Machine Learning 10-701
http://alex.smola.org/teaching/cmu2013-10-701x/
Carnegie Mellon University

Homework 4 Solutions

---

1 Graphical Models [Jing; 25 pts]

1.1 Directed Graphical Models (Bayesian Networks)

1. Independence Relations

   (a) C $\perp$ E | B
   Yes, we have serial path where B is given.

   (b) A $\perp$ E
   Yes, We have converging path where B is not given.

   (c) A $\perp$ C | E
   We have a serial path where B is not given.

   (d) D $\perp$ F
   We have two paths from D to F, DCF which is not blocked, because C is not given for a diverging path, and DGF, which is blocked because it is a diverging path where G and H are not given. Since only one of the paths is blocked, not both, D and F are not independent.

   (e) F $\perp$ H | G,D
   Yes, here both paths FGH (serial path) is blocked by G, and FCDG (diverging + serial) is blocked by D.

2. The joint distribution is written as follows.

$$P(X_1 \dots X_n) = \prod_{i=1}^{n} P(X_i|\text{Pa}_{X_i})$$
$$P(A,B,C,D,E,F,G,H) = P(A)P(E)P(B|A,E)P(C|B)P(F|C)P(D|C)P(G|D,F)P(H|G)$$

3. 3. a) Moralized graph below:

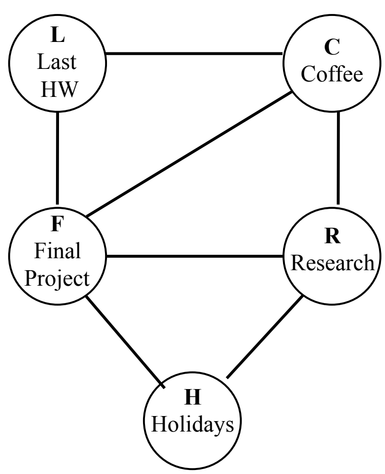

Figure 1: Moralized Bayesian Network

3. b) First eliminate F creating a clique of size 4, then any order of variables can follow.

3. c) L, C, F, R, H

3. d)

$$P(H|l,c) = \sum_{F,R,L,C} P(H|F,R)P(F|C,L)P(R|C)\delta(C=c)\delta(L=l)$$
$$P(H|l,c) = \sum_{F,R} P(H|F,R) \sum_{L,C} P(F|C,L)P(R|C)\delta(C=c)\delta(L=l)$$
$$P(H|l,c) = \sum_{F,R} P(H|F,R)P(F|C=c,L=l)P(R|C=c)$$

Reading off the conditional probability tables and computing, we get the following.

F	R	P(H = T | L = T, C = T)	P(H = F | L = T, C = T)
T	T	$0.9 \cdot 0.6 \cdot 0.9 = 0.486$	$0.9 \cdot 0.6 \cdot 0.1 = 0.054$
T	F	$0.9 \cdot 0.4 \cdot 0.7 = 0.252$	$0.9 \cdot 0.4 \cdot 0.3 = 0.108$
F	T	$0.1 \cdot 0.6 \cdot 0.6 = 0.036$	$0.1 \cdot 0.6 \cdot 0.4 = 0.024$
F	F	$0.1 \cdot 0.4 \cdot 0.2 = 0.008$	$0.1 \cdot 0.4 \cdot 0.8 = 0.032$
		Total = 0.782	Total = 0.218

Thus, $P(H|F=T,C=T) =< 0.7820, 218 >$.

1.2 Undirected Graphical Models (Markov Networks)

1. $P(x_1, x_2, x_3, x_4, x_5) = \psi_{13}(x_1, x_3)\ \psi_{23}(x_2, x_3)\ \psi_{34}(x_3, x_4)\ \phi_4(x_4, x_5)$

2. First, compute the messages.

$$\begin{aligned} m_{13}(x_3) &= \sum_{x_1} \psi_{13}(x_1, x_3) \\ &= \sum_{x_1} \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 1.5 \\ 1.5 \end{bmatrix} = k \begin{bmatrix} 1 \\ 1 \end{bmatrix} \end{aligned}$$

$$\begin{aligned} m_{23}(x_3) &= \sum_{x_1} \psi_{23}(x_2, x_3) \\ &= \sum_{x_2} \begin{bmatrix} 0.5 & 1 \\ 1 & 0.5 \end{bmatrix} \\ &= \begin{bmatrix} 1.5 \\ 1.5 \end{bmatrix} = k \begin{bmatrix} 1 \\ 1 \end{bmatrix} \end{aligned}$$

$$\begin{aligned} m_{54}(x_4) &= \sum_{x_5} \phi_4(x_5, x_4) \\ &= \sum_{x_5} \delta(x_5, 1) \begin{bmatrix} 0.1 & 1 \\ 1 & 0.1 \end{bmatrix} \\ &= \begin{bmatrix} 1 \\ 0.1 \end{bmatrix} \end{aligned}$$

$$\begin{aligned} m_{34}(x_4) &= \sum_{x_3} \psi_{34}(x_3, x_4) m_{13}(x_3) m_{23}(x_3) \\ &= \sum_{x_3} \begin{bmatrix} 0.2 & 1 \\ 1 & 0.2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} .* \begin{bmatrix} 1 \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} 1.2 \\ 1.2 \end{bmatrix} = k \begin{bmatrix} 1 \\ 1 \end{bmatrix} \end{aligned}$$

$$\begin{aligned} m_{43}(x_3) &= \sum_{x_4} \psi_{43}(x_4, x_3) m_{54}(x_4) \\ &= \sum_{x_3} \begin{bmatrix} 0.2 & 1 \\ 1 & 0.2 \end{bmatrix} \begin{bmatrix} 1 \\ 0.1 \end{bmatrix} \\ &= \begin{bmatrix} 0.3 \\ 1.02 \end{bmatrix} \end{aligned}$$

$$\begin{aligned} m_{31}(x_1) &= \sum_{x_3} \psi_{31}(x_3, x_1) m_{43}(x_3) m_{23}(x_3) \\ &= \sum_{x_3} \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \begin{bmatrix} 0.3 \\ 1.02 \end{bmatrix} .* \begin{bmatrix} 1 \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} 0.81 \\ 1.17 \end{bmatrix} \end{aligned}$$

$$\begin{aligned} m_{32}(x_2) &= \sum_{x_3} \psi_{32}(x_3, x_2) m_{43}(x_3) m_{13}(x_3) \\ &= \sum_{x_3} \begin{bmatrix} 0.5 & 1 \\ 1 & 0.5 \end{bmatrix} \begin{bmatrix} 0.3 \\ 1.02 \end{bmatrix} .* \begin{bmatrix} 1 \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} 1.17 \\ 0.81 \end{bmatrix} \end{aligned}$$

Now, computing the marginals.

$$P(x_1) = km_{31}(x_1) = \frac{1}{1.98}\begin{bmatrix} 0.81 \\ 1.17 \end{bmatrix} = \begin{bmatrix} 0.409 \\ 0.591 \end{bmatrix}$$

$$P(x_2) = km_{32}(x_1) = \frac{1}{1.98}\begin{bmatrix} 1.17 \\ 0.81 \end{bmatrix} = \begin{bmatrix} 0.591 \\ 0.409 \end{bmatrix}$$

$$\begin{aligned} P(x_3) &= km_{13}(x_3)m_{23}(x_3)m_{43}(x_3) \\ &= k\begin{bmatrix} 1 \\ 1 \end{bmatrix} .* \begin{bmatrix} 1 \\ 1 \end{bmatrix} .* \begin{bmatrix} 0.3 \\ 1.02 \end{bmatrix} \\ &= \frac{1}{1.32}\begin{bmatrix} 0.3 \\ 1.02 \end{bmatrix} = \begin{bmatrix} 0.227 \\ 0.773 \end{bmatrix} \end{aligned}$$

$$\begin{aligned} P(x_4) &= km_{54}(x_4)m_{34}(x_4) \\ &= k\begin{bmatrix} 1 \\ 0.1 \end{bmatrix} .* \begin{bmatrix} 1 \\ 1 \end{bmatrix} \\ &= \frac{1}{1.1}\begin{bmatrix} 1 \\ 0.1 \end{bmatrix} \\ &= \begin{bmatrix} 0.909 \\ 0.091 \end{bmatrix} \end{aligned}$$

2 Bootstrap [Ahmed]

2.1 Two modes, one variance

1. Let $Y \in \{0, 1\}$ indicate whether X is generated from the first or second mode. Then

$$Y \sim Bernoulli(1-p)$$

Using the law of total variance

$$\begin{aligned} &V[X] = E[V[X|Y]] + V[E[X|Y]] \\ &E[V[X|Y]] = pV[X|Y=0] + (1-p)V[X|Y=1] = p\sigma_1^2 + (1-p)\sigma_2^2 = \sigma^2 \\ &V[E[X|Y]] = V[\mu_1 Y + \mu_2(1-Y)] = V[\mu_2 + (\mu_1 - \mu_2)Y] = 0 + (\mu_1-\mu_2)^2 p(1-p) \end{aligned}$$

Then

$$V[\bar{X}] = \frac{V[X]}{n} = \frac{\sigma^2 + p(1-p)(\mu_1-\mu_2)^2}{n}$$

2. Coding Question

3. See figure 2

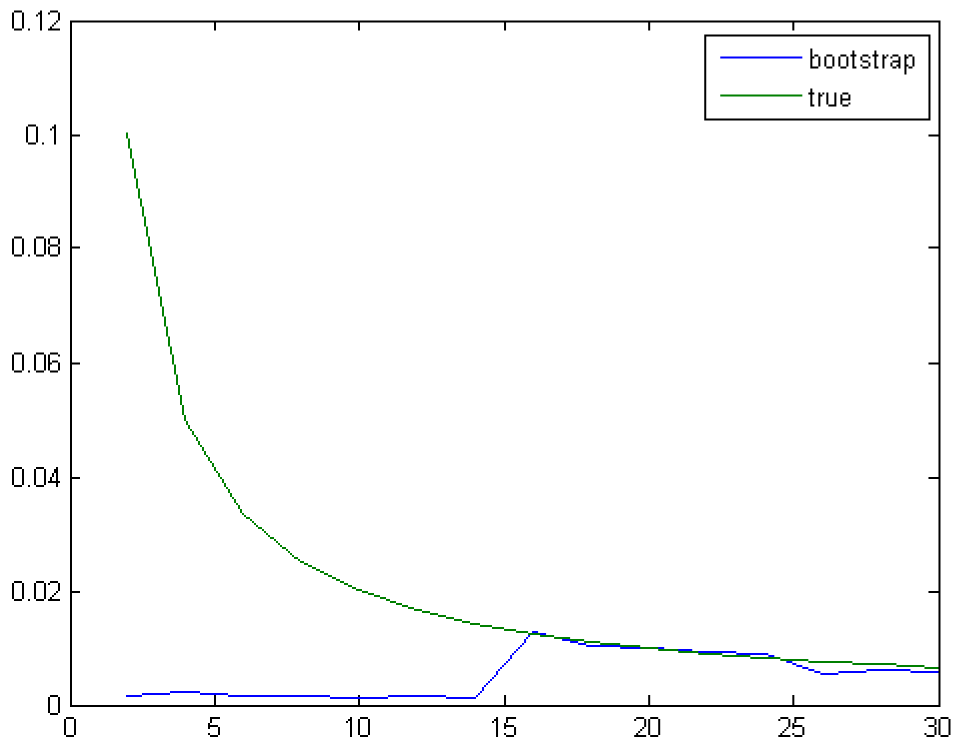

Figure 2: Variance estimate vs n for $p = 0.05$

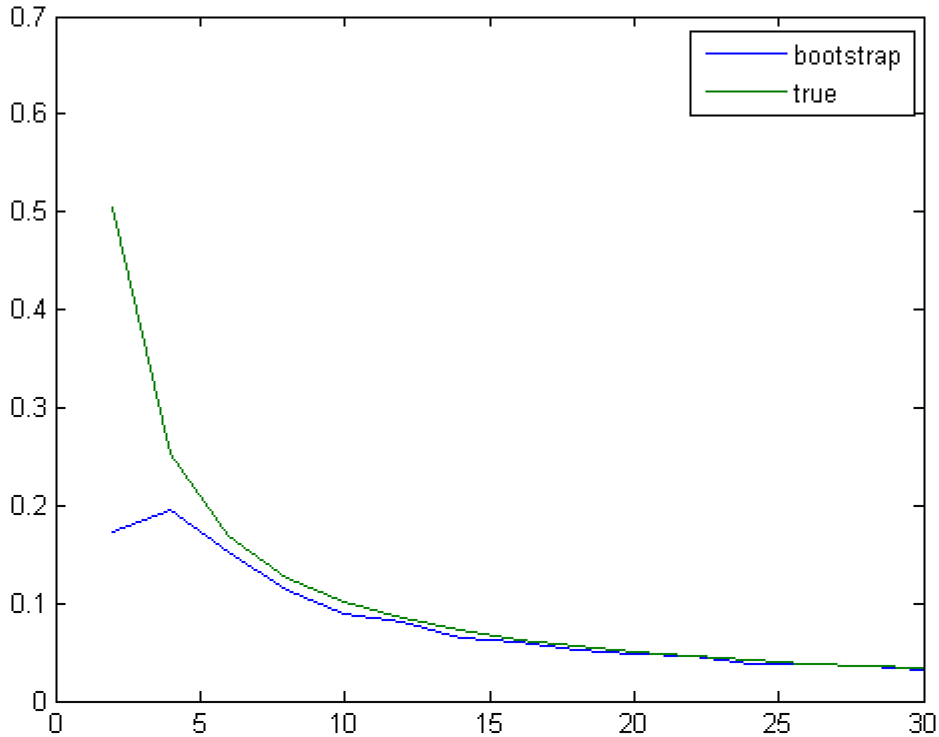

Figure 3: Variance estimate vs n for $p = 0.5$

4. We see that for small n, the median is severely below the true variance. For large n, however, the median matches the true variance pretty well. The behavior for small n can be explained by the fact that there is a substantial probability of having all points in the bootstrap resamples (or even in the original sample) from the high probability mode ($\mu = 1$). Which results in underestimation of the variance due to ignoring the other mode.

5. See figure 3. We see that the bootstrap matches the true variance (on median) even for small n. This is because both modes have equal probabilities which means our resamples are likely to contain points from both modes and hence better represent the actual distribution.

3 Bootstrap Cross-validation

Bootstrap resamples contain repeated points (one third on average) doing CV on these resamples mean that the test partition can contain points that are repeated in the training partition. This results in underestimating the generalization error especially for classifiers that are prone to overfitting (such as 1-NN).