\end{aligned}$$

Tactical objective as a linearisation

The expectation-maximisation algorithm can be viewed as follows:

- ▷ Guess model parameters $\boldsymbol{\Theta}^{(i)}$.
- ▷ Compute probability assignments $q_{\boldsymbol{\Theta}^{(i)}}(\boldsymbol{z}) = \mathrm{p}[\boldsymbol{z}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{\Theta}^{(i)}]$.
- ▷ Approximate $\mathrm{p}[\boldsymbol{z}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{\Theta}]$ with a linearisation $F(q_{\boldsymbol{\Theta}^{(i)}}, \boldsymbol{\Theta})$.
- ▷ Fix a new guess $\boldsymbol{\Theta}^{(i+1)}$ that maximises $F(q_{\boldsymbol{\Theta}^{(i)}}, \boldsymbol{\Theta})$.

As the actual value and linearisation can be expressed as

$$\log \mathrm{p}[\boldsymbol{\Theta}|\boldsymbol{x}_1, \ldots \boldsymbol{x}_n] = \sum_{\boldsymbol{z}} q_{\boldsymbol{\Theta}^{(i)}}(\boldsymbol{z}) \cdot \log \left(\frac{\mathrm{p}[\boldsymbol{\Theta}, \boldsymbol{z}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]}{\mathrm{p}[\boldsymbol{z}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{\Theta}]} \right)$$

$$F(q_{\boldsymbol{\Theta}^{(i)}}, \boldsymbol{\Theta}) = \sum_{\boldsymbol{z}} q_{\boldsymbol{\Theta}^{(i)}}(\boldsymbol{z}) \cdot \log \left(\frac{\mathrm{p}[\boldsymbol{\Theta}, \boldsymbol{z}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]}{\mathrm{p}[\boldsymbol{z}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{\Theta}^{(i)}]} \right)$$

Tactical objective as a linearisation

- ▷ Guess model parameters $\boldsymbol{\Theta}^{(i)}$.
- ▷ Compute probability assignments $q_{\boldsymbol{\Theta}^{(i)}}(\boldsymbol{z}) = \mathrm{p}[\boldsymbol{z}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{\Theta}^{(i)}]$.
- ▷ Approximate $\mathrm{p}[\boldsymbol{z}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{\Theta}]$ with the linear function $F(q_{\boldsymbol{\Theta}^{(i)}}, \boldsymbol{\Theta})$.
- ▷ Fix a new guess $\boldsymbol{\Theta}^{(i+1)}$ that maximises $F(q_{\boldsymbol{\Theta}^{(i)}}, \boldsymbol{\Theta})$

Kullback-Leibler divergence between probability assignments for label vectors

$$D(q_{\boldsymbol{\Theta}^{(i)}}||q_{\boldsymbol{\Theta}}) = \sum_{\boldsymbol{z}} q_{\boldsymbol{\Theta}^{(i)}}(\boldsymbol{z}) \cdot \log\left(\frac{\mathrm{p}[\boldsymbol{z}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{\Theta}^{(i)}]}{\mathrm{p}[\boldsymbol{z}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{\Theta}]}\right)$$

measures the linearisation error $\mathrm{p}[\boldsymbol{z}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{\Theta}^{(i)}] - F(q_{\boldsymbol{\Theta}^{(i)}}, \boldsymbol{\Theta})$.

Simplification of the lower bound

Observation I. The distribution $q_{\boldsymbol{\Theta}}$ decomposes into a product of posteriors:

$$q_{\boldsymbol{\Theta}}(\boldsymbol{z}) = \prod_{i=1}^{n} \mathrm{p}[z_i|\boldsymbol{x_i}, \boldsymbol{\Theta}] \ \ .$$

Observation II. Let $\boldsymbol{W}$ be matrix of weights $w_{ij} = \mathrm{p}[z_i = j|\boldsymbol{x_i}, \boldsymbol{\Theta}^{(*)}]$. The lower bound can be expressed only in terms of $\boldsymbol{W}$ and $\boldsymbol{\Theta}$:

$$\begin{aligned} F(q_{\boldsymbol{\Theta}^{(*)}}, \boldsymbol{\Theta}) = \log p[\boldsymbol{\Theta}] - \sum_{i=1}^{n}\sum_{j=1}^{k} w_{ij} \log w_{ij} \\ + \sum_{i=1}^{n}\sum_{j=1}^{k} w_{ij} \cdot \log \lambda_j + \sum_{i=1}^{n}\sum_{j=1}^{k} w_{ij} \cdot \log\left(p[\boldsymbol{x}_i|\boldsymbol{\Theta}_j]\right) \ \ . \end{aligned}$$

Parameter optimisation

Hard clustering finds model parameters of the jth cluster by solving

$$\sum_{i=1}^{n}[z_i = j] \cdot \log\left(p[\boldsymbol{x}_i|\boldsymbol{\Theta}_j]\right) \rightarrow \max \quad .$$

Soft clustering finds model parameters of the jth cluster by solving

$$\sum_{i=1}^{n}\sum_{j=1}^{k} w_{ij} \cdot \log\left(p[\boldsymbol{x}_i|\boldsymbol{\Theta}_j]\right) \rightarrow \max \quad .$$

and additionally updates mixture proportions $\lambda_1, \ldots, \lambda_k$.