Optimization and Machine Learning, Spring 2020
Homework 3
(Due Tuesday, Apr. 28 at 11:59pm (CST))

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)

(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)

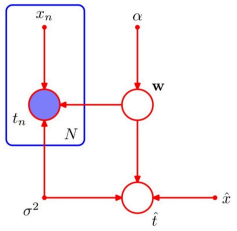

Figure 1: The graphical model.

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)

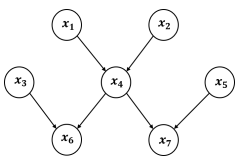

Figure 2: The Bayesian network for questions 2 and 3.

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)

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)

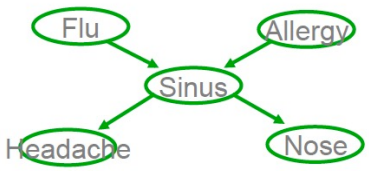

Figure 3: The Bayesian network for question 4.

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)

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)

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 $\text{Var}(Z) < \infty$. Show that

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

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

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: $m \geq \frac{1}{\epsilon}(4\log_2(2/\delta)+8\text{VC dim}(\mathcal{H})\log_2(13/\epsilon))$).