Homework 3
Due Friday March 29 3:00 pm
Submit a pdf file on Canvas

1. Get the iris data. (In R, use `data(iris)`.) There are 150 observations. The outcome is "Species" which has three values. The goal is to predict Species using the four covariates. Compare the following classifiers: (i) LDA, (ii) logistic regression, (iii) nearest neighbors. Note that you will need to figure out a way to deal with three classes when using logistic regression. Explain how you handled this. Summarize your results.

2. Use the iris data again but throw away the Species variable. Use k-means^{++} clustering and mean-shift clustering. Compare the clusterings to the true group defined by Species. Which method worked better?

3. Download the data from `http://www-bcf.usc.edu/~gareth/ISL/Ch10Ex11.csv`. This is a gene expression dataset. There are 40 tissue samples with measurements on 1,000 genes. The first 20 data points are from healthy people. The second 20 data points are from diseased people.

   (a) Use sparse logistic regression to classify the subject. (You may use the function glmnet in R if you like.) Explain how you chose λ. Summarize your findings.

   (b) Now use a Sparse Additive Model as described in class. Summarize your findings.

   (c) Now suppose we don't know which are healthy and which are diseased. Apply clustering to put the data into two groups. Applying k-means clustering may not work well because the dimension is so high. Instead, you will need to do some sort of dimension reduction or sparse clustering. One very simple method is Sparse Alternate Similarity (arXiv:1602.07277). But you may use any method you like. Describe what you chose to do and what the results are.

4. Let $X = (X_1, \ldots, X_d) \in \mathbb{R}^d$. Suppose that $X \sim N(\mu, \Sigma)$. Let $\Omega = \Sigma^{-1}$. Let $j \neq k$ be integers such that $1 \leq j < k \leq d$. Let $Z = (X_s : \ s \neq j, k)$.

   (a) Show that the distribution of $(X_j, X_k)|Z$ is $N(a, B)$ and find a and B explicitly.

   (b) Show that $X_j \amalg X_k|Z$ if and only if $\Omega_{jk} = 0$.

   (c) Now let $X_1, \ldots, X_n \sim N(\mu, \Sigma)$. Find the mle $\widehat{\Omega}$.

5. Let $X = (X_1, X_2, X_3, X_4, X_5)$ be a random vector distributed as $X \sim N(0, \Sigma)$ where

$$\Sigma^{-1} = \begin{pmatrix} 3 & 0 & 1 & 0 & 0 \\ 0 & 3 & 1 & 0 & 0 \\ 1 & 1 & 3 & 1 & 0 \\ 0 & 0 & 1 & 3 & 1 \\ 0 & 0 & 0 & 1 & 3 \end{pmatrix}.$$

(a) What is the graph for X, viewed as an undirected graphical model?

(b) List the maximal cliques of the graph.

(c) Which of the following independence statements are true?

i. $X_2 \perp\!\!\!\perp X_3 | X_1, X_2$

ii. $X_3 \perp\!\!\!\perp X_4 | X_5$

iii. $\{X_1, X_2\} \perp\!\!\!\perp X_3 | X_4, X_5$

iv. $X_1 \perp\!\!\!\perp X_5 | X_3$

(d) List the local Markov properties for this graphical model.

(e) Simulate 100 observations from this model. Construct a graph using hypothesis testing. Report your results. Include your code.

6. Let $X = (X_1, \ldots, X_4)$ where each variable is binary. Suppose the probability function is

$$\log p(x) = \psi_\emptyset + \psi_1(x_1) + \psi_{12}(x_1, x_2) + \psi_{13}(x_1, x_3) + \psi_{24}(x_2, x_4) + \psi_{34}(x_3, x_4).$$

(a) Draw the implied graph.

(b) Write down all the independence and conditional independence relations implies by the graph.

(c) Is the model graphical? Is the model hierarchical?

7. Let $X_1, \ldots, X_4$ be binary. Draw the independence graphs corresponding to the following log-linear models (where $\alpha \in \mathbb{R}$). Also, identify whether each is graphical and/or hierarchical (or neither).

(a) $\log p(x) = \alpha + 11x_1 + 2x_2 + 3x_3$

(b) $\log p(x) = \alpha + 2x_2 + 1.5x_3 + 17x_4 + 12x_2x_3 + 78x_2x_4 + 3x_3x_4 + 32x_2x_3x_4$

(c) $\log p(x) = \alpha + 9x_1 + 2x_2 + 1.5x_3 + 17x_4 + 12x_2x_3 + 3x_3x_4 + x_1x_4 + 2x_1x_2$

(d) $\log p(x) = \alpha + 115x_1x_2x_3x_4$.

8. Consider the log-linear model

$$\log p(x) = \beta_0 + x_1x_2 + x_2x_3 + x_3x_4.$$

Simulate $n = 1000$ random vectors from this distribution. (Show your code.) Fit the model

$$\log p(x) = \beta_0 + \sum_j \beta_j x_j + \sum_{k<\ell} \beta_{k\ell} x_k x_\ell$$

using maximum likelihood. Report your estimators. Use hypothesis testing to decide which parameters are non-zero. Compare the selected model to the true model.

9. Let $X_1, \ldots, X_n \in \mathbb{R}^d$. Let Σ be the $d \times d$ covariance matrix for X_i. The covariance graph G puts an edge between (j,k) if $\Sigma_{jk} \neq 0$. Here we will us the bootstrap to estimate the covariance graph.

   Let Σ have the following form: $\Sigma_{jj} = 1$, $\Sigma_{j,k} = a$ if $|j-k| = 1$ and $\Sigma_{j,k} = 0$ otherwise. Here, $a = 1/4$.

   Let $d = 100$ and $n = 50$. Generate n observations. Compute a 95 percent bootstrap confidence set for Σ using the bootstrap distribution

$$\mathbb{P}\Big(\max_{j,k} \sqrt{n}|\widehat{\Sigma}^*_{jk} - \widehat{\Sigma}_{jk}| \leq t \;\Big|\; X_1, \ldots, X_n\Big).$$

   This gives (uniform) confidence intervals for all the elements of Σ_{jk}. For each (j,k), put an edge if the confidence interval for Σ_{jk} excludes 0. Plot your graph. Try this for different values of a. Summarize your results.

10. Let $A \in \{0,1\}$ be a binary treatment variable and let $Y \in \mathbb{R}$ be the response variable. Let $(Y(0), Y(1))$ be the counterfactual variables where $Y = AY(1) + (1-A)Y(0)$. Assume that

$$Y = \alpha + \gamma A + \sum_{j=1}^{d} \beta_j X_j + \epsilon$$

    where $(X_1, \ldots, X_d)$ are confounding variables and $\mathbb{E}[\epsilon | X_1, \ldots, X_d] = 0$. Assume there are no unmeasured variables.

    (a) Let $\theta = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$. Show that $\theta = \gamma$.

    (b) Suppose now that we do not observe the confounding variables X_j. All we observe is $(A_1, Y_1), \ldots, (A_n, Y_n)$. Suppose, unaware of the confounding variables, we fit the linear model $Y = \alpha + \rho A + \delta$ where $\mathbb{E}[\delta] = 0$. Let $\widehat{\rho}$ be the least squares estimator. Show that $\widehat{\rho} \xrightarrow{P} \gamma + \Delta$ for some Δ. Find an explicit expression for Δ.

11. Consider a sequence of time ordered random variables

$$X_1, A_1, Y_1, X_2, A_2, Y_2, X_3, A_3, Y_3, \ldots, X_T, A_T, Y_T.$$

    Here, the X_j's are covariates, the A_j's are binary treatment variables and the Y_j's are the response of interest. Assume there are no unobserved confounding variables. The DAG for this model has all directed arrows from the past into the future. That is, the parents for each variables are all variables in its past, For example, the parent of A_1 is X_1. The parents of Y_1 are (X_1, A_1). The parents of X_2 are (X_1, A_1, Y_1) and so on. Let p denote the joint density of all these variables.

    (a) Find an explicit expression (in terms of p) for

$$E[Y_T | A_1 = a_1, \ldots, A_T = a_T].$$

    (b) Find an explicit expression (in terms of p) for

$$E[Y_T | \text{ set } (A_1 = a_1, A_2 = a_2, \ldots, A_T = a_T)].$$