UC Berkeley
Department of Electrical Engineering and Computer Science
Department of Statistics

EECS 281A / STAT 241A Statistical Learning Theory

Solutions to Problem Set 1

Fall 2011

Issued: Thurs, September 8, 2011 **Due:** Monday, September 19, 2011

Reading: For this problem set: Chapters 2–4.
Total: 100 points.

Problem 1.1
(15 pts) Larry and Consuela love to challenge each other to coin flipping contests. Suppose that Consuela flips $n+1$ coins, and Larry flips with n coins, and that all coins are fair. Letting E be the event that Consuela flips more heads than Larry, show that $\mathbb{P}[E] = \frac{1}{2}$.
Solution:

We define X_i as the outcome of the i^{th} throw by Consuela: $X_i = 1$ if Consuela flips heads and $X_i = 0$ if Consuela flips tails. Let $X = \sum_{i=1}^{n} X_i$ be the sum of Consuela's first n throws, and define Y_i and Y similarly for Larry. There are two cases when Consuela gets more heads than Larry: (1) Consuela gets more heads in the first n throws than Larry, or (2) Consuela gets the same number of heads in the first n throws as Larry, and gets a head in the last throw. Note that individual throws are independent from each other. Thus,

$$\mathbb{P}(E) = \mathbb{P}(X > Y) + \mathbb{P}(X = Y)\mathbb{P}(X_{i+1} = 1).$$

To compute the probabilities $\mathbb{P}(X > Y)$ and $\mathbb{P}(X = Y)$, note that X and Y are two i.i.d. Binomial$(n, \frac{1}{2})$ random variables. Thus,

$$\mathbb{P}(X > Y) = \mathbb{P}(X < Y).$$

Denoting $P(X = Y) = p$, we then have

$$\mathbb{P}(X > Y) = \frac{1-p}{2},$$

so

$$\mathbb{P}(E) = \frac{1-p}{2} + p \times \frac{1}{2} = \frac{1}{2}.$$

Problem 1.2
(15 pts) John's office is in a building that has 5 doors. Due to personal peculiarity, John refuses to use the same door twice in a row. In fact, he will choose the door to the left or the right of the last door he used, with probability p and $1-p$ respectively, and independently of what he has done in the past. For example, if he just chose door 5, there is a probability p that he will choose door 4, and a probability $1-p$ that he will choose door 1 next.

(a) Explain why the above process is a Markov chain.

Solution: Denote by X_t the index of the door John chooses at time t. Clearly,

$$P(X_t | X_1, X_2, \ldots, X_{t-1}) = P(X_t | X_{t-1}),$$

since the probability of the current choice only depends on the last door chosen. Hence, $\{X_t\}$ forms a Markov chain.

(b) Find the transition probability matrix.

Solution: The transition probability matrix A (defined by $A_{ij} = P(X_t = j | X_{t-1} = i)$) is as follows:

$$A = \begin{bmatrix} 0 & 1-p & 0 & 0 & p \\ p & 0 & 1-p & 0 & 0 \\ 0 & p & 0 & 1-p & 0 \\ 0 & 0 & p & 0 & 1-p \\ 1-p & 0 & 0 & p & 0 \end{bmatrix}.$$

(c) Find the steady state probabilities.

Solution: Denote the steady state probability as π. Then π must satisfy the equations $\mathbf{1}^\top \pi = 1$ and $\pi = \pi A$. Solving these equations gives

$$\pi = (0.2, 0.2, 0.2, 0.2, 0.2).$$

Problem 1.3

(15 pts) Let X_1, X_2, X_3, X_4 be random variables. For each of the following statements, either give a proof of its correctness, or a counterexample to show incorrectness.

(a) If $X_1 \perp X_4 \mid \{X_2, X_3\}$, then $X_2 \perp X_3 \mid \{X_1, X_4\}$.

Solution: This is false. Let X_1, X_2, X_4 be independent draws from $N(0,1)$, and let $X_3 = X_2$. Then $X_1 \perp X_4 \mid \{X_2, X_3\}$, but X_2 and X_3 remain dependent, even when conditioned on X_1 and X_4.

(b) If $X_1 \perp X_2 \mid X_4$ and $X_1 \perp X_3 \mid X_4$, then $X_1 \perp (X_2, X_3) \mid X_4$.

Solution: This is false. Let X_4 be independent of the other three random variables. Let X_2, X_3 be two independent Bernoulli($\frac{1}{2}$) random variables. Finally, let X_1 be the binary random variable that takes value 0 when $X_2 = X_3$ and takes value 1 otherwise. One can verify that $X_1 \perp X_2$ and $X_1 \perp X_2$, but X_1 is completely determined by observing the pair (X_2, X_3).

(c) If $X_1 \perp (X_2, X_3) \mid X_4$, then $X_1 \perp X_2 \mid X_4$.

Solution: This is true, since

$$p(x_2|x_1, x_4) = \sum_{x_3} p(x_2, x_3|x_1, x_4) = \sum_{x_3} p(x_1, x_2, x_3|x_4) = p(x_1, x_2|x_4),$$

where the second equality uses the assumption $X_1 \perp (X_2, X_3) \mid X_4$.

(d) Given three discrete random variables (X, Y, Z) and their PMF p, the following three statements are all equivalent (that is, $(i) \Leftrightarrow (ii) \Leftrightarrow (iii)$)

(i) $X \perp Y \mid Z$

(ii) $p(x, y, z) = h(x, z)k(y, z)$ for some functions h and k.

(iii) $p(x, y, z) = p(x, z)p(y, z)/p(z)$.

Solution: This is true, which we prove as follows:

$[(i) \Rightarrow (ii)]$: $X \perp Y \mid Z$ implies $p(x, y|z) = p(x|z)p(y|z)$. Hence,

$$p(x, y, z) = p(x, y|z)p(z) = p(x|z)p(y|z)p(z) = h(x, z)k(y, z),$$

for $h(x, z) = p(x|z)$ and $k(y, z) = p(y|z)p(z)$.

$[(ii) \Rightarrow (iii)]$: Since $p(x, y, z) = h(x, z)k(y, z)$, we have

$$\begin{aligned} p(x, z) &= h(x, z) \sum_y k(y, z) \\ p(y, z) &= k(y, z) \sum_x h(x, z) \\ p(z) &= \sum_x p(x, z) = \sum_x h(x, z) \sum_y k(y, z). \end{aligned}$$

Therefore,

$$\begin{aligned} \frac{p(x, z)p(y, z)}{p(z)} &= \frac{h(x, z)k(y, z) \sum_{y'} k(y', z) \sum_{x'} h(x', z)}{\sum_{x'} h(x', z) \sum_{y'} k(y', z)} \\ &= h(x, z)k(y, z) \\ &= p(x, y, z). \end{aligned}$$

$[(iii) \Rightarrow (i)]$: From the definition of conditional probability and assumption (iii),

$$p(x, y|z) = \frac{p(x, y, z)}{p(z)} = \frac{p(x, z)}{p(z)} \frac{p(y, z)}{p(z)} = p(x|z)p(y|z).$$

Hence, $X \perp Y \mid Z$.

Problem 1.4

(20 pts) *Directed graphical models:* Consider the graph in Figure 1(a).

(a) Is the ordering $\{1, 2, \ldots, 10\}$ topological? If yes, justify your answer; if not, find a topological ordering.

Solution: The given ordering is not topological. One possible topological ordering is

$$\{1, 2, 7, 8, 6, 4, 10, 3, 9, 5\}.$$

This ordering is topological because since every parent node comes before its children.

(b) Write down the standard factorization for the given graph.

Solution: The standard factorization for any directed graphical model can be written as $p(x) = \prod_{v \in V} p(x_v | x_{\pi_v})$, which yields

$$\begin{aligned} p(x) \quad = \quad & p(x_1)p(x_2)p(x_3|x_{10})p(x_4|x_2, x_6, x_7)p(x_5|x_9)p(x_6|x_1, x_2) \\ & p(x_7)p(x_8)p(x_9|x_3, x_7, x_8)p(x_{10}|x_2). \end{aligned}$$

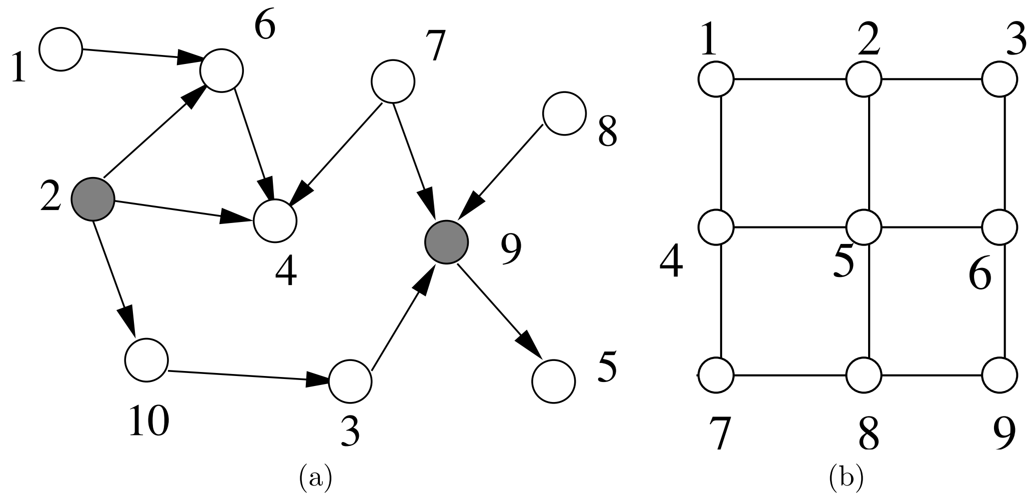


Figure 1: (a) A directed graph. (b) An undirected graphical model: a 3×3 grid or lattice graph. (Two-dimensional lattices frequently arise in spatial statistics.)

(c) For what pairs (i, j) does the statement $X_i \perp X_j$ hold? (Don't assume any conditioning in this part.)

Solution: The goal is to find all pairs (i, j) such that $X_i \perp X_j$. For node 1, an application of the Bayes Ball Algorithm shows that the ball can get to nodes 6 and 4 from node 1. For node 2, the ball can reach nodes 6, 4, 10, 3, 9, and 5. The ball starting at nodes 3 and 10 can reach the sames nodes as node 2. Node 4 can reach every node but node 8. Node 5 can reach every node but node 1. Node 6 can reach any node but nodes 7 and 8. Node 7 can reach node 9, 4, and 5. Node 8 can only reach nodes 9 and 5. Node 9 can't reach node 1. Finally, node 10 can't reach nodes 1, 7, and 8. So the pairs are $(1, 2)$, $(1, 3)$, $(1, 5)$, $(1, 7)$, $(1, 8)$, $(1, 9)$, $(1, 10)$, $(2, 7)$, $(2, 8)$, $(3, 7)$, $(3, 8)$, $(4, 8)$, $(6, 7)$, $(6, 8)$, $(7, 8)$, $(7, 10)$, and $(8, 10)$. In all there are 17 distinct pairs.

(d) Suppose that we condition on $\{X_2, X_9\}$, shown shaded in the graph. What is the largest set A for which the statement $X_1 \perp X_A \mid \{X_2, X_9\}$ holds. The Bayes ball algorithm could be helpful.

Solution: Conditioned on $\{X_2, X_9\}$, a ball starting at node 1 cannot reach nodes 3, 10, 7, 8 and 5. Hence, $A = \{3, 5, 7, 8, 10\}$. Note that nodes 2 and 9 are not elements of the set A because we are conditioning on them.

(e) What is the largest set B for which $X_8 \perp X_B \mid \{X_2, X_9\}$ holds?

Solution: Conditioned on $\{X_2, X_9\}$, a ball starting at node 8 cannot reach nodes 1, 5, and 6. Therefore, $B = \{1, 5, 6\}$.

(f) Suppose that I wanted to draw a sample from the marginal distribution $p(x_5) = \mathbb{P}[X_5 = x_5]$. (Don't assume that X_2 and X_9 are observed.) Describe an efficient algorithm to do so without actually computing the marginal. (Hint: Use the factorization from (b).)

```
sample(j,G)
   if (j has no parents in G)
      return x_j distributed according to p(x_j)
   else
      for each i parent of j
        parents[i]=sample(i,G)
        return x_j distributed according to p(x_j | parents)
```

Figure 2: Psuedo-code for generating a sample for a node in a given directed graphical model

Solution: In this problem, our goal is to somehow generate a sample of x_5 from the marginal distribution $p(x_5)$. One method to accomplish this is to apply eliminate and calculate the marginal distribution of node x_5. However, we are performing computations that are unnecessary in this case, since we do not need to explicitly calculate the marginal. Instead, we use the algorithm given in Figure 2 for sampling from any node j in a DAG G, given a topological ordering.

Hence, in order to generate a sample from X_5, we can first generate a sample (x_2, x_7, x_8) from X_2, X_7, and X_8. Then, using the factorization from part (b), we can generate a sample from the distribution $p(x_{10}|x_2)$, followed by a sample from the distribution $p(x_3|x_{10})$, by using the obtained samples x_2 and x_{10}. Next, we can generate a sample of x_9 from the distribution $p(x_9|x_7, x_8, x_2)$. Finally, we can obtain a sample for x_5 by sampling from the distribution $p(x_5|x_9)$. Note that generating a sample of x_5 in this way does not require sampling from x_1, x_6, or x_4, because $X_5 \perp X_1, X_4, X_6 \mid X_2, X_7, X_8$. In fact, we have generated a sample from the joint distribution

$$p(x_2, x_3, x_{10}, x_7, x_8, x_9, x_5) = p(x_2)p(x_7)p(x_8)p(x_{10}|x_2)p(x_3|x_{10})p(x_9|x_3, x_8, x_7)p(x_5|x_9).$$

Problem 1.5
(20 pts) *Graphs and independence relations:* For $i = 1, 2, 3$, let X_i be an indicator variable for the event that a coin toss comes up heads (which occurs with probability q). Supposing that that the X_i are independent, define $Z_4 = X_1 \oplus X_2$ and $Z_5 = X_2 \oplus X_3$ where $\oplus$ denotes addition in modulo two arithmetic.

(a) Compute the conditional distribution of (X_2, X_3) given $Z_5 = 0$; then, compute the conditional distribution of (X_2, X_3) given $Z_5 = 1$.

Solution: [4 pts] Note that

$$p(x_2, x_3|z_5) = \frac{p(x_2, x_3, z_5)}{p(z_5)} = \frac{p(x_2, x_3)}{p(z_5)}.$$

Furthermore,

$$P(Z_5 = 0) = P(X_2 = 0, X_3 = 0) + P(X_2 = 1, X_3 = 1),$$
$$P(Z_5 = 1) = P(X_2 = 0, X_3 = 1) + P(X_2 = 1, X_3 = 0).$$

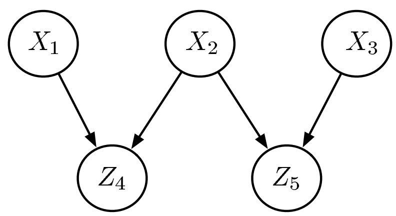
Using these facts, we easily obtain

$$\mathbb{P}(X_2 = 0, X_3 = 0 \mid Z_5 = 0) = \frac{(1-q)^2}{(1-q)^2 + q^2}$$
$$\mathbb{P}(X_2 = 1, X_3 = 1 \mid Z_5 = 0) = \frac{q^2}{(1-q)^2 + q^2}$$
$$\mathbb{P}(X_2 = 1, X_3 = 0 \mid Z_5 = 0) = 0$$
$$\mathbb{P}(X_2 = 0, X_3 = 1 \mid Z_5 = 0) = 0,$$

$$\mathbb{P}(X_2 = 0, X_3 = 0 \mid Z_5 = 1) = 0$$
$$\mathbb{P}(X_2 = 1, X_3 = 1 \mid Z_5 = 1) = 0$$
$$\mathbb{P}(X_2 = 1, X_3 = 0 \mid Z_5 = 1) = \frac{1}{2}$$
$$\mathbb{P}(X_2 = 0, X_3 = 1 \mid Z_5 = 1) = \frac{1}{2}.$$

(b) Draw a directed graphical model (the graph and conditional probability tables) for these five random variables. What independence relations does the graph imply?

Solution: [7 pts]



The conditional probability tables display conditional probabilities distributions for each node given its parents. The tables for this problem are listed below. Note that the tables for the X_i's are simply the marginal distributions, and that given their parents, the nodes Z_j are specified exactly.

- X_1, X_2, and X_3:

$X_i = 0$	$X_i = 1$
$1-q$	q

- Z_4, Z_5:

	$Z_4 = 0$	$Z_4 = 1$
$X_1 = 0, X_2 = 0$	1	0
$X_1 = 0, X_2 = 1$	0	1
$X_1 = 1, X_2 = 0$	0	1
$X_1 = 1, X_2 = 1$	1	0

	$Z_5 = 0$	$Z_5 = 1$
$X_2 = 0, X_3 = 0$	1	0
$X_2 = 0, X_3 = 1$	0	1
$X_2 = 1, X_3 = 0$	0	1
$X_2 = 1, X_3 = 1$	1	0

We list the conditional independence assertions implied by the graph below. Note that if $X_A \perp X_B \mid X_C$, then certainly $X_{A'} \perp X_{B'} \mid X_C$ for any $A' \subseteq A, B' \subseteq B$ (cf. Problem 1.3c). Hence, we will only list the largest sets X_A and X_B which are conditionally independent given X_C; all other conditional independence assertions follow trivially from this fact.

- $X_1 \perp (X_2, X_3, Z_5)$
- $X_1 \perp X_2 \mid (X_3, Z_5)$
- $X_1 \perp Z_5 \mid (X_2, X_3, Z_4)$
- $X_1 \perp (X_2, X_3) \mid Z_5$
- $X_1 \perp (X_2, Z_5) \mid X_3$
- $X_1 \perp (X_3, Z_5) \mid (X_2, Z_4)$
- $X_2 \perp (X_1, X_3)$
- $X_2 \perp X_3 \mid (X_1, Z_4)$
- $X_3 \perp (X_1, X_2, Z_4)$
- $X_3 \perp (X_1, X_2) \mid Z_4$
- $X_3 \perp (X_2, Z_4) \mid X_1$
- $X_3 \perp (X_1, Z_4) \mid (X_2, Z_5)$
- $X_3 \perp Z_4 \mid (X_1, X_2, Z_5)$
- $Z_4 \perp (X_3, Z_5) \mid (X_1, X_2)$
- $Z_4 \perp Z_5 \mid (X_1, X_2, X_3)$
- $Z_5 \perp (X_1, Z_4) \mid (X_2, X_3)$
- $(X_1, Z_4) \perp (X_3, Z_5) \mid X_2$

(c) Draw an undirected graphical model (the graph and compatibility functions) for these five variables. What independence relations does it imply?

Solution: [6 pts] The graphical model is the moralized form of the graph in part (b).

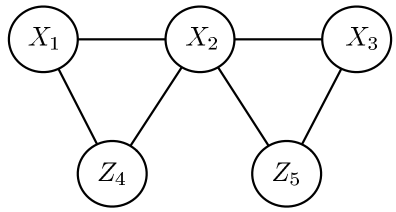
The maximal cliques in this graphical model can be parametrized in two ways, either as

$$\Psi_{\{X_1,Z_4,X_2\}}(x_1, z_4, x_2) = p(x_1)p(x_2)p(z_4|x_1, x_2) \tag{1}$$
$$\Psi_{\{X_2,Z_5,X_3\}}(x_2, z_5, x_3) = p(x_3)p(z_5|x_2, x_3), \tag{2}$$

or

$$\Psi_{\{X_1,Z_4,X_2\}}(x_1, z_4, x_2) = p(x_1)p(z_4|x_1, x_2) \tag{3}$$
$$\Psi_{\{X_2,Z_5,X_3\}}(x_2, z_5, x_3) = p(x_3)p(x_2)p(z_5|x_2, x_3). \tag{4}$$

Th conditional independence assertions implied by the graph are



- $(X_1, Z_4) \perp (X_3, Z_5) \mid X_2$
- $Z_4 \perp (X_3, Z_5) \mid (X_1, X_2)$
- $(X_1, Z_4) \perp Z_5 \mid (X_2, X_3)$
- $X_1 \perp (X_3, Z_5) \mid (Z_4, X_2)$
- $(X_1, Z_4) \perp X_3 \mid (X_2, Z_5)$
- $Z_4 \perp Z_5 \mid (X_1, X_2, X_3)$
- $Z_4 \perp X_3 \mid (X_1, X_2, Z_5)$
- $X_1 \perp Z_5 \mid (Z_4, X_2, X_3)$
- $X_1 \perp X_3 \mid (Z_4, X_2, Z_5)$.

As before, we have omitted the independencies that may be derived directly from the decomposition rule of Problem 1.3c.

(d) Under what conditions on q do we have $Z_5 \perp X_3$ and $Z_4 \perp X_1$? Are either of these marginal independence assertions implied by the graphs in (b) or (c)?

Solution: [3 pts] First note that when $q \in \{0, 1\}$, X_1, X_2, and X_3 are all constant random variables, which implies that Z_4 and Z_5 are also constant, so $Z_5 \perp X_3$ and $Z_4 \perp X_1$.

Next, consider $0 < q < 1$. For independence to hold, we need

$$P(Z_5 = 1) = P(Z_5 = 1 \mid X_3 = 0) = P(Z_5 = 1 \mid X_3 = 1).$$

Note that

$$\begin{aligned}
P(Z_5 = 1 \mid X_3 = 0) &= P(X_2 = 1) = q \\
P(Z_5 = 1 \mid X_3 = 1) &= P(X_2 = 0) = 1 - q \\
P(Z_5 = 1) &= P(Z_5 = 1 \mid X_3 = 0)P(X_3 = 0) \\
&\quad + P(Z_5 = 1 \mid X_3 = 1)P(X_3 = 1) = 2q(1 - q).
\end{aligned}$$

Thus, we have independence for $0 < q < 1$ if and only if $q = \frac{1}{2}$. These marginal independence assertions are not implied by the graphs.

Problem 1.6
(15 pts) *Car problems:* Alice takes her her car to a dishonest mechanic, who claims that it requires \$2500 dollars in repairs. Doubting this diagnosis, Alice decides to verify it with a graphical model.

Suppose that the car can have three possible problems: brake trouble (Br), muffler trouble (Mu), and low oil (Oi), all which are a priori marginally independent. The diagnosis is performed by testing for four possible symptoms: squealing noises (Sq), smoke (Sm), shaking (Sh), and engine light (Li). The conditional probabilities of these symptoms are related to the underlying causes as follows. Squealing depends only on brake problems, whereas smoke depends on brake problems and low oil. Shaking is related to brake problems and muffler problems, whereas the engine light depends only on the oil level.

(a) Draw a directed graphical model (graph only) for this problem.

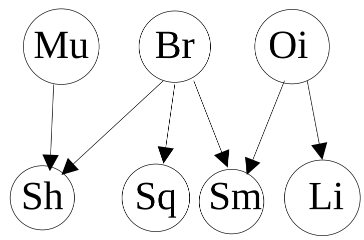


Figure 3: Model for Problem 1.8a.

Solution: [4 pts] See Figure 3.

(b) For which car problems do we gain information by observing the engine light? Suppose that we see the car belching oily clouds of smoke. What does seeing the engine light tell us now?

Solution: [4 pts] Observing the engine light by itself gives us information about possible low oil. Observing oily clouds of smoke tells us that there are possible problems with brakes and/or low oil. In this case, observing the engine light decreases the likelihood of a problem with brakes and increases the likelihood of low oil.

(c) Suppose that, in addition to smoking, the car also shakes violently. What does the flickering engine light tell us about brake problems or muffler problems?

Solution: [4 pts] Shaking and smoking indicates that all three problems are possible. Observing the engine light then decreases the likelihood of a brake problem (since smoking is now explained away) and increases the likelihood of a muffler problem (to explain shaking). Note, however, that unless we know the probability distribution associated with the model, we do not know whether a muffler or brake problem is a likelier explanation for shaking.

(d) Now suppose that Alice knows that the car has a brake problem. How useful is it to measure the smoke level in assessing whether it also has a muffler problem?

Solution: [3 pts] This is not at all useful, since conditional on a brake problem, the smoke level and muffler problems are independent (use the Bayes Ball Algorithm to see this).