CS 475 Machine Learning: Homework 5 Analytical (70 points)

Assigned: Tuesday, Nov. 16th, 2021
Due: Tuesday, Nov. 30th, 2021, 11:59 pm US/Eastern

Partner 1: NAME (JHED), Partner 2: NAME (JHED)

Instructions

We have provided this LaTeX document for turning in this homework. We give you one or more boxes to answer each question. The question to answer for each box will be noted in the title of the box. You can change the size of the box if you need more space.

Other than your name, do not type anything outside the boxes. Leave the rest of the document unchanged.

Do not add text outside of the answer boxes. You are allowed to make boxes larger if needed.

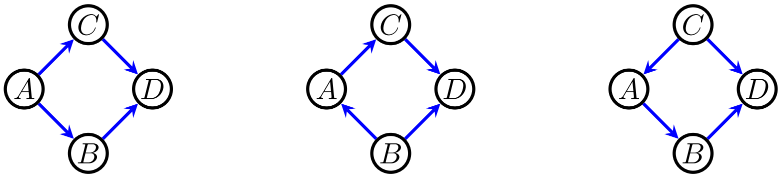
We strongly recommend you review your answers in the generated PDF to ensure they appear correct. We will grade what appears in the answer boxes in the submitted PDF, NOT the original latex file.

Observational Equivalence of DAGs and the GES Algorithm

Question 1.

Two DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ are said to be *observationally equivalent* if the list of conditional independences implied by the d-separation criterion in both graphs is the same. In other words, $\mathcal{G}_1$ and $\mathcal{G}_2$ are observationally equivalent if they imply the same statistical model.

A result by Thomas Verma and Judea Pearl states that $\mathcal{G}_1$ and $\mathcal{G}_2$ are observationally equivalent if and only if they agree on edge adjacencies (in other words, if V_i and V_j share an edge in $\mathcal{G}_1$, then V_i and V_j share an edge in $\mathcal{G}_2$, and vice versa – ignoring edge orientation), and agree on unshielded colliders. An unshielded collider is a structure of the form $V_i \rightarrow V_k \leftarrow V_j$ such that V_i and V_j do not share an edge. As an example, the following three graphs are observationally equivalent:



The above three DAGs give the same model: $(B \perp\!\!\!\perp C \mid A)$, $(D \perp\!\!\!\perp A \mid B, C)$.

(a) Write out all equivalence classes for DAGs with three vertices. How many equivalence classes are there?

(b) Assume all data is binary. Write down the dimension of each model corresponding to each equivalence class in (a).

(c) Create an undirected graph representing the discrete state space for structure learning, where vertices represent equivalence classes in (a), and there is an edge connecting any two classes where a DAG in one class differs from a DAG in another class by *addition or deletion* precisely one $\rightarrow$ edge.

(d) Assume the GES algorithm performs the following sequence of edge additions and deletions (starting from the empty graph): add $A \rightarrow B$, add $B \rightarrow C$, add $A \rightarrow C$, remove $B \rightarrow C$. Write down all sequences of equivalence classes consistent with this set of edge additions and removals. Note: there could be more than one such sequence.

(e) Consider a DAG $V_1 \rightarrow V_2 \rightarrow V_3 \rightarrow V_4 \rightarrow V_k$. How many DAGs are observationally equivalent to this DAG? Explain.

Missing Data

Question 2.

(a) Consider the following observed data likelihood:

$$\mathcal{L}_{[D]}(\beta) = \prod_{i=1}^{n} \sum_{x_{2i}^{(1)} \text{ if } r_{2i}=0} \sum_{x_{4i}^{(1)} \text{ if } r_{4i}=0} p(x_{1i}, x_{3i}) p(r_{4i} \mid x_{1i}, x_{3i}) p(r_{2i} \mid x_{1i}) p(x_{2i}^{(1)} \mid x_{1i}) p(x_{4i}^{(1)} \mid x_{3i})$$
$$p(x_{2i} \mid r_{2i}, x_{2i}^{(1)}) p(x_{4i} \mid r_{4i}, x_{4i}^{(1)})$$

Draw the missing data graph for this model.

(b) Does this likelihood represent a missing at random (MAR) model or missing not at random (MNAR) model?

(c) Express $p(x_1, x_2^{(1)}, x_3, x_4^{(1)})$ as a function $p(x_1, x_2, x_3, x_4, r_2, r_4)$.

(d) Consider the following observed data likelihood:

$$\mathcal{L}_{[D]}(\beta) = \prod_{i=1}^{n} \sum_{x_{1i}^{(1)} \text{ if } r_{1i}=0} \sum_{x_{2i}^{(1)} \text{ if } r_{2i}=0} \sum_{x_{3i}^{(1)} \text{ if } r_{3i}=0} p(x_{1i}^{(1)}) p(x_{2i}^{(1)} \mid x_{1i}^{(1)}) p(x_{3i}^{(1)} \mid x_{2i}^{(1)}, x_{1i}^{(1)})$$
$$p(r_{1i} \mid x_{2i}^{(1)}, x_{3i}^{(1)}) p(r_{2i} \mid x_{1i}^{(1)}, r_{3i}) p(r_{3i} \mid x_{2i}^{(1)}, r_{1i})$$
$$p(x_{1i} \mid r_{1i}, x_{1i}^{(1)}) p(x_{2i} \mid r_{2i}, x_{2i}^{(1)}) p(x_{3i} \mid r_{3i}, x_{3i}^{(1)})$$

Draw the missing data graph for this model.

(e) Does this likelihood represent a missing at random (MAR) model or missing not at random (MNAR) model?

(f) Does the observed data likelihood have a unique global maximum? In other words, is $p(x_1^{(1)}, x_2^{(1)}, x_3^{(1)}, r_1, r_2, r_3)$ a function of the observed data distribution $p(x_1, x_2, x_3, r_1, r_2, r_3)$? Why? (This is a yes/no question with an explanation: if such a function exists, you don't have to give it)

PCA.

(a) Assume we performed PCA on a centered dataset $[D^*]$ of size n by k, and kept the first m eigenvalues of the sample covariance matrix $\hat{C}$. In other words, the new dataset $[\tilde{D}] = [D^*][V]_m$ is a matrix of size n by m, where $[V]_m$ is a k by m matrix containing the m eigenvectors corresponding to the largest eigenvalues of $\hat{C}$. Let i and j be different column indices in $[V]_m$. Show that the sample covariance between the corresponding columns of $[\tilde{D}]$ is 0 (meaning that the ith and jth features in $[\tilde{D}]$ are uncorrelated.

Hints: the sample covariance of any dataset $[X]$ is proportional to $[X]^T[X]$. You may use the fact that eigenvectors are all orthogonal to each other, meaning that if $\mathbf{v}_i$, $\mathbf{v}_j$ are eigenvectors, then $\mathbf{v}_i\mathbf{v}_j = 0$. Finally, note that given a matrix $[A]$, and its eigenvalue λ_i, and the corresponding eigenvector $\mathbf{v}_i$, $[A]\mathbf{v}_i = \lambda\mathbf{v}_i$.

(b) PCA is a dimension reduction method that aims to find a k-dimensional description of m-dimensional data (where k is hopefully much smaller than m). Kernel PCA aims to project m into a very high dimensional (possibly infinite dimensional) space using a kernel $K(x_i, x_j)$. Given that the goal is to reduce dimension of the original data $[D]$, what is the point of projecting to a high dimensional space prior to reducing dimension to k?