10-708 L6

12/05/2020

Review readings

Targetted

(*) coupling in POGMs, FOGMs
(*) HMM supervised ML estimation ✓
(*) pseudocounts ✓
(*) mixture models
(*) GMMs
(*) EM for GMMs / soft K-means
(*) EM - compl.; incomplete; expected complete; free-energy entropy co-ord ascent.
(*) Jensen, KL div. ✓
(*) Baum-Welch, BNs, CMM (*)(*)

(*) clarity MAR, MCAR etc.

(*) KL div.

- Wikipedia on pseudocounts (thorough exp.)
- pseudocount / additive smoothing
- "Amt added to no. observed cases in order to change the expected probability in a model of those data, when not known to be 0."
- If frequency of item i is x_i out of N samples: -

$$P_{i,\text{empirical}} = \frac{x_i}{N} \qquad P_{i,\alpha\text{-smoothed}} = \frac{x_i + \alpha}{N + \alpha d}$$

- posterior probability after add. smoothing; effectively increases count x_i by α a priori.
- pseudocount → any non-neg. finite value
- rel. value of pseudocounts represent relative prior exp. prob. of their poss. (?)
- motivated in lecture as a way of addressing natural tendency (Bishop) of ML ~~est~~ point estimation to overfit
- used in Naive Bayes (known as Laplace smoothing)

*) Sum of pseudocounts, may be large, represents estimated weight of prior knowledge compared with all the actual obs. (one for each) when determining expected probability.

(*) HMM supervised estimation,
comfortable with trickier unobserved states setting

Materials (readings)

Jordan (2003) :- ch 10 mixtures and cond. mixtures
(covers GMM, K-means, mixture; heuristic pres. of EM)
+ CMMS + conj grad, N-R methods

Jordan (2003) : ch 11 EM (*)

- Formal pres
+ interpretations

Neal + Hinton (20 : statistical physics interpretation (*)

Borman

Koller (2009) :- ch 19

Jordan (2003) ch 11 - EM

- EM critical to graphical models
- divide and conquer
- (11) : complex dependencies ; model "top-down" using latent variables
- unobserved latent variables :
  - (*) likelihood is marginal probability ; obtained by summing/integrating over latent variables
  - (*) marginalisation couples parameters, obscures underlying structure in likelihood fn.
- E - "inference" → compute expected sufficient statistics
  - → for multinomial (latent) variables ; reduces to computing probability
  - → OR : calculating probability of latent variables given observed variables and current parameter values.
- M-step :- update parameters based on inferred latent variables

11.2 - General setting

- $\underline{X}$ - observables $\underline{Z}$ - latent
- often X, Z decompose into sets of IID pairs
- $\underline{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_N \end{pmatrix}$ $\underline{Z} = \begin{pmatrix} Z_1 \\ \vdots \\ Z_N \end{pmatrix}$ - Note: the X_i are IID variables - observations

$$\underline{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix}$$

- $\underline{X}$ - totality of observed variables $\underline{x}$ - entire observed dataset.
- $\underline{Z}$ - set of all latent

If $\underline{Z}$ observed; ML estimation:-

- Complete log likelihood (CLL):- $\ell_c(\theta; \underline{x}, \underline{z}) := \log p(\underline{x}, \underline{z} | \theta)$ (11.1)
- MLE maximises (11.1); if $p(\underline{x}, \underline{z} | \theta)$ factors in some way, such that sep. components of θ occur in separate factors (*), log has effect of separating likelihood into terms that can be maximised independently

(*) This is what is meant by the est. prob. decouples

- $\underline{Z}$ is not observed, hence; (ILL)
- Incomplete log-likelihood:- $\ell(\theta; \underline{x}) = \log p(\underline{x} | \theta) = \log \sum_{\underline{z}} p(\underline{x}, \underline{z} | \theta)$ (11.2)
- Probability of data $\underline{x}$ - marginal
- No decoupling: log separated from $p(\underline{x}, \underline{z} | \theta)$ by summation.

- $\underline{Z}$ not observed $\Rightarrow$ complete log likelihood is a random quantity; (note $\underline{x}$ is obs.)
- complete-log like cannot be maximised directly

(*) Average over $\underline{z}$ to remove randomness (using an "averaging distri")

The expected-complete log likelihood (*)

$$\langle \ell_c(\underline{\theta}; \underline{x}, \underline{z}) \rangle_q = \sum_{\underline{z}} q(\underline{z}|\underline{x}, \underline{\theta}) \log p(\underline{x}, \underline{z}|\underline{\theta}) \qquad (11.3)$$

(*) This is a deterministic function of parameters $\underline{\theta}$

(OR): If q is well chosen; perhaps expected complete log like (ECLL) will not be far from the log-likelihood (which one?) (probably ILL); serves as surrogate for ILL*

(OR) maximising surrogate does not guarantee a value of $\underline{\theta}$ that maximises likelihood.

BUT (*)(OR): If it may yield an improvement from an initial value of $\underline{\theta}$

(**): If so iterate process and hill-climb

(*) use of averaging distri $q(z|x)$ can provide lower bound on log-likelihood ( )

$$\ell(\underline{\theta}; \underline{x}) = \log p(\underline{x}|\underline{\theta})$$

$$= \log \sum_{\underline{z}} p(\underline{x}, \underline{z}|\underline{\theta})$$

$$= \log \sum_{\underline{z}} q(\underline{z}|\underline{x}) \frac{p(\underline{x}, \underline{z}|\underline{\theta})}{q(\underline{z}|\underline{x})}$$

$$\geq \sum_{\underline{z}} q(\underline{z}|\underline{x}) \log \frac{p(\underline{x}, \underline{z}|\underline{\theta})}{q(\underline{z}|\underline{x})} \qquad (*)$$

$$:= \mathcal{L}(q, \underline{\theta})$$

"ILL (marginal) is a CLL (joint); marginalised over latent"

"importance sampling trick"

(*) Jensen:- $\log(\cdot)$ is concave $\Rightarrow \log[\mathbb{E}_q[\cdot]] \geq \mathbb{E}_q[\log[\cdot]]$

- $\mathcal{L}(q, \underline{\theta})$ - auxiliary function
- for arbitrary $q(\cdot)$ distri; auxiliary function $\mathcal{L}(q, \underline{\theta})$ is a lower bound for the (incomplete) log likelihood.

EM (*) EM as coordinate ascent on auxiliary $\mathcal{L}(q, \underline{\theta})$

E-step: $q^{(t+1)} = \underset{q}{\operatorname{argmax}}\ \mathcal{L}(q, \underline{\theta}^{(t)})$

M-step: $\underline{\theta}^{(t+1)} = \underset{\underline{\theta}}{\operatorname{argmax}}\ \mathcal{L}(q^{(t+1)}, \underline{\theta})$

(*) Max lower bound on $\ell(\underline{\theta}; \underline{x})$ (LL) can maximise $\ell(\underline{\theta}; \underline{x})$

(*) M-step as maximisation of ECLL

- decompose auxiliary $\mathcal{L}(q, \underline{\theta}^{(t+1)})$ // (lower bound)

$$\mathcal{L}(q, \underline{\theta}) = \sum_{\underline{z}} q(\underline{z}|\underline{x}) \frac{\log p(\underline{x}, \underline{z}|\underline{\theta})}{q(\underline{z}|\underline{x})}$$

$$= \sum_{\underline{z}} q(\underline{z}|\underline{x}) \log p(\underline{x}, \underline{z}|\underline{\theta}) - \sum_{\underline{z}} q(\underline{z}|\underline{x}) \log q(\underline{z}|\underline{x})$$

$$= \langle \ell_c(\underline{\theta}; \underline{x}, \underline{z}) \rangle_q - H_q$$

(**) $\mathcal{L}(q, \underline{\theta})$ is made up of expected-complete log likelihood

(*) And also H_q, which is both indep of param $\underline{\theta}$

(*) H_q - entropy of $q(\cdot)$; - KL div of distn with itself (?)

(*) E-step: setting $q(\cdot)$ as posterior distn over latent variables, given data and parameters yields maximum: $q^{(t+1)}(\underline{z}|\underline{x}) = p(\underline{z}|\underline{x}, \underline{\theta}^{(t)})$

Evaluate $\mathcal{L}(q, \underline{\theta}^{(t)})$ for $q^{(t+1)} = p(\underline{z}|\underline{x}, \underline{\theta}^{(t+1)})$

$\underline{\theta}^{(t)}$?

$$\mathcal{L}(p(\underline{z}|\underline{x}, \underline{\theta}^{(t)}) | \underline{\theta}^{(t)}) = \sum_{\underline{z}} p(\underline{z}|\underline{x}, \underline{\theta}^{(t)}) \log \frac{p(\underline{x}, \underline{z}|\underline{\theta})}{p(\underline{z}|\underline{x}, \underline{\theta}^{(t)})}$$

stack exchange

$$= \sum_{\underline{z}} p(\underline{z}|\underline{x}, \underline{\theta}^{(t)}) \log p(\underline{x}|\underline{\theta}^{(t)})$$

(2) (W) (AI) (?) - Bayes rule

$\rightarrow \mathbb{E}_{p(\underline{z}|\underline{x}, \underline{\theta}^{(t)})}[\log p(\underline{x}|\underline{\theta})]$

$$= \log p(\underline{x}|\underline{\theta}^{(t)})$$

$$= \ell(\underline{\theta}^{(t)}; \underline{x})$$

(*) $\ell(\underline{\theta}; \underline{x})$ is an upper bound for $\mathcal{L}(q, \underline{\theta}^{(t)}) \Rightarrow \mathcal{L}(q, \underline{\theta}^{(t)})$ maximised by setting $q(\cdot) = p(\underline{z}|\underline{x}, \underline{\theta}^{(t)})$

setting

(*) M-step: $q(\cdot) = p(\underline{z} | \underline{x}, \underline{\theta}^{(t)})$ maximises $l(\underline{\theta}; \underline{x})$ and hence $\mathcal{L}(q, \underline{\theta}^{(t)})$ via KL-divergences

(*) Appendix material on KL-divergence perspective + alternating minimisation

---

(*) intuitions/exp:-

(*) $p(\underline{z} | \underline{x}, \underline{\theta}^{(t)})$ as best guess of latents, conditioned on data

(*) use best guess distn to compute ECLL (E-step)

(*) Maximise ECLL wrt parameters to yield new $\underline{\theta}^{(t+1)}$ (M-step)

(*) Given improvement, make better guess $p(\underline{z} | \underline{x}, \underline{\theta}^{(t+1)})$; iterate.

(*) Intuition:-

- Effect of EM iteration on log-likelihood $l(\underline{\theta}; \underline{x})$?
- M step: select $\underline{\theta}$ parameters to increase a lower bound ($\mathcal{L}(q^{(t+1)}, \underline{\theta})$)

(x)(x) , on likelihood $l(\underline{\theta}; \underline{x})$

- increasing lower bound on function $\not\Rightarrow$ increasing function itself

(*) BUT: in E-step, close gap with approp. choice of $q(\cdot)$ distn

$$l(\underline{\theta}^{(t)}; \underline{x}) = \mathcal{L}(q^{(t+1)}, \underline{\theta}^{(t)})$$

(N): some graphics/illustr would help

(*) EM - hill climbing

in log likelihood $l(\underline{\theta}; \underline{x})$

- ~~via~~ indirect hill climbing by co-ordinate ascent in auxiliary $\mathcal{L}(q, \underline{\theta})$
- tactic → maximisation of ECLL rather than ILL