Midterm

CS536, Machine Learning, Spring 2015

April 4, 2015

This test is a take-home test. You have until 11:59pm on April 9 to answer the questions and submit your test. The test should be submitted electronically through Sakai.

Name:	
ID:	

Problem	Score	Max. score
1		100
2		100
Max. Total		200 (200 perfect grade)

Problem 1

In this problem you will consider different clustering methods and investigate their respective advantages and disadvantages. Complete each of the specified tasks and fully answer all questions. For programming parts of the problem, you can use either MATLAB or Python. Submit all code and data.

1. Read the paper J. Friedman, T. Hastie, and R. Tibshirani, ”Sparse inverse covariance estimation with the graphical lasso,” Biostat. (2008) 9 (3): 432-441.

2. Create a covariance matrix $\mathbf{\Sigma}_0$ according to the following specification:

   - Uniformly sample $x_1, ..., x_{20}$ (a 2-dim random variable) from the unit square $[0, 1]^2$.
   - Construct a **symmetric** binary adjacency matrix $\boldsymbol{A} = [a_{ij}]_{20\times20}$, $a_{ij} \in \{0, 1\}$, such that
   $$Pr(a_{ij} = 1) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2s}\|x_i - x_j\|^2)$$
   and zero otherwise. Use $s = 1/8$. Note again that the matrix needs to be symmetric, so sample only the lower triangular part of the matrix and then replicate it in the upper triangle.
   - Construct the inverse covariance (precision) matrix $\mathbf{\Lambda}_0 = [\lambda_{0,ij}]_{20\times20}$ such that
   $$\lambda_{0,ij} = \begin{cases} 1, & i = j \\ 0.245, & a_{ij} = 1 \\ 0, & a_{ij} = 0 \end{cases}$$
   - Now set $\mathbf{\Sigma}_0 = \mathbf{\Lambda}_0^{-1}$.

   Show (plot) the structure of $\boldsymbol{A}$ as an adjacency graph (use x_i as the node coordinates) and show $\mathbf{\Lambda}_0$ and $\mathbf{\Sigma}_0$ as images. Comment on the density of all structures.

3. Create a dataset $\mathcal{D}_0$ of $N = 1000$ points by sampling from the Gaussian density $\mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma_0})$, where $\mu_i = 1.5,\ i = 1, \ldots, 20$. Plot the sample covariance of this data as well as the sample precision. Comment on their similarities/differences compared to the ”true” model.

4. Create a covariance matrix $\mathbf{\Sigma}_1$ according to the following specification:

   - Set precision matrix $\mathbf{\Lambda}_1 = [\lambda_{1,ij}]_{24\times24}$ as
   $$\mathbf{\Lambda}_1 = \begin{bmatrix} \mathbf{\Lambda}_0 & \boldsymbol{\lambda}_1 \\ \boldsymbol{\lambda}_1^T & \boldsymbol{I} \end{bmatrix}.$$
   where $\mathbf{\Lambda}_0$ was constructed in the previous task.
   Set the off-diagonal blocks $\boldsymbol{\lambda}_1$ such that each of the new four elements/nodes $i = 21, 22, 23, 24$ is connected to a subset of nodes $i = 1, \ldots, 5$, $i = 6, \ldots, 10$, $i = 11, \ldots, 15$, and $i = 16, \ldots, 20$, respectively, with $\lambda_{1,ij} = 0.45$. Note that, for example, this means that $\lambda_{1,21,8} = 0$ because node 21 is not connected to node 8.
   - Now set $\mathbf{\Sigma}_1 = \mathbf{\Lambda}_1^{-1}$.

   Show (plot) the structure of $\mathbf{\Lambda}_1$ and $\mathbf{\Sigma}_1$ as images. Comment on the density of these new structures.

5. Create a dataset $\mathcal{D}_1$ of $N = 1000$ points by sampling from the Gaussian density $\mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma_1})$, where $\mu_i = 1.5,\ i = 1, \ldots, 24$ and **retaining only** the first 20 dimensions of each sample (i.e., discarding dimensions $21 - 24$.). Plot the sample covariance of this data as well as the sample precision. Comment on their similarities/differences compared to the ”true” model.

6. Cluster the first $N = 750$ samples in datasets $\mathcal{D}_0$ and $\mathcal{D}_1$ using k-means. Set $k = 2, \ldots, 6$ and initialize each model using 10 different random point assignments (i.e., you will construct ten different models for each k). Select the best model among the ten obtained models, according to the training reconstruction error. Then calculate and plot the $k - means$ reconstruction error on the holdout data (remaining 250 samples) for the five models $k = 2, \ldots, 6$, for each dataset.

- Plot the clustered points, for $k = 2, \ldots, 6$ in the x coordinate system (defined in Item 2) and color them according to the cluster membership.
- Comment on the results of this clustering. How does it relate to the original adjacency structure used to synthesize the data?

7. Repeat the k-means task using the Gaussian mixture model clustering approach. Report the cluster means and inverse covariances. Then answer the same two questions as in the k-means problem.

8. Repeat the clustering task using spectral clustering. Use 5 nearest neighbor (5-NN) affinity constructed from the sample inverse covariance (precision) matrices.

    - Plot the 5-NN adjacency graph used to cluster the data.
    - Plot the clustered points in the x coordinate system (defined in Item 2) and color them according to the cluster membership.
    - Comment on the results of this clustering. How does it relate to the original adjacency structure used to synthesize the data?

9. Use the Graphical Lasso method described in the paper above to find the sparse inverse covariance matrix, then use spectral clustering on this matrix to cluster the points in the two datasets. (Note: you can use an existing implementation of Graphical Lasso, e.g., in sklearn).

    - Explain, in your own words, how the Graphical Lasso method works. One paragraph explanation is sufficient.
    - Plot the reconstructed inverse adjacency matrix.
    - Plot the clustered points in the x coordinate system (defined in Item 2) and color them according to the recovered cluster membership.
    - Comment on the results of this clustering. How does it relate to the original adjacency structure used to synthesize the data?

10. Intuitively explain what the synthetic model in Item 4 does and what is the role of the removed dimensions $21 - 24$. With that in mind, which of the four clustering methods worked best for the two data sets and why?

Problem 2

In this problem you will consider the task of modeling the data that arises from processes with latent variables (i.e., when we only observe partial information of the full process). Complete each of the specified tasks and fully answer all questions.

1. Read the paper M. Jordan et al., ”An Introduction to Variational Methods for Graphical Models,” Machine Learning, 37, 183233 (1999). In particular, focus on Section 4, ”Basics of variational methodology” and Section 6, ”The block approach”.

2. Explain briefly how the EM algorithm arises from the convex duality approach and the KL-divergence minimization in eq(41) of the paper and why the KL was used as the objective. Why do we want to have an approximate density Q (e.g., a simpler density than the true $P(H|E)$)?

3. In the paper the authors use the KL-divergence $KL(Q\|P)$. How does the variance of the approximating Q under constraints and the optimal Q when no constraints exist (i.e., $P(H|E)$) compare? Is it always smaller / larger / equal to the true variance? Clearly justify your answer.

4. Could one use the KL-divergence $KL(P\|Q)$? How would this impact the posterior variance of Q?

5. Consider the task of modeling a dataset $\mathcal{D} = \{x\}_{i=1}^{N}$ with a mixture of K Gaussians. Let z_i denote that latent variables (assignments) of the i-th data point to one of the K clusters (mixture components), $z_i \in \{1, 2, \dots, K\}$. Suppose that the component models have known variances, i.e., $x|z \sim \mathcal{N}(x; \mu_z, 1)$. Also suppose that $\mu_z \sim \mathcal{N}(0, 1)$.

   In class, when we discussed Gaussian mixture learning, we focused on finding the posterior distribution $P(z|x, \mu, \mathcal{D})$ and then estimating the "best" values of μ_k (so called point-estimates) using an EM-algorithm. We did not care to find the distribution of each component's mean, μ_k. Suppose now that we want to accomplish a slightly more challenging task, i.e., find the full posterior of both z's and μ's. In other words, find $Pr(z_1, \dots, z_N, \mu_1, \dots, \mu_K|\mathcal{D})$. In that context answer the following questions:

   - Describe how you would compute the true $Q(z_1, \dots, z_N, \mu_1, \dots, \mu_K|\mathcal{D})$ and why this could be computationally challenging.
   - Consider a simpler, approximating distribution $Pr(z_1, \dots, z_N, \mu_1, \dots, \mu_K|\mathcal{D}) = \prod_{i=1}^{N} Q(z_i) \prod_{k=1}^{K} Q(\mu_k)$. Specify each factor's form (i.e., $Q(z_i)$ and $Q(\mu_k)$) and justify why you picked those forms (distributions).
   - Justify why the proposed factorization may help resolve the (computational) challenge of computing the true posterior.
   - Describe how you would compute the parameters of this approximating distribution ($Q(z_i)$ and $Q(\mu_k)$) under the variational inference framework. You do not need to compute the actual estimation equations, just clearly (mathematically) specify how you would do it and comment on how complex this would be. (In fact, for some reasonable choices of Q's this is very easy to do analytically).