Parameter Estimation for Log-linear Models as D.C. Optimisation

Tran The Truyen
thetruyen.tran@postgrad.curtin.edu.au

April 27, 2008

Contents

1 Problem Statement **1**
1.1 Standard Problem 1
1.2 Some Variants 2

2 Applications **2**
2.1 Log-linear Modelling of Stochastic Systems 2

3 State-of-the-Arts **3**
3.1 EM algorithm and DCA 3

1 Problem Statement

1.1 Standard Problem

Let $x = (x_1, x_2, ..., x_K) \in \mathbb{R}^K$; $o \in \mathcal{O}$; $v \in \mathcal{V}$; $w \in \mathcal{W}$, where $\mathcal{O}$, $\mathcal{V}$ and $\mathcal{W}$ are sets of finite size. Let $m_k(v, w, o)$ be some real or binary functions. We want to solve the following optimisation problem

$$x^* = \arg\min_x f(x) \tag{1}$$

where

$$\begin{aligned} f(x) &= \sum_{o \in \mathcal{O}} \Big(g(x, o) - h(x, v, o) \Big) \\ g(x, o) &= \log \sum_{v \in \mathcal{V}} \sum_{w \in \mathcal{W}} \exp \left(\sum_{k=1}^{K} x_k m_k(v, w, o) \right) \qquad (2) \\ h(x, v, o) &= \log \sum_{w \in \mathcal{W}} \exp \left(\sum_{k=1}^{K} x_k m_k(v, w, o) \right) \qquad (3) \end{aligned}$$

It can be shown that this is a D.C. optimisation problem[1] as g and h are convex functions with respect to x. A special case is when $w = \emptyset$ and the problem is reduced to convex programming. However, the challenges are:

- The problem may be very large-scale, e.g. $|\mathcal{O}| = 10^6$ and $K = 10^7$. With these scales, representing a vector x and a gradient ∇f can be difficult for some computing tools.
- f and its gradient may only be computed approximately.
- The problem can be ill-posed for some x_k, that is, a very large change in x_k may lead only to a little change in $f(x)$.
- Sometimes, m_k are positive numbers, so large positive x_k may lead to numerical overflow in some machines when the size of $\mathcal{V}$ and $\mathcal{W}$ is exponentially large.

1.2 Some Variants

Since the problem can be ill-posed, we may introduce a regularisation term as follows

$$f'(x) \quad = \quad f(x) + \alpha||x||^2$$

or

$$f'(x) = f(x) + \alpha||x||$$

for some $\alpha > 0$. The latter setting often results in many zeros variables.

Another popular practice to deal the ill-posed problem is that we may iteratively select only those variables that are influential in $f(x)$. This practice is known as *feature selection* in data processing. This may result in a non-smooth optimisation problem.

2 Applications

2.1 Log-linear Modelling of Stochastic Systems

In statistical modelling of physical systems with the discrete state $s = (v, w)$, we have some measurements $m_{k=1}^{K}(v, w, o)$ where o is some external data on which the system is dependent. Often, some part v of the system state is visible, and the part w is hidden from us. We are often interested in the following distribution

$$\Pr(v, w|o; x) = \frac{1}{Z(x, o)} \exp\left(\sum_{k=1}^{K} x_k m_k(v, w, o)\right) \tag{4}$$

[1]D.C. stands for *difference-of-convex.*

where $Z(o;x) = \sum_{v,w} \exp\left(\sum_{k=1}^{K} x_k m_k(v,w,o)\right)$ is the normalisation term. This distribution is also known as the Gibbs distribution, or the Maximum Entropy distribution [5]. This type of model has very important applications in natural language processing and pattern recognition [1, 7, 8, 9, 11, 12].

If we have multiple external data points $\mathcal{O}$ then the log-likelihood of the data is given as

$$
\begin{aligned}
\mathcal{L}(\mathcal{O}|x) &= \sum_{o\in\mathcal{O}} \log \sum_{w} \Pr(v,w|o;x) \\
&= \sum_{o\in\mathcal{O}} \left(\log \sum_{w} \exp\left(\sum_{k=1}^{K} x_k m_k(v,w,o)\right) - \log \sum_{v,w} \exp\left(\sum_{k=1}^{K} x_k m_k(v,w,o)\right)\right)
\end{aligned}
$$

In standard statistics, estimation of x_k is often done by maximising the data likelihood. With appropriate arrangement, this is exactly the problem in Equation 1.

3 State-of-the-Arts

There have been some well-known methods for solving the Equation 1:

- Generalised Iterative Scaling (GIS) and variants [2, 1]. However, the behaviours of these algorithms are similar to gradient descent, which is relatively slow. Beside, this is only applicable for the case when $w = \emptyset$.
- Conjugate gradients [4].
- A quasi-Newton method known as L-BFGS [6].

3.1 EM algorithm and DCA

The EM algorithm [3] is a generic technique for maximum likelihood learning with missing variables. In this subsection, we show that the EM is equivalent to the DCA [10] in the log-linear models.

Let us start from the Jensen's inequality applied to the log-likelihood

$$
\begin{aligned}
\log \sum_{w} \Pr(v,w|o;x) &= \log \sum_{w} Q(w) \Pr(v,w|o;x) \frac{1}{Q(w)} \\
&\geq \sum_{w} Q(w) \log \Pr(v,w|o;x) \frac{1}{Q(w)} \qquad (5)
\end{aligned}
$$

where $Q(w)$ is a distribution, i.e. $\sum_w Q(w) = 1$ and $Q(w) > 0$. The equality holds when

$$Q(w) = \Pr(w|v,o;x) \qquad (6)$$

The EM algorithm operates by looping through two steps:

1. *E-step*: maximising the following expectation

$$\mathcal{Q}(x) = \sum_w Q^{t-1}(w) \log \Pr(v, w|o; x)$$

This improves the lower bound of the log-likelihood as in (5). Note that we have ignored the term $\sum_w Q^{t-1}(w) \log Q^{t-1}(w)$ in the lower bound because it does not depend on x. Let the solution be x^t.

2. *M-step*: filling the lower bound gap by setting

$$Q^t(w) = \Pr(w|v, o; x^t)$$

The net result of these two steps is that the log-likelihood is monotonically increasing until reaching a local maximum.

When the distribution has the log-linear form as in (4), we have

$$\mathcal{Q}(x) = \sum_w \Pr(w|v, o; x^{t-1}) \left(\sum_{k=1}^{K} x_k m_k(v, w, o) \right) - \log Z(v, o)$$

This has the advantage that it is a convex programming problem. Then at the maximum of $\mathcal{Q}(x)$ we have

$$\begin{aligned} \left.\frac{\partial \mathcal{Q}(x)}{\partial x_k}\right|_{x_k^t} &= \sum_w \Pr(w|v, o; x^{t-1}) m_k(v, w, o) - \sum_{v,w} \Pr(v, w|o; x^t) m_k(v, w, o) \\ &= 0 \end{aligned} \tag{7}$$

Recall the definition of g and h in (2) and (3), respectively. Using some algebra, we can arrive that solving the maximum log-likelihood is equivalent to minimising the D.C. function

$$s(x, v, o) = g(x, o) - h(x, v, o)$$

and

$$\begin{aligned} \left.\frac{\partial g}{\partial x_k}\right|_{x_k^t} &= \sum_{v,w} \Pr(v, w|o; x^t) m_k(v, w, o) \\ \left.\frac{\partial h}{\partial x_k}\right|_{x_k^{t-1}} &= \sum_w \Pr(w|v, o; x^{t-1}) m_k(v, w, o) \end{aligned}$$

Thus, Equation 7 becomes

$$\left.\frac{\partial g}{\partial x_k}\right|_{x_k^t} = \left.\frac{\partial h}{\partial x_k}\right|_{x_k^{t-1}}$$

which is essentially the DCA.

In summary, the EM and DCA are local methods that iteratively maximise the concave lower bound of the log-likelihood. However, solving the Equation 7 does not have a closed form solution, so it is unclear that it offers any advantage over direct optimisation using gradient-based methods applied for the log-likelihood.

References

[1] A. L. Berger, S. A. Della Pietra, and V. J. Della Pietra. A Maximum Entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71, 1996.

[2] J. Darroch and D. Ratcliff. Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics*, 42:1470–1480, 1972.

[3] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society*, 39(1):1–38, 1977.

[4] M.R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *Journal of Research of the National Bureau of Standards*, 49(6):409–436, 1952.

[5] E. T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106:620–630, 1957.

[6] D. C. Liu and J. Nocedal. On the limited memory BFGS method for large scale optimization methods. *Mathematical Programming*, 45:503–528, 1989.

[7] K. Nigam, J. Lafferty, and A. McCallum. Using maximum entropy for text classification. In *Proceedings of the IJCAI-99 Workshop on Machine Learning for Information Filtering*, pages 61–67, 1999.

[8] F.J. Och and H. Ney. Discriminative training and maximum entropy models for statistical machine translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 295–302, 2001.

[9] A. Ratnaparkhi. A maximum entropy part-of-speech tagger. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 17–18, May 1996.

[10] Pham Dinh Tao and Le Thi Hoai An. A D.C. optimization algorithm for solving the trust-region subproblem. *SIAM Journal on Optimization*, 8(2):476–505, 1998.

[11] S.C. Zhu, Y. Wu, and D. Mumford. Filters, Random Fields and Maximum Entropy (FRAME): towards a unified theory for texture modeling. *International Journal of Computer Vision*, 27(2):107–126, 1998.

[12] C.L. Zitnick and T. Kanade. Maximum entropy for collaborative filtering. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 636–643. AUAI Press Arlington, Virginia, United States, 2004.