Practice Exam

(1) Suppose that $(X_1, Y_1), \ldots, (X_n, Y_n)$ are from the following model:

$$Y_i = \beta X_i + \epsilon_i$$

where $\epsilon_i \sim N(0,1)$. (Note that there is no intercept.) Suppose you accidentally flip the signs of the Y_i's with probability 1/2. In more detail, let $U_1, \ldots, U_n$ be independent data such that $U_i \in \{-1, +1\}$ and $\mathbb{P}(U_i = -1) = \mathbb{P}(U_i = +1) = 1/2$. Define $W_i = U_i Y_i$. Let

$$\widehat{\beta} = \frac{\sum_{i=1}^n W_i X_i}{\sum_{i=1}^n X_i^2}.$$

Find the MSE of $\widehat{\beta}$.

---

(2) Let

$$\mathbb{X} = \begin{pmatrix} 1 & 2 \\ 1 & 2 \\ 1 & 2 \\ 1 & 2 \end{pmatrix}.$$

(a) Describe the column space $\mathcal{L}$ corresponding to $\mathbb{X}$.
(b) Let

$$Y = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}.$$

Find the projection of Y onto the column space $\mathcal{L}$. That is, find the point $\widehat{Y} \in \mathcal{L}$ closest to Y.

---

(3) Suppose we have data $(X_1, Y_1), \ldots, (X_n, Y_n)$ and

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_1 X_{ip} + \epsilon_i.$$

You may assume that the variables are standardized so that

$$\frac{1}{n}\sum_{i=1}^n X_{ij}^2 = 1, \qquad j = 1, \ldots, p.$$

Suppose we regress Y on the j^{th} covariate, ignoring all the other covariates:

$$\widehat{\beta}_j = \frac{\sum_{i=1}^n Y_i X_{ij}}{\sum_{i=1}^n X_{ij}^2}.$$

Find the mean and variance of $\widehat{\beta}_j$. When will $\widehat{\beta}_j$ be an unbiased estimate of β_j?

---

(4) Let

$$Y_i = r(X_i) + \epsilon_i$$

where $\epsilon_i \sim N(0,1)$. Let $\widehat{r}_n(x)$ denote the kernel regression estimator of $r(x)$. Suppose that $r(x) = c$, that is, r is constant. Find the mean and variance of $\widehat{r}_n(x)$.

---

(5) Suppose that $X|Y=0 \sim N(0,1)$ and $X|Y=1 \sim N(0,4)$. Thus,

$$f_0(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \qquad f_1(x) = \frac{1}{2\sqrt{2\pi}} e^{-x^2/8}.$$

Also assume that $\mathbb{P}(Y=0) = \mathbb{P}(Y=1) = 1/2$.

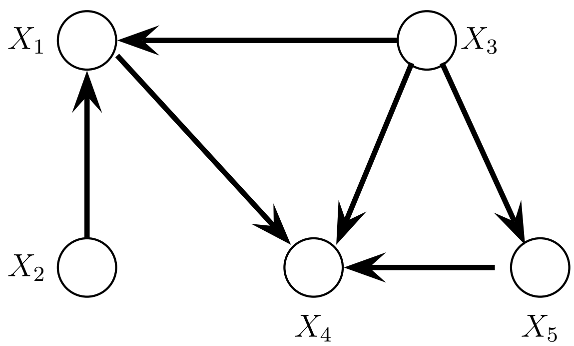
(a) Find the Bayes rule.

(b) Find an expression for the Bayes risk.

(c) Find the regression function $r(x) = \mathbb{P}(Y=1|X=x)$.

---

(6) Here is a DAG for 5 variables:

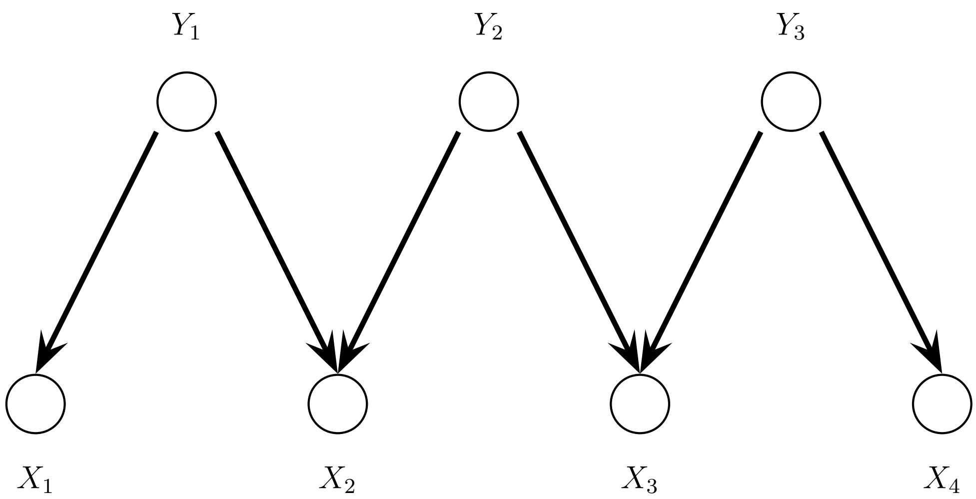


(a) Write down the factorization for the distribution $p(x_1, x_2, x_3, x_4, x_5)$.

(b) Based on the graph, fill in YES or NO in this table:

Independence statement	YES/NO
$X_1 \perp\!\!\!\perp X_2$	
$X_1 \perp\!\!\!\perp X_3$	
$X_1 \perp\!\!\!\perp X_5$	
$X_1 \perp\!\!\!\perp X_5 \mid X_3$	
$X_1 \perp\!\!\!\perp X_5 \mid X_4$	
$X_1 \perp\!\!\!\perp X_5 \mid X_3, X_4$	

(7) Here is a DAG for 7 variables:

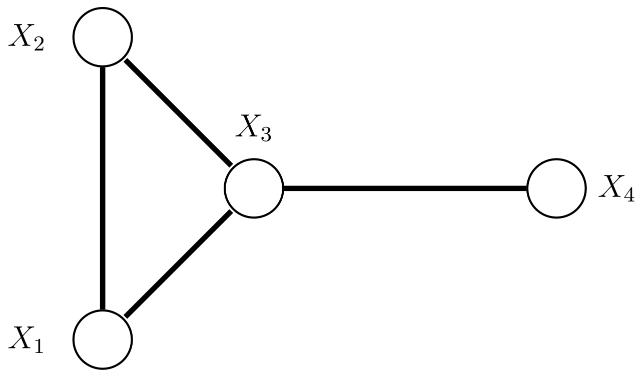


(a) Write down the factorization for the distribution $p(x_1, x_2, x_3, x_4, y_1, y_2, y_3)$.

(b) Based on the graph, fill in YES or NO in this table:

Independence statement	YES/NO
$X_1 \perp\!\!\!\perp X_2$	
$X_1 \perp\!\!\!\perp X_2 \mid Y_1$	
$X_1 \perp\!\!\!\perp X_3$	
$X_1 \perp\!\!\!\perp X_3 \mid X_2$	
$X_1 \perp\!\!\!\perp X_3 \mid Y_1$	
$X_1 \perp\!\!\!\perp X_4 \mid X_2$	
$(X_1, X_2) \perp\!\!\!\perp X_4 \mid X_3$	

(8) Here is an undirected graph for 4 variables:



(a) Identify the maximal cliques and write down a factorization for the distribution $p(x_1, x_2, x_3, x_4)$.
(b) Write down a graphical, loglinear model for this graph.
(c) Write down a nongraphical, hierarchical, loglinear model for this graph.
(d) Write down a nonhierarchical loglinear model for this graph.