Variational noise-contrastive estimation of unnormalised latent variable models

Michael Gutmann

michael.gutmann@ed.ac.uk

Institute for Adaptive and Neural Computation
School of Informatics, University of Edinburgh

11th June 2019

Reference

B. Rhodes and M.U. Gutmann
Variational Noise-Contrastive Estimation
In Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS) 2019
`http://proceedings.mlr.press/v89/rhodes19a`

Overall goal: density estimation

- Given observed data $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ estimate parameters $\boldsymbol{\theta}$ of a statistical model $p(\mathbf{x}; \boldsymbol{\theta})$.
- We assume that $p(\mathbf{x}; \boldsymbol{\theta})$ is not directly available but specified in terms of an unnormalised latent variable model.
- Classical example: restricted Boltzmann machine
- Importance: such models are highly flexible and widely applicable

Latent variable models

- Latent variables = variables $\mathbf{z}$ in the model for which we do not have observed data
- Latent variable model: we model the joint behaviour of $(\mathbf{x}, \mathbf{z})$ and specify $p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})$, rather than $p(\mathbf{x}; \boldsymbol{\theta})$.
- Obtain model for the observables $\mathbf{x}$ by marginalising out $\mathbf{z}$

$$p(\mathbf{x}; \boldsymbol{\theta}) = \int p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) \, \mathrm{d}\mathbf{z} \tag{1}$$

but integral often too expensive to compute/approximate

Latent variable models are important

- Modelling tool: explain structure (dependencies) in observed data in terms of unobserved explanatory factors
  - PCA, ICA, factor analysis, HMMs, topic models, variational autoencoders, ...
- Probabilistic treatment of missing data
  - model missing values ✗ as unobserved (latent) random variables

$$\text{data matrix} = \begin{pmatrix} \checkmark & \checkmark & \text{✗} & \dots & \text{✗} \\ \text{✗} & \checkmark & \text{✗} & \dots & \checkmark \\ \checkmark & \text{✗} & \checkmark & \dots & \checkmark \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \text{✗} & \text{✗} & \checkmark & \dots & \checkmark \end{pmatrix}$$

Unnormalised models

- Model $p(\mathbf{x}; \boldsymbol{\theta})$ must satisfy for all parameter values $\boldsymbol{\theta}$

$$\int p(\mathbf{x}; \boldsymbol{\theta}) \, \mathrm{d}\mathbf{x} = 1 \tag{2}$$

- Unnormalised models $\phi(\mathbf{x}; \boldsymbol{\theta}) \propto p(\mathbf{x}; \boldsymbol{\theta})$ do not impose this constraint.
- Obtain $p(\mathbf{x}; \boldsymbol{\theta})$ by dividing by the partition function $Z(\boldsymbol{\theta})$.

$$p(\mathbf{x}; \boldsymbol{\theta}) = \frac{\phi(\mathbf{x}; \boldsymbol{\theta})}{\int \phi(\mathbf{x}; \boldsymbol{\theta}) \, \mathrm{d}\mathbf{x}} \tag{3}$$

but integral often too expensive to compute/approximate

Unnormalised models are important

- Removing normalisation constraint gives more flexibility in model specification ("energy-based modelling")
- Widely used:
  - Large class of undirected graphical models (e.g. Markov networks) are typically unnormalised.
  - Unsupervised representation learning (including word and graph embeddings)
  - Machine translation (e.g. Zoph et al, 2016[1])
  - Product recommendation: (e.g. Tschiatschek et al, 2016[2])
  - ...

[1] Simple, fast noise-contrastive estimation for large RNN vocabularies

[2] Learning probabilistic submodular diversity models via noise contrastive estimation

Unnormalised latent variable models

- Unnormalised latent variable models $\phi(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})$ are latent variable models that are unnormalised

$$\int \phi(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) \, \mathrm{d}\mathbf{z} \, \mathrm{d}\mathbf{x} = Z(\boldsymbol{\theta}) \neq 1 \tag{4}$$

- They are important:
  - estimation of unnormalised models from data with missing values
  - increased modelling flexibility (e.g. latent variable models do not need to satisfy normalisation constraint)

Can we use maximum likelihood estimation? *sometimes*

- Since model pdf $p(\mathbf{x}; \boldsymbol{\theta})$ is defined via integrals, (log) likelihood evaluations are expensive/intractable

$$p(\mathbf{x}; \boldsymbol{\theta}) = \frac{\int \phi(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) \, \mathrm{d}\mathbf{z}}{\int \phi(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) \, \mathrm{d}\mathbf{z} \, \mathrm{d}\mathbf{x}} \quad \ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log p(\mathbf{x}_i; \boldsymbol{\theta}) \quad (5)$$

- Gradient $\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta})$ can be expressed as

$$\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z}|\mathbf{x}_i; \boldsymbol{\theta})} \left[\nabla_{\boldsymbol{\theta}} \log \phi(\mathbf{x}_i, \mathbf{z}; \boldsymbol{\theta}) \right] - \mathbb{E}_{\mathbf{x}, \mathbf{z} \sim p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})} \left[\nabla_{\boldsymbol{\theta}} \log \phi(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) \right] \quad (6)$$

- Enables gradient ascent on the log-likelihood *IF* computing the expectations (e.g. via sampling) is reasonably efficient.

Alternative: variational noise-contrastive estimation

(Rhodes and Gutmann, AISTATS, 2019)

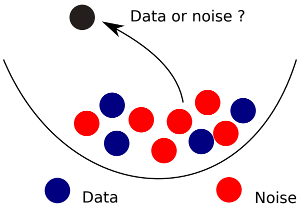
- New method for learning parameters of unnormalised latent variable models.
- Variational theory for noise-contrastive estimation (NCE), which is an estimation framework for unnormalised models.

Noise-contrastive estimation (for unnormalised models)

(Gutmann and Hyvärinen, 2010, 2012)

- Formulates the estimation problem as a classification problem: observed data vs. auxiliary "noise" (reference data with known properties)
- Successful classification $\equiv$ learn the differences between the data and the noise
- differences $+$ known noise properties $\Rightarrow$ properties of the data

- Unsupervised learning by supervised learning
- We used (nonlinear) logistic regression for classification



Noise-contrastive estimation (for unnormalised models)

(Gutmann and Hyvärinen, 2010, 2012)

- NCE procedure:
  1. Choose auxiliary noise distribution $p_{\mathbf{y}}$
  2. Generate auxiliary data $\{\mathbf{y}_1, \ldots, \mathbf{y}_m\}$, $\mathbf{y}_i \sim p_{\mathbf{y}}$
  3. Estimate $\boldsymbol{\theta}$ by maximising

$$J_{\mathrm{NCE}}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{\phi(\mathbf{x}_i; \boldsymbol{\theta})}{\phi(\mathbf{x}_i; \boldsymbol{\theta}) + \nu p_{\mathbf{y}}(\mathbf{x}_i)} + \nu \frac{1}{m}\sum_{i=1}^{m} \log \frac{\nu p_{\mathbf{y}}(\mathbf{y}_i)}{\phi(\mathbf{y}_i; \boldsymbol{\theta}) + \nu p_{\mathbf{y}}(\mathbf{y}_i)} \tag{7}$$

  where $\nu = m/n$

- Nonlinear logistic regression (classification) to learn the differences between the observed data $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ and the auxiliary data $\{\mathbf{y}_1, \ldots, \mathbf{y}_m\}$.

Noise-contrastive estimation (for unnormalised models)

- Choice of $p_{\mathbf{y}}$:
  - simple distributions (e.g. uniform, Gaussian) work surprisingly well
  - can be adaptively chosen to make classification harder or we can take the model learned in the previous iteration (Gutmann and Hyvärinen, 2010), $\rightsquigarrow$ GANs
  - can be chosen dependent on the observed data (Ciwan and Gutmann, ICML, 2018)
  - ...
- NCE has provable convergence guarantees, including MLE as limit for $\nu \to \infty$

  (Gutmann and Hyvärinen, 2012; Riou-Durand and Chopin, 2018)

NCE for latent variables?

$$J_{\mathrm{NCE}}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{\phi(\mathbf{x}_i;\boldsymbol{\theta})}{\phi(\mathbf{x}_i;\boldsymbol{\theta})+\nu p_{\mathbf{y}}(\mathbf{x}_i)} + \nu \frac{1}{m}\sum_{i=1}^{m} \log \frac{\nu p_{\mathbf{y}}(\mathbf{y}_i)}{\phi(\mathbf{y}_i;\boldsymbol{\theta})+\nu p_{\mathbf{y}}(\mathbf{y}_i)}$$

- NCE cannot be used for latent variables models $\phi(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})$. Issue:

$$\phi(\mathbf{u}; \boldsymbol{\theta}) = \int \phi(\mathbf{u}, \mathbf{z}; \boldsymbol{\theta}) \, \mathrm{d}\mathbf{z} \tag{8}$$

  generally intractable

- Approach: derive a variational lower bound $J_{\mathrm{VNCE}}(\boldsymbol{\theta}, q)$ on $J_{\mathrm{NCE}}(\boldsymbol{\theta})$ such that

$$J_{\mathrm{NCE}}(\boldsymbol{\theta}) = \max_{q} J_{\mathrm{VNCE}}(\boldsymbol{\theta}, q), \tag{9}$$

  where J_{VNCE} is computable and defined in terms of $\phi(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})$.
- $q(\mathbf{z}|\mathbf{x})$ is a variational distribution

Variational noise-contrastive estimation (VNCE)

(Rhodes and Gutmann, 2019)

- (Skipping lots of details) Derivation of the bound based on Jensen's inequality, analogue to but not the same as standard variational inference with the log likelihood.
- The variational lower bound is

$$
\begin{aligned}
J_{\mathrm{VNCE}}(\boldsymbol{\theta}, q) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z}|\mathbf{x}_i)} \log \left(\frac{\phi(\mathbf{x}_i, \mathbf{z}; \boldsymbol{\theta})}{\phi(\mathbf{x}_i, \mathbf{z}; \boldsymbol{\theta}) + \nu q(\mathbf{z} \mid \mathbf{x}_i) p_{\mathbf{y}}(\mathbf{x}_i)} \right) \\
+ \nu \frac{1}{m} \sum_{i=1}^{m} \log \left(\frac{\nu p_{\mathbf{y}}(\mathbf{y}_i)}{\nu p_{\mathbf{y}}(\mathbf{y}_i) + \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z}|\mathbf{y}_i)} \left[\frac{\phi(\mathbf{y}_i, \mathbf{z}; \boldsymbol{\theta})}{q(\mathbf{z}|\mathbf{y}_i)} \right]} \right).
\end{aligned} \tag{10}
$$

where $\mathbf{y}_i \sim p_{\mathbf{y}}$ as in NCE.

Variational noise-contrastive estimation (VNCE)

(Rhodes and Gutmann, 2019)

- Key properties of VNCE:
  1. Parameter estimation for unnormalised latent variable models

$$\max_{\boldsymbol{\theta}} J_{\mathrm{NCE}}(\boldsymbol{\theta}) = \max_{\boldsymbol{\theta}, q} J_{\mathrm{VNCE}}(\boldsymbol{\theta}, q) \tag{11}$$

  2. Posterior estimation: optimal q is the true posterior

$$p(\mathbf{z} \mid \mathbf{x}; \boldsymbol{\theta}) = \arg\max_{q} J_{\mathrm{VNCE}}(\boldsymbol{\theta}, q) \tag{12}$$

- Results parallel to those for standard variational inference (VI) for normalised models (see paper for details)
- Significance: Allows us to apply the tricks and tools from standard VI to the unnormalised setting (e.g. EM algorithm, VAEs, etc)

Application: Structure learning with missing data

- Lin et al. (2016) learn undirected graphs for gene expressions in RNAseq data.
- Unnormalised model (truncated normal):

$$\phi(\mathbf{x}; \mathbf{K}, c) = \exp\left(-\frac{1}{2}\mathbf{x}^\top \mathbf{K}\mathbf{x} - c\right) \mathbb{I}(\mathbf{x} \in A), \quad A \subset \mathbb{R}^d \quad (13)$$

$$x_i \perp\!\!\!\perp x_j \mid \text{other variables} \iff K_{ij} = 0 \quad (14)$$

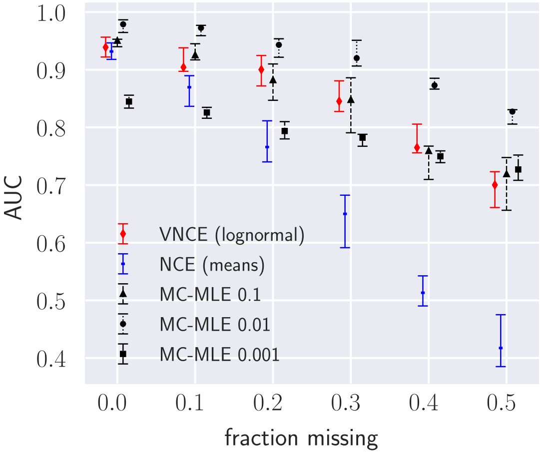
  Cannot compute partition function.
- Previous work used non-negative Score Matching (Hyvärinen, 2007) to estimate the model. Not applicable to data with missing values.
- Data points with missing values were omitted.
- With VNCE, we can treat the missing values as latent variables.

Application: Structure learning with missing data

- Results on synthetic data (20 dimensions, $n = 1000$ samples) with different fractions of missing data
- Graph: ring structure with 10% densely connected nodes (hubs)
- Criterion: accuracy of the learned graph in terms of AUC.
- Learned matrix $\hat{\mathbf{K}}$ yields a graph:
  - If $\hat{K}_{ij}$ below a threshold, then we predict no edge between x_i & x_j.
  - Comparing to ground-truth graph, we obtain a true & false positive rate.
  - Varying the threshold yields curve; area under the curve (AUC) is the criterion (1: best, 0: worst)
- Comparison:
  - mean imputation plus NCE
  - generally infeasible MLE-based gold standard — $\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta})$ can here be approximated using sampling.

Results

- VNCE is significantly better than NCE + fixed imputation ($\nu = 10$)
- Close to a (generally infeasible) MLE-based gold standard.



Summary

- Unnormalised latent variable models: what they are and why they are important
- Estimation is difficult because of two intractable integrals
  - due to the partition function
  - due to marginalisation of latent variables.
- Reviewed noise-contrastive estimation for unnormalised models (previous work)
  - density estimation by classifying between data and noise.
- Theory of variational noise-contrastive estimation
  - a theory that parallels standard (likelihood-based) variational inference but for unnormalised latent variable models
- Application to structure learning from missing data