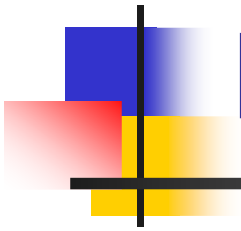
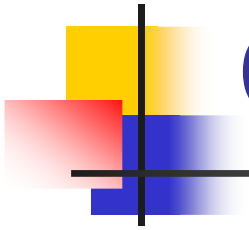


Recitation 7

Hetunandan



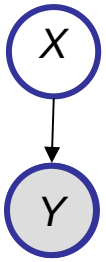
CRF learning

- Review of CRF likelihood
- CRF learning
- Proof that gradient of the log-partition function in an exponential family is the expectation of the sufficient statistics

Factor analysis

- In regression, both x and y are observed and task is to find relation between them
- In factor analysis, y is observed and we wish to obtain latent (low-dim) x
- Can treat this as unsupervised regression

Factor analysis



$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; 0, I)$$

$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}; \mu + \Lambda\mathbf{x}, \Psi)$$

Λ is called a factor loading matrix
Ψ is diagonal.

- The loading matrix Λ transforms low dim latent x to high dim y
- Ψ can be treated as noise parameter

Review of Linear Algebra

$$M = \begin{bmatrix} E & F \\ G & H \end{bmatrix}$$

- M^{-1} can be expressed in terms of H^{-1} and $(E-FH^{-1}G)^{-1}$
- $(E-FH^{-1}G)$ is called the *Schur's complement* of H in M(written M/H)
- Matrix Inverse Lemma

$$\left(E - FH^{-1}G\right)^{-1} = E^{-1} + E^{-1}F\left(H - GE^{-1}F\right)^{-1}GE^{-1}$$

EM for Factor Analysis

- Complete log likelihood

$$l_c(\theta, D) = -\frac{N}{2}\log|I| - \frac{1}{2}\sum_n x_n^T x_n - \frac{N}{2}\log|\Psi| - \frac{1}{2}\sum_n (y_n - \Lambda x_n)^T \Psi^{-1}(y_n - \Lambda x_n)$$

writing scalars as traces(trace trick)

$$= -\frac{N}{2}\log|\Psi| - \frac{1}{2}\sum_n \mathrm{tr}\left[x_n^T x_n\right] - \frac{1}{2}\sum_n \mathrm{tr}[(y_n - \Lambda x_n)^T \Psi^{-1}(y_n - \Lambda x_n)],$$

using $\mathrm{tr}[ABC] = \mathrm{tr}[CAB] = \mathrm{tr}[BCA]$ on second and third terms

$$= -\frac{N}{2}\log|\Psi| - \frac{1}{2}\sum_n \mathrm{tr}\left[x_n x_n^T\right] - \frac{1}{2}\sum_n \mathrm{tr}[(y_n - \Lambda x_n)^T (y_n - \Lambda x_n)\Psi^{-1}]$$

$$= -\frac{N}{2}\log|\Psi| - \frac{1}{2}\sum_n \mathrm{tr}\left[x_n x_n^T\right] - \frac{N}{2}\mathrm{tr}\left[\mathbf{S}\Psi^{-1}\right], \qquad \text{where } \mathbf{S} = \frac{1}{N}\sum_n (y_n - \Lambda x_n)(y_n - \Lambda x_n)^T$$

E step

- Expected log likelihood

$$\langle \ell_c(\theta, D) \rangle = -\frac{N}{2}\log|\Psi| - \frac{1}{2}\sum_n \mathrm{tr}\left[\left\langle X_n X_n^T \right\rangle\right] - \frac{N}{2}\mathrm{tr}\left[\langle S \rangle \Psi^{-1}\right]$$

$$\langle S \rangle = \frac{1}{N}\sum_n (y_n y_n^T - y_n \left\langle X_n^T \right\rangle \Lambda^T - \Lambda \left\langle X_n^T \right\rangle y_n^T + \Lambda \left\langle X_n X_n^T \right\rangle \Lambda^T)$$

- Therefore, sufficient statistics for E step

$$\langle X_n \rangle = E[X_n \mid y_n]$$

$$\left\langle X_n X_n^T \right\rangle = Var[X_n \mid y_n] + E[X_n \mid y_n] E[X_n \mid y_n]^T$$

M-step for Factor Analysis

- Taking derivative wrt. parameters.

$$\frac{\partial}{\partial \Psi^{-1}}\langle \ell_c \rangle = \frac{\partial}{\partial \Psi^{-1}}\left(-\frac{N}{2}\log|\Psi| - \frac{1}{2}\sum_n \mathrm{tr}\left[\langle X_n X_n^T \rangle\right] - \frac{N}{2}\mathrm{tr}\left[\langle \mathbf{S} \rangle \Psi^{-1}\right]\right)$$

$$= \frac{\partial}{\partial \Psi^{-1}}\left(-\frac{N}{2}\log|\Psi| - \frac{N}{2}\mathrm{tr}\left[\langle \mathbf{S} \rangle \Psi^{-1}\right]\right)$$

$$= -\frac{N}{2}\frac{\partial}{\partial \Psi^{-1}}\log|\Psi| - \frac{N}{2}\frac{\partial}{\partial \Psi^{-1}}\mathrm{tr}\left[\langle \mathbf{S} \rangle \Psi^{-1}\right]$$

Ψ is diagonal, so $|\Psi| = 1/|\Psi^{-1}|$; use $\frac{\partial}{\partial A}\log|A| = A^{-T}$

$$= \frac{N}{2}\Psi - \frac{N}{2}\frac{\partial}{\partial \Psi^{-1}}\mathrm{tr}\left[\langle \mathbf{S} \rangle \Psi^{-1}\right]$$

using $\frac{\partial}{\partial A}\mathrm{tr}[\mathrm{BA}] = \mathrm{B}^{\mathrm{T}}$

$$= \frac{N}{2}\Psi - \frac{N}{2}\langle \mathbf{S} \rangle^T \qquad \Rightarrow \qquad \Psi^{t+1} = diag(\langle \mathbf{S} \rangle^T) = diag(\langle \mathbf{S} \rangle)$$

M step contd

$$\frac{\partial}{\partial \Lambda}\langle \ell_c \rangle = \frac{\partial}{\partial \Lambda}\left(-\frac{N}{2}\log|\Psi| - \frac{1}{2}\sum_n \mathrm{tr}\left[\langle X_n X_n^T \rangle\right] - \frac{N}{2}\mathrm{tr}\left[\langle \mathbf{S} \rangle \Psi^{-1}\right] \right) = -\frac{N}{2}\Psi^{-1}\frac{\partial}{\partial \Lambda}\langle \mathbf{S} \rangle$$

$$= -\frac{N}{2}\Psi^{-1}\frac{\partial}{\partial \Lambda}\left(\frac{1}{N}\sum_n (y_n y_n^T - y_n \langle X_n^T \rangle \Lambda^T - \Lambda \langle X_n^T \rangle y_n^T + \Lambda \langle X_n X_n^T \rangle \Lambda^T) \right)$$

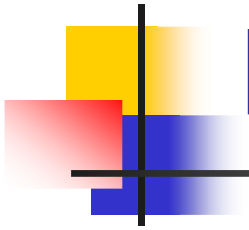
$$= -\frac{N}{2}\Psi^{-1}\frac{\partial}{\partial \Lambda}\left(\frac{1}{N}\sum_n (-y_n \langle X_n^T \rangle \Lambda^T - \Lambda \langle X_n^T \rangle y_n^T + \Lambda \langle X_n X_n^T \rangle \Lambda^T) \right)$$

$$\text{using } \frac{\partial A}{\partial \Lambda} = \left(\frac{\partial A}{\partial \Lambda^T} \right)^T \text{ and } \frac{\partial}{\partial \Lambda}\mathrm{tr}[BA^T CA] = 2CAB$$

$$= \Psi^{-1}\sum_n y_n \langle X_n^T \rangle - \Psi^{-1}\Lambda \sum_n \langle X_n X_n^T \rangle$$

$$\Rightarrow \quad \Lambda^{t+1} = \left(\sum_n y_n \langle X_n^T \rangle \right)\left(\sum_n \langle X_n X_n^T \rangle \right)^{-1}$$

Additional References Mentioned



- Old and New Matrix Algebra Useful for Statistics, Tom Minka http://research.microsoft.com/~minka/papers/matrix/
- The Matrix cookbook http://matrixcookbook.com/
- A Unifying Review of Linear Gaussian Models, Sam Roweis and Zoubin Ghahramani, *Neural Computation*, 1999