Date: Wednesday, November 11, 2020. 4:00 pm - 5:10 pm
There are 5 questions. All Questions Carry 6 points.
Attempt any 4 out of 5. Max Points: 24.
You must start answer to each question on a new page.

If answering all questions, You MUST clearly specify the question that you do not want to be graded (in your answer sheet). Otherwise, we may randomly pick any 4.

1. **Logistic Regression with Non-Linear Separator**: Consider a learning problem where the training data is not linearly separable. Let the training data be given as $\{x^{(i)}, y^{(i)}\}_{i=1}^{m}$. Assume that class labels are Boolean valued, and each $x^{(i)} \in \mathcal{R}^n$. Assume that there is a polynomial curve of degree K which can separate the data. One way to learn this classifier is to first transform the input features in a higher dimensional polynomial space with degree K and learn a linear classifier there. We would like to learn a model using logistic regression.

   (a) Describe the general form of the polynomial classifier of degree K as described above.

   (b) Consider the polynomial transformation $\phi(x) = \sum_{k=0}^{K} a_k x^k$, such that for $x, z \in \mathcal{R}^n$, $\phi(x)^T \phi(z) = x^T z + c$, where $c \in \mathcal{R}$. Show that entire parameter update algorithm for logistic regression can be written in terms of $\phi(x)^T \phi(z)$, without ever explicitly computing $\phi(x)$ for any given input x.

   (c) Give one practical advantage of the scheme as described above in terms of the computational complexity of the learning algorithm (as opposed to explicitly computing $\phi(x)$'s).

   Hint: You will have to think of a scheme where you can implicitly represent θ's in terms of the coefficients of input features $x^{(i)}$'s.

2. **Locally Weighted Linear Regression and Positive Semi-Definiteness**

   (a) Consider a learning problem with training data $\{x^{(i)}, y^{(i)}\}_{i=1}^{m}$. Write down the objective function $J_x(\theta)$ for locally weighted linear regression as defined in the class. Assume that the objective is being defined for prediction at a give point x. Next, convert this objective into the matrix form using design matrix X, label vector Y and parameter vector θ. Recall that each row of X represents an example, Y is a vector of all the class labels, and θ represents the entire parameter vector. You will also need to define a weight matrix W to incorporate the weights assigned to each point $x^{(i)}$.

   (b) Show that the objective function as defined above is a convex function of the parameters θ. You may want to do this by showing that the corresponding Hessian matrix is positive semi-definite.

3. **Gaussian Discriminant Analysis (GDA) with more than 2 classes**: Consider a learning problem with training data $\{x^{(i)}, y^{(i)}\}_{i=1}^{m}$. Let there be $2r$ class labels,i.e.,

$y^{(i)} \in \{1, 2, \cdots 2r\}$. Consider modeling this problem using GDA. Let the mean vectors for conditional distribution of $x|y = k$ be given by μ_k $(1 \le k \le 2r)$. Let the co-variance parameters be tied such that for the conditional distribution $x|y = k$, the co-variance matrix is given by Σ_1 for $1 \le k \le r$, and Σ_{r+1} for $r + 1 \le k \le 2r$. In other words, the co-variance matrix for feature distributions for the first r class labels is Σ_1, and for the remaining it is Σ_{r+1}. Assume that the class prior probabilities are given by vector Φ where $\Phi \in \mathcal{R}^{2r}$, such that $0 \le \phi_k \le 1$, and the ϕ_k's add up to 1.

(a) Mathematically describe how would you find the decision boundary for class label 1. Your answer needs to be mathematically precise and not just a bunch of English statements. Characterize the shape of this boundary in terms of whether it is linear, quadratic, or some other form. What factors does it depend on?

(b) Clearly characterize the expression for the decision boundary in terms of the mean-parameters $(\mu_1, \mu_2, \cdots, \mu_{2r})$ and co-variance parameters Σ_1 and Σ_{r+1}.

4. **Naive Bayes and MAP Estimation:** Assume that we are interested in computing the parameters of text classification under a Multinoulli event model. Assume the training data is given as $\{x^{(i)}, y^{(i)}\}_{i=1}^{m}$.

   (a) Write down the expression for the maximum-likelihood (ML) estimate of the parameters $\theta_{j=l|k}$ $(1 \le l \le |V|, k \in \{1, \cdots r\})$.

   (b) Modify the expression derived in the part above to add the appropriate smoothing term, as described in class. Briefly justify the term that you add to the numerator and the denominator.

   (c) In this part, we will work with the MAP estimate of the parameters $\theta_{j=l|k}$ $(1 \le l \le |V|, k \in \{1, 2, \cdots, r\}$. Assume that the vocabulary size is $|V| = 2$ (this is a toy problem so that you can work out the maths). In this case, for every k, there are two parameters, $\theta_{j=1|k}$, and $\theta_{j=2|k}$, such that $(\theta_{j=1|k} + \theta_{j=2|k}) = 1$. Let the prior over the parameters $\theta_{j=1|k}$ be given as $\mathrm{P}(\theta_{j=1|k}) \propto \theta_{j=1|k}(1 - \theta_{j=1|k})$. Show that using this prior for the MAP estimation leads to an estimate which is equivalent to the smoothed estimate under ML-likelihood estimation, as derived above. Your answer should be mathematically justified.

5. **Solve the following Problems**:

   (a) Assume that you are dealing with a population where the fraction of COVID +ve patients is given by 0.01 (1 in 100). Further, assume that your RTPCR test has a success rate of 0.98 if the person is COVID +ve, and has a success rate of 0.97 if the person is COVID -ve. Here, by success rate we mean the probability with which the test gives the correct result. Given that two persons were independently tested for COVID, and both labeled COVID +ve by the test, compute the overall probability that both of them are COVID +ve in reality. You can assume that the prior probabilities of them being COVID +ve are independent of each other.

   (b) Consider the Bayesian network shown in Figure 1 (see next page).

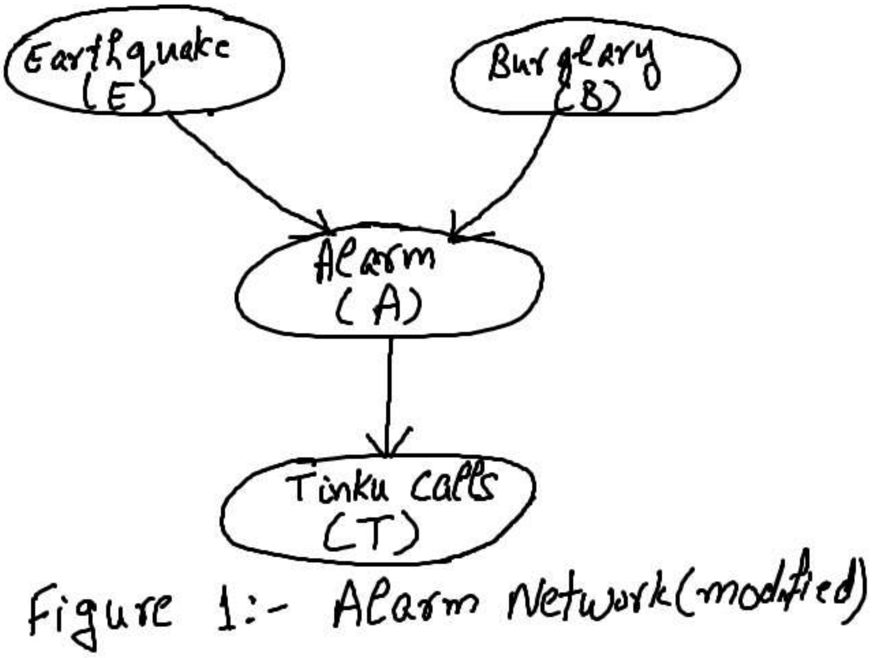

Figure 1: Alarm Network (modified)

i. Write down the expression for the joint distribution for this network in its factorized form. Describe the exact number of (independent) parameters in this network. Assume all the variables to be Boolean valued.

ii. Is Earthquake (E) independent of Burglary (B) given Tinu Calls (T)? Why or why not? Justify.