**Date: Monday, August 31st, 2020. 9:30am - 11:45 pm**
**There are 6 questions. All Questions Carry 5 points. Max Points: 30**

1. **Elliptical SVMs:** Consider a binary classification with training data $\{x^{(i)}, y^{(i)}\}_{i=1}^{m}$. Assume, $x^{(i)} \in \mathcal{R}^n, \forall i$. In this problem, we would like to learn a non-linear (quadratic) separator using the principle of max-margin in SVMs, and correspondingly design a kernel function for the same. Specifically, we are interested in learning the hypothesis of the following form:
$h_\theta(x) = \mathbb{1}\{\sum_{j=1}^{n} \frac{{x_j}^2}{{a_j}^2} <= 1\ \}$. Here, $(a_1, \cdots, a_n)$ are the parameters of the model. This is the generalization of an elliptical boundary centered at origin, to $\mathcal{R}^n$.

   (a) Design an appropriate kernel function such that a linear SVM in the transformed space can learn functions of the form given above. Your kernel should be as precise as possible so as to be able to learn quadratic functions exactly of the above form (and nothing outside).

   (b) Formulate the linear SVM problem in the transformed space. Clearly specify your objective, and the constraints. Note that additional constraints on parameters may be required to ensure that only boundaries of the form as specified above are learned, and nothing else.

   (c) Derive the dual problem corresponding to the primal problem obtained in part (b) above.

   In the above setting, you can assume that data is separable using the form of the decision boundary specified as above.

2. **Generalizing PCA:** Consider a set of data points $\{x^{(i)}\}_{i=1}^{m}$, where each $x^{(i)} \in \mathcal{R}^n$. We have covered PCA in class, which transforms the data linearly, so as to maximize the variance along principal components. In this problem, we will do a step by step computation, to generalize PCA to non-linear spaces. Let $\phi : \mathcal{R}^n \rightarrow \mathcal{R}^N$ be a feature transformation, and we are interested in finding the principal components of the data in this transformed space. Assume that $\phi$ is such that if original data is 0 mean, so is data in the transformed space. Assume that there is a function $K : \mathcal{R}^n \times \mathcal{R}^n \rightarrow \mathcal{R}$, where $K(x^{(i)}, x^{(j)}) = \phi(x^{(i)})^T \phi(x^{(j)})$. Let $\{u_l^{\phi}\}_{l=1}^{N}$ denote the set of principal components in the transformed space (there are $N$ of them).

   (a) Write down the objective function for PCA in terms of the co-variance matrix $\Sigma^{\phi}$, where $\Sigma^{\phi}$ denotes the co-variance matrix of the data in the transformed space. Show that each principal component $u_l^{\phi}$ can be written as a linear combination of the given set of data points in the transformed space, i.e., $u_l^{\phi} = \sum_i a_l^i \phi(x^{(i)})$.

   (b) Show that the $a_l$'s can be purely written only in terms of dot products of the form $\phi(x^{(i)})^T \phi(x^{(j)})$, i.e., only using the function $K$ (dot-products), as defined earlier. There should be no other terms involving individual $\phi(x^{(i)})$ terms in your expression.

   (c) Use the part (b) above to write down a Matrix, expressed only in terms of the $K$ function, such that $a_l$'s can be computed as eigen-vectors of this matrix.

(d) Show that the projections, $\phi(x^{(i)})^T u_l^{\phi}$ can be computed purely in terms of the function $K$, without even explicitly referring to the feature transformation.

3. **Logistic Regression and Convexity:** Consider a very simple learning problem with only two data points : $\{x^{(i)}, y^{(i)}\}_{i=1}^{2}$. $x^{(i)} \in \mathcal{R}^n$. Assume Boolean target labels. Assume that the two points are distinct, one with a positive label and one with a negative label. Consider learning a logistic regression model over this simple setting. Without loss of generality, assume the first point to have a positive label, and second point to have a negative label.

   (a) Explicitly write down the negative log-likelihood objective $LL(\theta)$ for the above setting. Show that the objective $LL(\theta)$ is a convex function of the $\theta$ parameters.

   (b) Prove that the optimal decision boundary will be equidistant from the two points, and will assign equal probability of being +ve and -ve, to the two points, respectively.

   (c) Show that objective function $LL(\theta)$, though convex, has no local minima in the above setting. Further, show that this fact this does not pose any issue for learning the optimal decision boundary, from a practical stand point.

4. **Gradients in CNN with tied weights:** Consider a CNN architecture, with input image given as a 3-D feature map of size $w \times w \times r$, where $w$ is the height/width of the feature map, and $r$ is its depth. Assume that the feature map is convolved with $r$ kernels, of size $k \times k$, followed by *average-pooling* over a $2 \times 2$ neighborhood, and a stride of 1. Assume that the size of feature map remains the same after convolution and pooling operation (due to sufficient 0 padding on the right and bottom side). This set of operations is applied $d$ times, with tied set of parameters. In other words, the weights of the $r$ kernels are identical across different layers (note that the $r$ kernels within a single layer are free to have independent parameters). Assume that the application of alternating convolutional and pooling layers results in a final feature map $F$ of size $w \times w \times r$, which is processed further using a fully connected layer, followed by a softmax layer. You have been given the gradient $\nabla_F \mathcal{L}$ of the final loss $\mathcal{L}$ with respect to the feature map $F$. Derive the gradients with respect to the kernel parameters $k \times k \times r$ in this architecture, in terms of $\nabla_F \mathcal{L}$, and the parameters given above, using back-propagation.

5. **Mixture Model for Naive Bayes:** Assume that you are given training data $\mathcal{D} = \{x^{(i)}, y^{(i)}\}_{i=1}^{m}$, over which you would like to learn a mixture of Naïve Bayes model. The intuition for this mixture model comes from the Gaussian Mixture Model (GMM) learned. Recall that GMM learned a mixture of Gaussian models, where given the latent mixture component $z^{(i)} = l$, the conditional distribution $P(x^{(i)}|z^{(i)} = l)$ was assumed to be a multi-variate Gaussian. In this problem, we will replace the conditional Gaussian distributions in the mixture, by Naïve Bayes models. Assume that we are learning a mixture of Naïve Bayes models, with $k$ components (latent). Note that, in this case, since we are learning a mixture over Naïve Bayes, $z^{(i)}$'s denote the latent mixture components, and $y^{(i)}$'s play the role of the class label in Naïve Bayes, and

$x^{(i)}$'s are the features as earlier. Once the mixture component is known, we specify a different Naïve Bayes model for each mixture component. To keep things simple you can assume both $y^{(i)}$'s and $x^{(i)}$'s are Boolean valued. Further, you can assume that prior probability $P(y = 1)$ (denoted by parameter $\phi$ using the notation in class), is independent of the mixture component chosen.

(a) Given training data $\mathcal{D} = \{x^{(i)}, y^{(i)}\}_{i=1}^{m}$, write down the complete generative process for the mixture of Naïve Bayes model as described above. Clearly specify the parameters of your model. Write down the expression for log-likelihood for the above model.

(b) Consider learning the parameters of the above model, using the EM (Expectation Maximization) set up as described in the class. Specifically, the E-step would estimate the mixture component, and $E$ step would estimate the parameters of the corresponding Naïve Bayes model. Clearly describe the expression for E-step, and the M-step.

✓ 6. **VC-dimension of axis parallel cuboids:** In the class, we derived the VC-dimension of the hypothesis class $\mathcal{H}_{||}^{2}$, consisting of axis parallel rectangles. Here, each rectangular hypothesis is specified by the set of parameters $\theta = ((x_1, y_1), (x_2, y_2))$ where $x_1, y_1, x_2, y_2 \in \mathcal{R}$. Given a point $(x, y) \in \mathcal{R}^2$, and $h_\theta \in \mathcal{H}_{||}^{2}$, $h_\theta(x, y) = 1$ if $x_1 \leq x \leq x_2 \wedge y_1 \leq y \leq y_2$, 0 otherwise. We will now extend this hypothesis class to axis parallel cuboids in 3-D, denoted by $\mathcal{H}_{||}^{3}$. Here, each hypothesis is now parameterized by set of parameters $\theta = ((x_1, y_1, z_1), (x_2, y_2, z_2))$, where, $x_1, y_1, z_1, x_2, y_2, z_2 \in \mathcal{R}$. Correspondingly, given a point $(x, y, z) \in \mathcal{R}^3$, and $h_\theta \in \mathcal{H}_{||}^{3}$, we have $h_\theta(x, y, z) = 1$ if $x_1 \leq x \leq x_2 \wedge y_1 \leq y \leq y_2 \wedge z_1 \leq z \leq z_2$, 0 otherwise. Derive the VC-dimension of this hypothesis $\mathcal{H}_{||}^{3}$. How would you extend this argument to the hypothesis class $\mathcal{H}_{||}^{n}$ of axis parallel cuboids in $\mathcal{R}^n$: give an argument with justification.