Optimization and Machine Learning, Spring 2020
Reference Solutions for Homework 3

1. (a) Consider the linear regression from a probabilistic perspective. Suppose we are given a set of N observations of the input vector $\mathbf{x}$, which we denote collectively by a data matrix $\mathbf{X}$ whose n-th row is $\mathbf{x}_n^T$ with $n = 1, \cdots, N$. The corresponding target values are $\boldsymbol{t} = (t_1, \cdots, t_N)^T$. We can express uncertainty over the value of target variable using a probability distribution. Assume that given the data $\mathbf{x}_n$ and coefficient vector $\mathbf{w}$, the corresponding value of t_n has a Gaussian distribution with variance σ^2. If the data are assumed to be drawn independently, then the likelihood function is given by

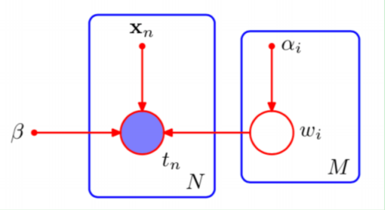
$$p(\mathbf{t} \mid \mathbf{X}, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} p(t_n \mid \mathbf{x}_n, \mathbf{w}, \sigma^2). \tag{1}$$

Next we similarly introduce a prior distribution over the parameter vector $\mathbf{w}$, we shall consider a zero-mean Gaussian prior with variance α_i for each w_i. Assume that the parameter variables are independent. Thus the parameter prior takes the form

$$p(\mathbf{w} \mid \alpha) = \prod_{n=1}^{M} \mathcal{N}(w_i \mid 0, \alpha_i^{-1}). \tag{2}$$

Draw a directed probabilistic graphical model corresponding to the relevance vector machine described by equations (1) and (2). (5 points)

Solution: Introduce a graphical notation that allows such multiple nodes to be expressed more compactly, in which we draw a single representative node tn and then surround this with a box, called a plate, labelled with N indicating that there are N nodes of this kind.



(b) Consider the model defined in (a). Suppose we are given a new input data $\hat{x}$ and we wish to find the corresponding probability distribution for $\hat{t}$ conditioned on the observed data. The graphical model that describes this problem is shown in following Fig. 1. Please give the corresponding joint distribution of all of the random variables in this model and conditioned on the deterministic parameters, i.e., $p(\hat{t}, \mathbf{t}, \mathbf{w} \mid \hat{x}, \mathbf{x}, \alpha, \sigma^2)$. (5 points)

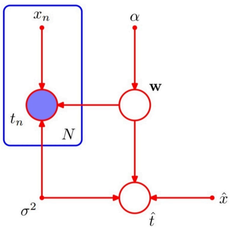


Figure 1: The graphical model.

Solution:

$$p(\hat{t}, \mathbf{t}, \mathbf{w} \mid \hat{x}, \mathbf{x}, \alpha, \sigma^2) = \left[\prod_{n=1}^{N} p(t_n \mid \mathbf{x}_n, \mathbf{w}, \sigma^2)\right] p(\mathbf{w} \mid \alpha) p(\hat{t} \mid \hat{x}, \mathbf{w}, \sigma^2).$$

2. According to the following Fig. 2, use the D-separation to analyze the following cases:

   (a) Given x_4, $\{x_1, x_2\}$ and $\{x_6, x_7\}$ are conditionally independent. (5 points)

   (b) Given $\{x_6, x_7\}$, x_3 and x_5 are conditionally independent. (5 points)

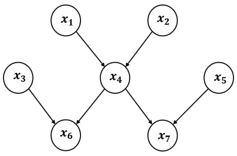


Figure 2: The Bayesian network for questions 2 and 3.

Solution:

(a) The statement is True. According to D-separation, $\{x_1, x_2\}$ and $\{x_6, x_7\}$ can be regarded as two sets A and B. All the arrows on the path form A to B meet head-to-tail, therefore all the paths are blocked given x_4.

(b) The statement is False. The arrow on the path from x_3 to x_4 meets head-to-head. Since the node x_6 is observed, the path from x_3 to x_4 is not blocked. The path from x_3 to x_4 is the same. The path from x_6 to x_7 is also unblocked, therefore x_3 and x_5 are not conditionally independent.

3. According to the Fig. 2, if all the nodes are observed and boolean variables, please complete the process of learning the parameter $\theta_{x_4|i,j}$ by using **MLE**, where $\theta_{x_4|i,j} = p(x_4 = 1 \mid x_1 = i, x_2 = j), i, j \in \{0, 1\}$. (15 points)

Solution: Suppose we observed K data points. Let $\theta = \{\theta_{x_1}, \theta_{x_2}, \theta_{x_3}, \theta_{x_5}, \theta_{x_4|i,j}, \theta_{x_6|i,j}, \theta_{x_7|i,j}\}$, then

$$\begin{aligned}
\log p(\mathcal{D} \mid \theta) &= \log \prod_{k=1}^{K} p(x_{1k}, x_{2k}, x_{3k}, x_{4k}, x_{5k}, x_{6k}, x_{7k} \mid \theta) \\
&= \log \prod_{k=1}^{K} p(x_{1k} \mid \theta) p(x_{2k} \mid \theta) p(x_{3k} \mid \theta) p(x_{5k} \mid \theta) p(x_{4k} \mid x_{1k}, x_{2k}, \theta) p(x_{6k} \mid x_{3k}, x_{4k}, \theta) p(x_{7k} \mid x_{4k}, x_{5k}, \theta) \\
&= \sum_{k=1}^{K} \log p(x_{1k} \mid \theta) + \log p(x_{2k} \mid \theta) + \log p(x_{3k} \mid \theta) + \log p(x_{5k} \mid \theta) + \log p(x_{4k} \mid x_{1k}, x_{2k}, \theta) \\
&\quad + \log p(x_{6k} \mid x_{3k}, x_{4k}, \theta) + \log p(x_{7k} \mid x_{4k}, x_{5k}, \theta).
\end{aligned}$$

Then we derive the gradient of $\log p(\mathcal{D} \mid \theta)$ with respect to $\theta_{x_4|i,j}$

$$\frac{\partial \log p(\mathcal{D} \mid \theta)}{\partial \theta_{x_4|i,j}} = \sum_{ik=1}^{K} \frac{\partial p(x_{4k} \mid x_{1k}, x_{2k}, \theta)}{\partial \theta_{x_4|i,j}}$$

Set the derivative to 0 and then obtain the parameter $\theta_{x_4|i,j}$

$$\theta_{x_4|i,j} = \frac{\sum_{k=1}^{K} \mathbb{I}(x_{4k} = 1, x_{1k} = i, x_{2k} = j)}{\sum_{k=1}^{K} \mathbb{I}(x_{1k} = i, x_{2k} = j)},$$

where $\mathbb{I}(\cdot)$ is the indicator function.

4. Define a Bayesian network with five discrete variables, represented by $\{F, A, S, H, N\}$. $\{F, A, H, N\}$ are 0/1 binary variables and $S \in \{0, 1, 2\}$, as illustrated in Fig. 3. Among them, $\{F, A, N\}$ are observed variables and $\{S, H\}$ are latent variables. Now we implement EM algorithm for this model.

   (a) If all five variables are observed, derive MLE of this model. You should state the close-form solution for each parameter you define. (5 points)

   (b) At least how many parameters should be defined for EM algorithm? (2 points)

   (c) Derive the E-step. You should enumerate each term. (4 points)

   (d) Derive the M-step. (4 points)

Solution:

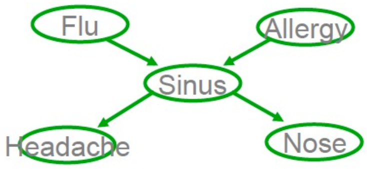


Figure 3: The Bayesian network for question 4.

(a) Define $\theta_f = P(F=1), \theta_a = P(A=1), \theta_{s|f,a} = P(S=s|F=f, A=a), \theta_{h|s} = P(H=1|S=s), \theta_{n|s} = P(N=1|S=s)$. Suppose there are K data points. The likelihood function is

$$\begin{aligned} l(\theta) &= \prod_{k=1}^{K} P(x_k|\theta) = \prod_{k=1}^{K} P(f_k, a_k, s_k, h_k, n_k|\theta) \\ &= \prod_{k=1}^{K} P(f_k)P(a_k)P(s_k|f_k, a_k)P(h_k|s_k)P(n_k|s_k) \end{aligned}$$

Set the derivative of log likelihood function with respect to each parameter to 0, the solutions are:

$$\begin{aligned} \theta_f &= \frac{\sum_{k=1}^{K} \delta(f_k = 1)}{K}, \\ \theta_a &= \frac{\sum_{k=1}^{K} \delta(a_k = 1)}{K}, \\ \theta_{s|f,a} &= \frac{\sum_{k=1}^{K} \delta(s_k = s|f_k = f, a_k = a)}{\sum_{k=1}^{K} \delta(f_k = f, a_k = a)}, \\ \theta_{h|s} &= \frac{\sum_{k=1}^{K} \delta(h_k = 1|s_k = s)}{\sum_{k=1}^{K} \delta(s_k = s)}, \\ \theta_{n|s} &= \frac{\sum_{k=1}^{K} \delta(n_k = 1|s_k = s)}{\sum_{k=1}^{K} \delta(s_k = s)}. \end{aligned}$$

(b) At least 16 variables for θ.

(c) In E-step, calculate $P(S, H|F, A, N, \theta)$.

$$\begin{aligned} P(s_k = 0, h_k = 0|f_k, a_k, n_k, \theta) &= \frac{P(s_k = 0, h_k = 0, f_k, a_k, n_k|\theta)}{\sum_{i=0}^{2}\sum_{j=0}^{1} P(s_k = i, h_k = j, f_k, a_k, n_k|\theta)}, \\ P(s_k = 0, h_k = 1|f_k, a_k, n_k, \theta) &= \frac{P(s_k = 0, h_k = 1, f_k, a_k, n_k|\theta)}{\sum_{i=0}^{2}\sum_{j=0}^{1} P(s_k = i, h_k = j, f_k, a_k, n_k|\theta)}, \\ P(s_k = 1, h_k = 0|f_k, a_k, n_k, \theta) &= \frac{P(s_k = 1, h_k = 0, f_k, a_k, n_k|\theta)}{\sum_{i=0}^{2}\sum_{j=0}^{1} P(s_k = i, h_k = j, f_k, a_k, n_k|\theta)}, \\ P(s_k = 1, h_k = 1|f_k, a_k, n_k, \theta) &= \frac{P(s_k = 1, h_k = 1, f_k, a_k, n_k|\theta)}{\sum_{i=0}^{2}\sum_{j=0}^{1} P(s_k = i, h_k = j, f_k, a_k, n_k|\theta)}, \\ P(s_k = 2, h_k = 0|f_k, a_k, n_k, \theta) &= \frac{P(s_k = 2, h_k = 0, f_k, a_k, n_k|\theta)}{\sum_{i=0}^{2}\sum_{j=0}^{1} P(s_k = i, h_k = j, f_k, a_k, n_k|\theta)}, \\ P(s_k = 2, h_k = 1|f_k, a_k, n_k, \theta) &= \frac{P(s_k = 2, h_k = 1, f_k, a_k, n_k|\theta)}{\sum_{i=0}^{2}\sum_{j=0}^{1} P(s_k = i, h_k = j, f_k, a_k, n_k|\theta)}. \end{aligned}$$

(d) In M-step, choose θ' which maximize $E_{P(S,H|F,A,N,\theta)} \log P(S, H, F, A, N|\theta')$, where

$$\begin{aligned} &E_{P(S,H|F,A,N,\theta)} \log P(S, H, F, A, N|\theta') \\ &= \sum_{k=1}^{K}\sum_{i=0}^{2}\sum_{j=0}^{1} P(s_k = i, h_k = j|f_k, a_k, n_k, \theta)[\log P(f_k) + \log P(a_k) + \log P(s_k|f_k, a_k) + \log P(h_k|s_k) + \log P(n_k|s_k)]. \end{aligned}$$

5. Consider a set of K binary variables x_i, where $i = \{1, ..., K\}$, each variable $x_i \sim Bern(\mu_i)$. So $P(\mathbf{x}|\boldsymbol{\mu}) = \prod_{i=1}^{K} \mu_i^{x_i}(1-\mu_i)^{1-x_i}$, where $\mathbf{x} = (x_1, ..., x_K)^T$ *and* $\boldsymbol{\mu} = (\mu_1, ..., \mu_K)^T$. The mean and covariance of this distribution are easily seen to be $\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}$ and $\text{cov}[\mathbf{x}] = \text{diag}\{\mu_i(1-\mu_i)\}$.

   Now define a finite mixture of N Bernoullis given by $P(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \pi_n P(\mathbf{x}|\boldsymbol{\mu_n})$ where $\boldsymbol{\mu} = \{\boldsymbol{\mu_1}, ..., \boldsymbol{\mu_N}\}$, $\boldsymbol{\pi} = \{\pi_i, ..., \pi_N\}$ and $P(\mathbf{x}|\boldsymbol{\mu_n}) = \prod_{i=1}^{K} \mu_{ni}^{x_i}(1-\mu_{ni})^{1-x_i}$.

   (a) Derive the mean of the mixture distribution. (5 points)

   (b) Show the covariance of the mixture distribution equals $\sum_{n=1}^{N} \pi_n\{\boldsymbol{\Sigma_n} + \boldsymbol{\mu_n}\boldsymbol{\mu_n}^T\} - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{x}]^T$, where $\boldsymbol{\Sigma_n} = \text{diag}\{\mu_{ni}(1-\mu_{ni})$. (5 points)

   Solution:

   (a)

$$\begin{aligned}\mathbb{E}(P(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\pi})) &= \sum_{n=1}^{N} \pi_n \mathbb{E}(P(\mathbf{x}|\boldsymbol{\mu_n})) \\ &= \sum_{n=1}^{N} \pi_n \boldsymbol{\mu_n}\end{aligned}$$

   (b)

$$\begin{aligned}\text{cov}[\mathbf{x}] &= \mathbb{E}[\mathbf{x}\mathbf{x}^T] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{x}]^T \\ &= \sum_{n=1}^{N} \pi_n \mathbb{E}_n[\mathbf{x}\mathbf{x}^T] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{x}]^T \\ &= \sum_{n=1}^{N} \pi_n\{\boldsymbol{\Sigma_n} + \boldsymbol{\mu_n}\boldsymbol{\mu_n}^T\} - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{x}]^T\end{aligned}$$

6. Derive EM algorithm for the mixture of Bernoulli distributions above. There are D data points in total, where $\mathbf{X} = \{\mathbf{x_1}, ..., \mathbf{x_D}\}$. (15 points)

   Solution: The log likelihood function for the model is

$$\log P(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\pi}) = \sum_{d=1}^{D} \log\{\sum_{n=1}^{N} \pi_n P(\mathbf{x_d}|\boldsymbol{\mu_n})\}.$$

   Assume the latent variable $\mathbf{z} = (z_1, ..., z_N)^T$ is a binary N-dimensional variable having only a single component equal to 1. So we have

$$P(\mathbf{x}|\mathbf{z}, \boldsymbol{\mu}) = \prod_{n=1}^{N} P\left(\mathbf{x}|\boldsymbol{\mu}_n\right)^{z_n}, \ P(\mathbf{z}|\boldsymbol{\pi}) = \prod_{n=1}^{N} \pi_n^{z_n}.$$

   If we form the product of $P(\mathbf{x}|\mathbf{z}, \boldsymbol{\mu})$ and $P(\mathbf{z}|\boldsymbol{\pi})$ and then marginalize over $\mathbf{z}$, then we obtain $P(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \pi_n P\left(\mathbf{x}|\boldsymbol{\mu}_n\right)$.
   Then the log likelihood function for complete-data is

$$\log P(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\pi}) = \sum_{d=1}^{D}\sum_{n=1}^{N} z_{dn}\{\log \pi_n + \sum_{k=1}^{K}[x_{dk}\log \mu_{nk} + (1-x_{dk})\log(1-\mu_{nk})]\}.$$

   Expecting over $\mathbf{Z}$, we have

$$\mathbb{E}_{\mathbf{Z}}[\log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\pi})] = \sum_{d=1}^{D}\sum_{n=1}^{N} \gamma\left(z_{dn}\right)\left\{\log \pi_n + \sum_{k=1}^{K}\left[x_{dk}\log \mu_{nk} + (1-x_{dk})\log\left(1-\mu_{nk}\right)\right]\right\},$$

   where $\gamma\left(z_{dn}\right) = \mathbb{E}[z_{dn}] = \frac{\sum_{z_{dn}} z_{dn}[\pi_n p(\mathbf{x}_d|\boldsymbol{\mu}_n)]^{z_{dn}}}{\sum_{z_{dj}}[\pi_j p(\mathbf{x}_d|\boldsymbol{\mu}_j)]^{z_{dj}}} = \frac{\pi_n p(\mathbf{x}_d|\boldsymbol{\mu}_n)}{\sum_{j=1}^{N}\pi_j p(\mathbf{x}_d|\boldsymbol{\mu}_j)}$.

   In M-step, set the derivative with respect to π_n to 0, we obtain $\pi_n = \frac{\sum_{d=1}^{D}\gamma(z_{dn})}{D}$. Set the derivative with respect to $\boldsymbol{\mu_n}$ to 0, we obtain $\boldsymbol{\mu_n} = \frac{\sum_{d=1}^{D}\gamma(z_{dn})\mathbf{x_d}}{\sum_{d=1}^{D}\gamma(z_{dn})}$.

7. Hoeffding's inequality is a powerful technique—perhaps the most important inequality in learning theory for bounding the probability that sums of bounded random variables are too large or too small. Below are some related inequalities you are required to provide proof:

   (a) **(Markov's inequality).** Let $Z \geq 0$ be a non-negative random variable. Then for all $t \geq 0$, show that

$$\mathbb{P}(Z \geq t) \leq \frac{\mathbb{E}(Z)}{t}, \tag{3}$$

   where $\mathbb{E}$ denotes the expectation operator. (6 points)

   (b) **(Chebyshev's inequality).** Let $Z \geq 0$ be a random variable with $\mathrm{Var}(Z) < \infty$. Show that

$$\mathbb{P}(Z \geq \mathbb{E}[Z] + t \text{ or } Z \leq \mathbb{E}[Z] - t) \leq \frac{\mathrm{Var}(Z)}{t^2}, \qquad \text{for } t \geq 0, \tag{4}$$

   where $\mathrm{Var}(Z)$ denotes the variance of Z. (6 points)

Solution:

   (a) *Proof.* We note that $\mathbb{P}(Z \geq t) = \mathbb{E}[\mathbf{1}\{Z \geq t\}]$, and that if $Z \geq t$, then it must be the case that $Z/t \geq 1 \geq \mathbf{1}\{Z \geq t\}$, while of $Z < t$, then we still have $Z/t \geq 0 = \mathbf{1}\{Z \geq t\}$. Thus

$$\mathbb{P}(Z \geq t) = \mathbb{E}[\mathbf{1}\{Z \geq t\}] \leq \mathbb{E}\left[\frac{Z}{t} = \frac{\mathbb{E}[Z]}{t}\right],$$

   as desired. □

   (b) *Proof.* The result is an immediate consequence of Markov's inequality. We note that either $Z \geq \mathbb{E}(Z) + t$ or $Z \leq \mathbb{E}[Z] - t$, we have $(Z - \mathbb{E}(Z))^2 \geq t^2$. Thus,

$$\begin{aligned}\mathbb{P}(Z \geq \mathbb{E}[Z] + t \text{ or } Z \leq \mathbb{E}[Z] - t) &= \mathbb{P}((Z - \mathbb{E}(Z))^2 \geq t^2) \\ &\leq \frac{\mathbb{E}[(Z - \mathbb{E}(Z))^2]}{t^2} = \frac{\mathrm{Var}(Z)}{t^2},\end{aligned}$$

   where the inequality holds due to the Markov's inequality. □

8. Recall that to show VC dimension is d for hypotheses $\mathcal{H}$ can be done via showing that VC $\dim(\mathcal{H}) \leq d$ and VC $\dim(\mathcal{H}) \geq d$. More specifically, to prove that VC $\dim(\mathcal{H}) \geq d$ it suffices to give d examples that can be shattered; to prove VC $\dim(\mathcal{H}) \leq d$ one must show that no set $d + 1$ examples can be shattered.

   For each one of the following function classes, find the VC dimension. State your reasoning based on the presented hint above. (Note that: solutions with the correct answer but without adequate explanation will not earn marks.)

   (a) **Halfspaces in** $\mathbb{R}^2$. Examples lying in or on the halfspace are labeled $+1$, and the remaining examples are labeled -1. (3 points)

   (b) **Axis-parallel rectangles in** $\mathbb{R}^2$. Points lying on or inside the target rectangle are labeled $+1$, and points lying outside the target rectangle are labeled -1. (3 points)

   (c) **Closed sets in** $\mathbb{R}^2$. All points lying in the set or on the boundary of the set are labeled $+1$, and all points lying outside the set are labeled -1. (3 points)

   (d) How many training examples suffice to assure with probability 0.9 that a consistent learner using the function classes presented in (b) will learn the target function with accuracy of at least 0.95? (4 points) (Hint: we use the following bounds on sample complexity: $m \geq \frac{1}{\epsilon}(4\log_2(2/\delta) + 8\text{VC} \dim(\mathcal{H})\log_2(13/\epsilon))$).

Solution:

   (a) It is easily shown that any three non-collinear points (e.g., $(0, 1)$, $(0, 0)$, $(1, 0)$) are shattered by $\mathcal{H}$. Thus, VC $\dim(\mathcal{H}) \geq 3$. We now show that no set of size four can be shattered by $\mathcal{H}$. If at least three of the points are collinear then there is no halfspace that contains the two extreme points but does not contain the middle points. Thus the four points cannot be shattered if any three are collinear. Next, suppose that the points form a quadrilateral. There is no halfspace which labels one pair of diagonally opposite points positive and the other pair of diagonally opposite points negative. The final case is that one point p is in the triangle defined by the other three. In this case there is no halfspace which labels p differently from the other three. Thus clearly the four points cannot be shattered. Therefore we have demonstrated that VC $\dim(\mathcal{H}) = 3$.

(b) First, it is easily seen that there is a set of four points (e.g., $(0,1)$, $(0,-1)$, $(1,0)$, $(-1,0)$) that can be shattered. Thus VC $\dim(\mathcal{H}) \geq 4$. We now argue that no set of five points can be shattered. The smallest bounding axis-parallel rectangles defined by the five points is in fact defined by at most four of the points. For p a non-defining point in the set, we see that the set cannot be shattered since it is not possible for p to be classified as negative while also classifying the others as positive. Thus VC $\dim(\mathcal{H}) = 4$.

(c) Any set can be shattered by $\mathcal{H}$, since a closed set can assume any shape in $\mathbb{R}^n$. Thus, the largest set that can be shattered by $\mathcal{H}$ is infinite, and hence VC $\dim(\mathcal{H}) = \infty$.

(d) The bound is $m \geq \frac{1}{\epsilon}(4\log_2(2/\delta) + 8\text{VC }\dim(\mathcal{H})\log_2(13/\epsilon))$. Then just by plugging in the numbers (VC $\dim(\mathcal{H}) = 4$, $\delta = 0.1$ and $\epsilon = 0.05$), we have $m \geq 5480$.