Homework # 4 - Due Friday 26th of October

NAME:_____

Signature:_____

STD. NUM: _____

1. **Image Compression:** Convert a photo of yours to .png format. Then load it with python and compute its SVD compression as explained in the lecture notes and in class. Choose a number of eigenvalues (truncation threshold) and explain why you chose this value below. How much compression have you achieved? Answer this question below and hand in all code and your image before and after compression.

2. **PCA:** As explained in class in the context of digit images (the images of 2's), implement a PCA projection of about 10 or 20 data instances of your liking (images, text documents, objects in a excel document, wikipedia entities, etc.) to a 2D PCA layout (like the ones for countries, hand gestures and numbers in the lectures). If the data instances are images, transform them to be of the same size, say 40 by 40 pixels, and grayscale. For help on generating 2D PCA layouts, see the posting of Bobak Shahriari on "annotating scatter plots with images" in the google group or visit

   `http://matplotlib.org/examples/pylab_examples/demo_annotation_box.html`.

   Hand in all the code and your 2D PCA layout.

   For example: for a hotel browsing app, I would select 10 Vancouver hotels in *tripadvisor*. As the attributes for each hotel, I would use the 5 traveler ratings (excellent, very good, average, poor, terrible), the *tripadvisor* rank and the cheapest room price. I would then form a matrix of 10 rows by 7 columns, and use the SVD to compute the 2D PCA components $\mathbf{U}_2\mathbf{\Sigma}_2$. Finally, I would do a scatter plot of the hotels and, at the location of the hotel point in 2D, I would insert either the name of the hotel or the image of the hotel.

   Be creative and enjoy the exercise!

3. **Learning Bayesian networks:** For the Fritz network that was discussed in the lecture notes, the joint distribution of the i-th observation is:

$$P(T_i, M_i, S_i, F_i | \theta, \alpha, \gamma_{1:2}, \beta_{1:4}) = P(T_i | S_i, F_i, \beta_{1:4}) P(F_i | M_i, \gamma_{1:2}) P(M_i | \theta) P(S_i | \alpha),$$

where each distribution is Bernoulli:

$$
\begin{aligned}
P(M_i | \theta) &= \theta^{\mathbb{I}(M_i=1)} (1-\theta)^{\mathbb{I}(M_i=0)} \\
P(S_i | \alpha) &= \alpha^{\mathbb{I}(S_i=1)} (1-\alpha)^{\mathbb{I}(S_i=0)} \\
P(F_i | M_i=0, \gamma_1) &= \gamma_1^{\mathbb{I}(F_i=1|M_i=0)} (1-\gamma_1)^{\mathbb{I}(F_i=0|M_i=0)} \\
P(F_i | M_i=1, \gamma_2) &= \gamma_2^{\mathbb{I}(F_i=1|M_i=1)} (1-\gamma_2)^{\mathbb{I}(F_i=0|M_i=1)} \\
P(T_i | S_i=0, F_i=0, \beta_1) &= \beta_1^{\mathbb{I}(T_i=1|S_i=0,F_i=0)} (1-\beta_1)^{\mathbb{I}(T_i=0|S_i=0,F_i=0)} \\
P(T_i | S_i=0, F_i=1, \beta_2) &= \beta_2^{\mathbb{I}(T_i=1|S_i=0,F_i=1)} (1-\beta_2)^{\mathbb{I}(T_i=0|S_i=0,F_i=1)} \\
P(T_i | S_i=1, F_i=0, \beta_3) &= \beta_3^{\mathbb{I}(T_i=1|S_i=1,F_i=0)} (1-\beta_3)^{\mathbb{I}(T_i=0|S_i=1,F_i=0)} \\
P(T_i | S_i=1, F_i=1, \beta_4) &= \beta_4^{\mathbb{I}(T_i=1|S_i=1,F_i=1)} (1-\beta_4)^{\mathbb{I}(T_i=0|S_i=1,F_i=1)}
\end{aligned}
$$

(a) Derive the ML estimates of $\theta, \alpha, \gamma_{1:2}, \beta_{1:4}$ for the dataset in the slides. Hint: we did some of these estimates already in class.

(b) Derive the posterior mean estimates of $\theta, \alpha, \gamma_{1:2}, \beta_{1:4}$ assuming that we use the same $Beta(1,1)$ prior for each of the parameters. Start by writing each of the 8 posterior distributions. For example,

$$\begin{aligned} P(\theta|M_{1:5}) &\propto \prod_{i=1}^{5} P(M_i|\theta)p(\theta) \\ &\propto \theta^4(1-\theta)^1\theta^{1-1}(1-\theta)^{1-1} \\ &= \theta^{5-1}(1-\theta)^{2-1} \end{aligned}$$

and, consequently, the posterior mean estimate of θ is $\mathbb{E}(\theta|M_{1:5}) = \frac{5}{5+2} = 5/7$.

(c) Calculate $P(M|T = 0)$ from the ML estimates of the parameters.

(d) Calculate $P(M|T = 0)$ from the posterior mean estimates of the parameters.

(e) As explained in the last slide of the lectures on Bayesian network learning, assume we have another model with an arrow from M to S. In this case, $P(S_i|\alpha)$ gets replaced by $P(S_i|M_i, \alpha_1, \alpha_2)$. For the validation dataset in that slide, which model provides a better prediction? Hint: evaluate $\prod_{i=1}^{2} P(T_i, M_i, S_i, F_i|\theta, \alpha, \gamma_{1:2}, \beta_{1:4})$, where i is the index over the two items in the validation dataset, for both models. Note that for model 2, you first have to learn $\alpha_{1:2}$.

4. **Bayesian linear regression** Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ and $\mathbf{y} \in \mathbb{R}^{n \times 1}$. Assume the likelihood is Gaussian:

$$p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\Sigma}) = |2\pi\boldsymbol{\Sigma}|^{-1/2} e^{-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T \boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})}.$$

Assume that the prior for $\boldsymbol{\theta}$ is also Gaussian:

$$p(\boldsymbol{\theta}) = |2\pi\boldsymbol{\Delta}|^{-1/2} e^{-\frac{1}{2}\boldsymbol{\theta}^T \boldsymbol{\Delta}^{-1} \boldsymbol{\theta}}.$$

(a) Using Bayes rule and completing squares, derive an expression for the posterior distribution of $\boldsymbol{\theta}$. In this part, assume that the covariance $\boldsymbol{\Sigma}$ is given. State clearly what the mean and variance of the posterior are.

(b) State the conditions under which the posterior mean would be equivalent to the ridge and maximum likelihood estimators.

.