1 Bayesian network inference

1.1 Variable elimination [12 points]

Consider the Bayesian network shown in Figure 1. Suppose we want to compute the marginal for node H, i.e, Probability(H=h). We will examine the effect of order of variable elimination on the amount of computation required. Assume each variable can take only two values.

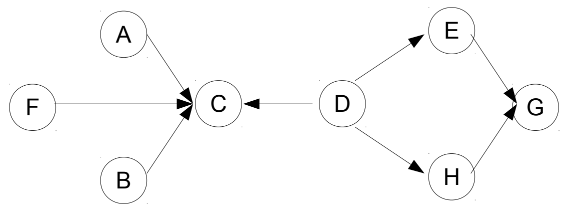


Figure 1: Bayesian network for variable elimination

- [**5 points**] Consider the elimination order: A,B,F,C,D,E,G. Write down the factors that you will encounter during this elimination. How many additions does it require?

  Answer:

Variable to be eliminated	Factors encountered	Number of additions required
A	$m_a(b, c, f)$	2^3=8
B	$m_b(c, f)$	2^2=4
F	$m_f(c)$	2
C	$m_c(d)$	2
D	$m_d(e, h)$	2^2=4
E	$m_e(g, h)$	2^2=4
G	$m_g(h)$	2

  Total number of additions: $8 + 4 + 2 + 2 + 4 + 4 + 2 = 26$.

- [**5 points**] Consider the elimination order: C,D,F,A,B,E,G. Write down the factors that you will encounter during this elimination. How many additions does it require?

  Answer:

Variable to be eliminated	Factors encountered	Number of additions required
C	$m_c(a, b, d, f)$	2^4=16
D	$m_d(a, b, f, e, h)$	2^5=32
F	$m_f(a, b, e, h)$	2^4=16
A	$m_a(b, e, h)$	2^3=8
B	$m_b(e, h)$	2^2=4
E	$m_e(g, h)$	2^2=4
G	$m_g(h)$	2

  Total number of additions: $16 + 32 + 16 + 8 + 4 + 4 + 2 = 82$.

- [**2 points**] Which elimination order is better in terms of computation and storage?

The first one.

1.2 Graph construction [9 points]

1. Suppose there is Bayesian network with n vertices $X_1, \cdots, X_n$. We want to compute $P(X_2 = x)$. Your task is to draw the edges in the Bayesian network so that the graph has the following two properties:

   Answer:
   The graph should be a star where X_1 is the center pointing to other variables.

   - [**2 points**] There exists an elimination order such that computing $P(X_2 = x)$ takes time linear in n.

     Answer: Elimination order: X_3, X_4 ... X_n, X_1.
   - [**2 points**] There exists an elimination order such that computing $P(X_2 = x)$ takes time exponential in n.

   Answer: Elimination order: X_1, X_3 ... X_n.

   Draw the graph ([**3 points**]) and write the two elimination orders we desire. Note that your set of edges should be the same for both cases.

2. [**2 points**] State true or false with brief explanation: If we have learnt a Naive Bayes classifier with n features and a class label, all of which are boolean, then for computing $P(X_1 = 0)$ there exists an order of elimination of the other variables that takes exponential time in terms of n.

   Answer: True. If you first eliminate the class variable C, then the computation is exponential.

1.3 Learning Bayes Nets [4 points]

Suppose you want to learn a Bayes net over two binary variables X_1 and X_2. You have N training pairs of X_1 and X_2, given as $\{(x_1^{(1)}, x_2^{(1)}), \cdots, (x_1^{(N)}, x_2^{(N)})\}$. For any Bayes net you learn its parameters using maximum likelihood estimation. Let A denote the BN with no edges, and B denote the BN with an edge from X_1 to X_2.

- [**2 points**] Describe a case when choosing B to model the data is better than choosing A.

  Answer: When X_1 and X_2 are not independent.

- [**2 points**] Describe a case when choosing A to model the data is better than choosing B. (Hint: Your choice may be determined by factors other than how well the BN fits the data).

  Answer: When X_1 and X_2 are actually independent, both Bayes nets would be able to learn the distribution. Then, learning A is easier than learning B.

2 Undirected Graphical Models [25pt, Nan Li]

A Markov Random Field is an undirected graphical model. Unlike Bayesian Networks, undirected edges in the graph simply give correlations between variables. Figure 2 is a Markov Random Field. Assume that all the variables are boolean. Each edge in the graph, $\forall i, j \in \{1, 2, 3, 4\}, i \neq j, e_{ij} = \langle x_i, x_j \rangle$, corresponds to a potential $\Psi_{e_{ij}}(x_i, x_j)$. We will explore the probability distribution encoded in this graph in this problem.

2.1 Representation [3 pt]

(a) Please write down the formula to calculate the joint probability of all variables defined by the above Markov Random Field in terms of the given potentials.
Answer:

$$P(X_1, X_2, X_3, X_4) = \frac{1}{Z}\Psi_{e_{12}}(X_1, X_2)\Psi_{e_{23}}(X_2, X_3)\Psi_{e_{34}}(X_3, X_4)\Psi_{e_{14}}(X_1, X_4) \tag{1}$$

where $Z = \sum_{X_1,X_2,X_3,X_4} \Psi_{e_{12}}(X_1, X_2)\Psi_{e_{23}}(X_2, X_3)\Psi_{e_{34}}(X_3, X_4)\Psi_{e_{14}}(X_1, X_4)$

(b) How many parameters do we need to store the potentials in the graph?
Answer: 16

2.2 Independence [?? pt]

(a) What is the Markov blanket of variable x_2?
Answer: $\{X_1, X_3\}$

(b) Is $x_1 \perp x_3 | x_2, x_4$? Please answer yes or no, and briefly explain why.
Answer: Yes. $\{X_2, X_4\}$ is the Markov blanket of X_1

2.3 Hammersley-Clifford Theorem [?? pt]

Now let's consider a probability distribution P over x_1, x_2, x_3, x_4, which gives probability 1/8 to each of the following configurations: $(0,0,0,0)$, $(1,0,0,0)$, $(1,1,0,0)$, $(1,1,1,0)$, $(0,0,0,1)$, $(0,0,1,1)$, $(0,1,1,1)$, $(1,1,1,1)$. All other configurations are given probability zero.

(a) Is $x_1 \perp x_3 | x_2, x_4$ true in the above distribution? Please answer yes or no, and briefly explain why.
Answer: Yes. We want to show $P(X_1, X_3|X_2, X_4) = P(X_1|X_2, X_4)P(X_3|X_2, X_4)$:
Similarly, you can compute $P(X_1, X_3|X_2, X_4)$ to show $P(X_1, X_3|X_2, X_4) = P(X_1|X_2, X_4)P(X_3|X_2, X_4)$.

(b) The MRF shown in Figure 2 is actually an I-map for P. Based on the Hammersley-Clifford Theorem taught in class, does this mean that the distribution P factorizes according to the

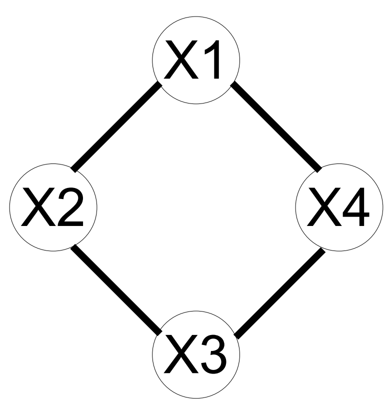


Figure 2: A Markov Random Field

(X_2, X_4)	$X_1 = 0$	$X_1 = 1$	$X_3 = 0$	$X_3 = 1$
$(0, 0)$	0.5	0.5	1	0
$(0, 1)$	1	0	0.5	0.5
$(1, 0)$	0	1	0.5	0.5
$(1, 1)$	0.5	0.5	0	1

given Markov Random Field, i.e. can you find a set of potential values for $\Psi_{e_{ij}(x_i,x_j)}$ so that it defines the distribution P? Please answer yes or no. If yes, please show the potentials. If no, please prove it.
Answer: No, the Hammersley-Clifford Theorem only holds for **positive** distribution. To prove this, let's define

$$\begin{aligned}\Psi_{e_{12}}(X_1 = 0, X_2 = 0) &= A_{00}\\ \Psi_{e_{12}}(X_1 = 0, X_2 = 1) &= A_{01}\\ \Psi_{e_{12}}(X_1 = 1, X_2 = 0) &= A_{10}\\ \Psi_{e_{12}}(X_1 = 1, X_2 = 1) &= A_{11}\end{aligned}$$

Similarly, define B for $\Psi_{e_{23}}(X_2, X_3)$, C for $\Psi_{e_{34}}(X_3, X_4)$ and D for $\Psi_{e_{14}}(X_1, X_4)$. Since we know $P(0, 0, 1, 0) = 0$, according to the factorization, we would expect $A_{00}B_{01}C_{10}D_{00} = 0$, which means at least one of $\{A_{00}, B_{01}, C_{10}, D_{00}\}$ will have to be 0. However, if either one of them is 0, this will contradict the distribution P we have. For example, w.l.o.g., let's assume $A_{00} = 0$. This will lead to $P(0, 0, 0, 0) = 0$ which contradicts the fact that $P(0, 0, 0, 0) = \frac{1}{8}$.

2.4 Partition Function [8 pt]

To define a probability distribution in an MRF, we need a need a normalization factor Z known as the partition function. Can we simply renormalize the values in each potential function so that they sum up to one, and then get rid of the normalization factor Z?

(a) If we scale only one potential function by a positive constant, how does it affect the distribution defined by the Markov network? Please prove your answer.
Answer: If we scale one potential function by a positive constant, and calculate the normalization factor Z with the updated potential, since both Z and the potential function are scaled by the same constant, the distribution will not change.

(b) If we scale all factors to sum to 1 locally, does that imply $Z = 1$? If yes, please prove it. If no, please give a counter example.
Answer: No. Use the previous model,

$$\begin{aligned}Z &= \frac{\sum_{X_1,X_2} \Psi_{e_{12}}(X_1, X_2) \sum_{X_2,X_3} \Psi_{e_{23}}(X_2, X_3) \sum_{X_3,X_4} \Psi_{e_{34}}(X_3, X_4) \sum_{X_1,X_4} \Psi_{e_{14}}(X_1, X_4)}{\sum_{X_1,X_2,X_3,X_4} \Psi_{e_{12}}(X_1, X_2)\Psi_{e_{23}}(X_2, X_3)\Psi_{e_{34}}(X_3, X_4)\Psi_{e_{14}}(X_1, X_4)}\\ &\neq 1\end{aligned}$$

3 SVMs [Qirong Ho, 25 points]

3.1 Feature Mappings

Suppose you are given 6 one-dimensional points: 3 with negative labels $x_1 = -1$, $x_2 = 0$, $x_3 = 1$ and 3 with positive labels $x_4 = -2$, $x_5 = 2$, $x_6 = 3$.

1. [**1 point**] Draw the 6 points on the one-dimensional line, using circles to represent the positive labels and squares to represent the negative labels.

   Answer:

   Refer to Figure 3

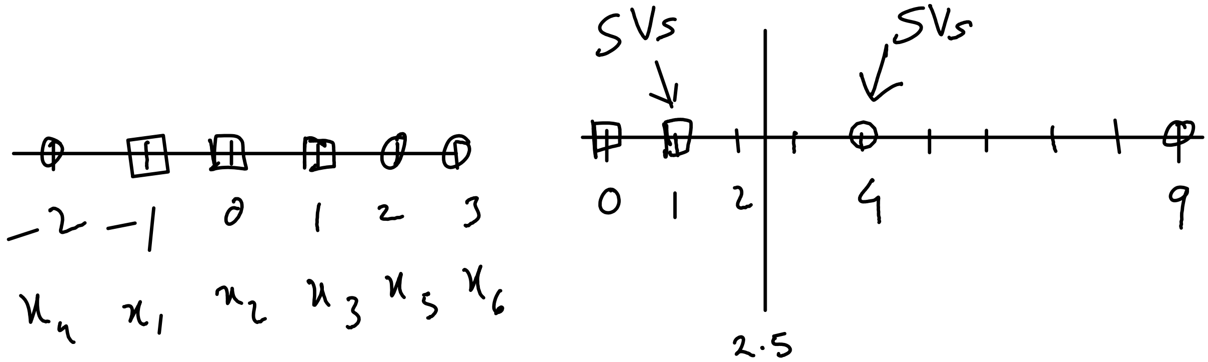


Figure 3: 1-d SVM plots

2. [**2 points**] The 6 points are not linearly separable. Write down a feature transformation $f(x)$ such that the transformed points $f(x_1), f(x_2), \ldots, f(x_6)$ are linearly separable.

   Answer:

   $f(x) = x^2$ will make the points linearly separable.

3. [**2 points**] Draw your 6 *transformed* points from part 2 on the one-dimensional line. Then, draw the decision boundary given by the hard-margin linear SVM, and indicate which points are support vectors.

   Answer:

   Refer to Figure 3

4. [**2 points**] Your decision boundary from part 3 has the form $w_0 + w_1 f(x)$. Give the values of w_0 and w_1.

   Answer:

   $w_0 = -5/3, w_1 = 2/3$

5. [**2 points**] Now suppose we transform the 6 points to the feature space $(x, f(x))$, where $f(x)$ is your feature transformation from part 2. In other words, you now have 6 two-

dimensional points $(x_1, f(x_1)), (x_2, f(x_2)), \ldots, (x_6, f(x_6))$. Draw these 6 points on the two-dimensional plane, along with the decision boundary given by the hard-margin linear SVM. Finally, indicate which points are support vectors.

Answer:

Refer to Figure 4

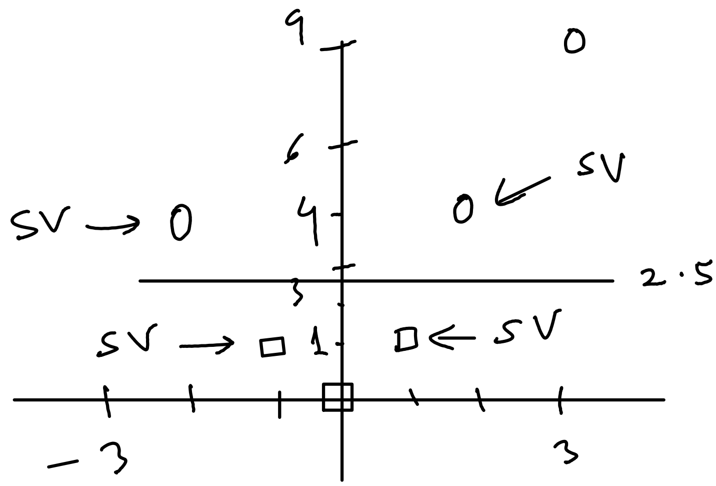


Figure 4: 2-d SVM plot

6. **[2 points]** Your decision boundary from part 5 has the form $w_0 + w_1 x + x_2 f(x)$. Give the values of w_0, w_1, w_2.

Answer:

$w_0 = -5/3, w_1 = 0, w_2 = 2/3$

7. **[2 points]** The feature mapping $x \longmapsto (x, f(x))$ in parts 5 and 6 is associated with a kernel $K(x, x')$ where x, x' are points in the original one-dimensional feature space. Write down this kernel.

Answer:

$$\begin{aligned} K\left(x, x'\right) &= \langle \phi(x), \phi(x') \rangle \\ &= \langle (x, x^2), (x', x'^2)) \rangle \\ &= x x' + x^2 x'^2 \end{aligned}$$

8. **[3 points]** What is the VC dimension of the hard-margin linear SVM in the feature space $(x, f(x))$? That is to say, what is the largest number of points in $(x, f(x))$ that can be shattered by a linear classifier?

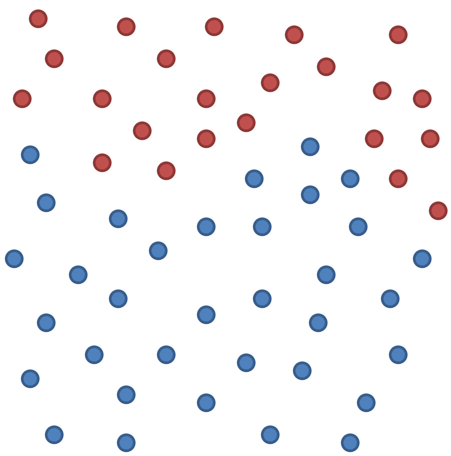
   Answer:

   Since the SVM is a linear hyperplane in the 2-d feature space, its VC dimension is 3.

3.2 Kernels

For each of the figures below, give a kernel that allows the hard-margin linear SVM to classify the red and blue points perfectly. No need to write down the mathematical form of the kernel, just state its type.

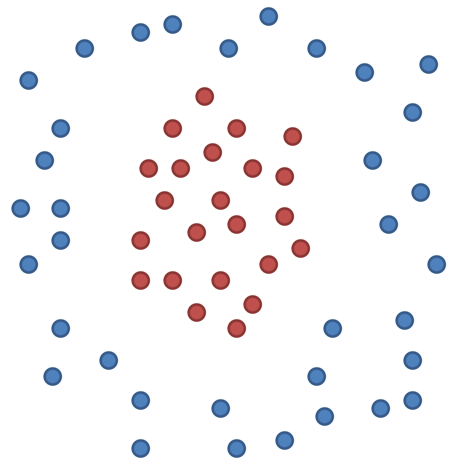
1. **[3 points]**

   Answer:

   A polynomial kernel .

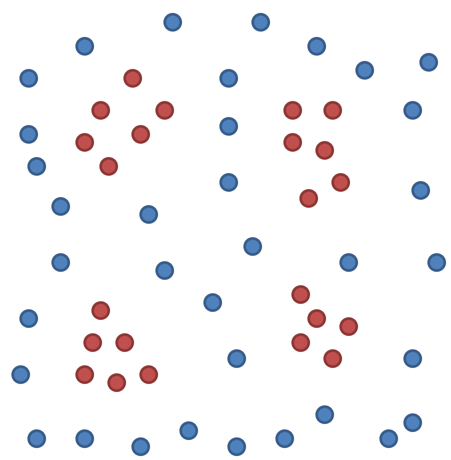
2. **[3 points]**



Answer:

A quadratic kernel centered at the center of mass of the red points.

3. **[3 points]**



Answer:

A gaussian RBF kernel, since a simple polynomial kernel will not be able to separate them.