Homework 4

Due Date: June 9 , 2017

Problem 1. In linear algebra, the singular value decomposition (**SVD**) is a factorization of a real matrix $\mathbf{X}$ as:

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T \tag{1}$$

If the dimension of $\mathbf{X}$ is $m \times n$, where without loss of generality $m \geq n$, $\mathbf{U}$ is an $m \times n$ matrix, $\mathbf{S}$ is an $n \times n$ diagonal matrix and $\mathbf{V}^T$ is also an $n \times n$ matrix. Furthermore, $\mathbf{U}$ and $\mathbf{V}$ are orthonormal matrices: $\mathbf{U}\mathbf{U}^T = \mathbf{I}$ and $\mathbf{V}\mathbf{V}^T = \mathbf{I}$.

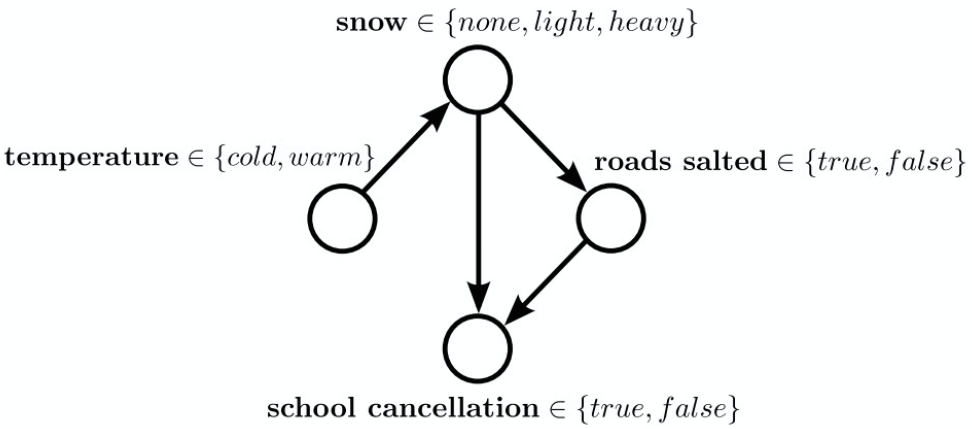
Consider a dataset of observations $\{\mathbf{x_n}\}$ where $n = 1, ..., N$. We assume that the examples are zero-centered such that $\bar{\mathbf{x}} = \sum_{n=1}^{N} \mathbf{x}_n = 0$. PCA algorithm computes the covariance matrix:

$$\mathbf{S} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n^T \tag{2}$$

The principal components $\{\mathbf{u}_i\}$ are eigenvectors of $\mathbf{S}$.
Let $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_N]$, a $D \times N$ matrix where each column is one example $\mathbf{x}_n$. If $\mathbf{U}\mathbf{S}'\mathbf{V}^T$ is a **SVD** of $\mathbf{X}$,

1. Show that the principal components $\{\mathbf{u}_i\}$ are columns of $\mathbf{U}$. This shows the relationship between PCA and SVD.
2. When the number of dimensions is much larger than the number of datapoints ($D \gg N$), is it better to do PCA by using the covariance matrix or using SVD?

Problem 2. Below is depicted a graphical model with four *discrete* random variables that can be used to predict whether school will be closed due to inclement weather.



a. Answer the following questions about the conditional independence structure in the model:

- **i.** Which variables are independent of **temperature** given that **snow** is observed?
- **ii.** Which variables are independent of **snow** given that no variables are observed?
- **iii.** Which variables are independent of **snow** given that **temperature** is observed?
- **iv.** Which variables are independent of **school cancellation** given that **snow** and **roads salted** are observed?

Suppose the random variables in the above graphical model have the following parameters:

The variable **temperature** does not depend on any other variable, and so it has the following prior distribution:

$p(\textbf{temperature} = cold)$	$p(\textbf{temperature} = warm)$
0.4	0.6

The variable **snow** only depends on the value of **temperature**:

temperature	$p(\textbf{snow} = none \mid \textbf{temp})$	$p(\textbf{snow} = light \mid \textbf{temp})$	$p(\textbf{snow} = heavy \mid \textbf{temp})$
cold	0.4	0.4	0.2
warm	0.9	0.08	0.02

The variable **roads salted** only depends on the value of **snow**:

snow	$p(\textbf{roads salted} = T \mid \textbf{snow})$	$p(\textbf{roads salted} = F \mid \textbf{snow})$
none	0.01	0.99
light	0.9	0.1
heavy	0.97	0.03

The variable **school cancellation** depends on both **snow** and **roads salted**. For brevity, the condition "**snow**, **roads slated**" is replaced with "...":

snow	**roads salted**	$p(\textbf{school cancellation} = T \mid \ldots)$	$p(\textbf{school cancellation} = F \mid \ldots)$
none	T	0.01	0.99
none	F	0.01	0.99
light	T	0.2	0.8
light	F	0.4	0.6
heavy	T	0.95	0.05
heavy	F	0.99	0.01

b. The joint probability is given by $p(\textbf{temperature}, \textbf{snow}, \textbf{roads salted}, \textbf{school cancellation})$. Write the factorized form of the joint probability (as a product of simpler probabilities) for the model above.

c. Compute the distribution $p(\textbf{school cancellation} \mid \textbf{snow} = \textit{light})$.

Problem 3. When attempting to run the EM algorithm, it may sometimes be difficult to perform the M step exactly recall that we often need to implement numerical optimization to perform the maximization, which can be costly. Therefore, instead of finding the global maximum of our lower bound on the log-likelihood, and alternative is to just increase this lower bound a little bit, by taking one step of gradient ascent, for example. This is commonly known as the Generalized EM (GEM) algorithm.

Put slightly more formally, recall that the M-step of the standard EM algorithm performs the maximization

$$\theta := \arg\max_\theta \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}$$

The GEM algorithm, in contrast, performs the following update in the M-step:

$$\theta := \theta + \alpha \nabla_\theta \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}$$

where α is a learning rate which we assume is chosen small enough such that we do not decrease the objective function when taking this gradient step.

1. Prove that the GEM algorithm described above converges. To do this, you should show that the the likelihood is monotonically improving, as it does for the EM algorithm — i.e., show that $\ell(\theta^{(t+1)}) \geq \ell(\theta^{(t)})$.

2. Instead of using the EM algorithm at all, suppose we just want to apply gradient ascent to maximize the log-likelihood directly. In other words, we are trying to maximize the (non-convex) function

$$\ell(\theta) = \sum_i \log \sum_{z^{(i)}} p(x^{(i)}, z^{(i)}; \theta)$$

so we could simply use the update

$$\theta := \theta + \alpha \nabla_\theta \sum_i \log \sum_{z^{(i)}} p(x^{(i)}, z^{(i)}; \theta)$$

Show that this procedure in fact gives the same update as the GEM algorithm described above.