PRNN 2024 - Minor 1

Questions by Prathosh A P

25 April 2024

Instructions: (a) The exam is for **30 + 1 (bonus) marks** for **180 minutes**. (b) Please refrain from writing a lot of text but answer to the point with appropriate mathematical equations.

1. **Estimation and Bayesian Decision theory**

(a) Consider a mixture model for a prior given as follows: $p(\theta) = \sum_k p(z = k)p(\theta|z = k)$ where each conditional $p(\theta|z = k)$ is conjugate to a given likelihood. Prove that the mixture is also a conjugate prior. (2)

(b) Given a dataset D, any function s of the dataset is called a statistic. A statistic is said to be 'sufficient' for a parameter θ if $p(D|s, \theta)$ is independent of θ. With this definition, Show that a statistic is sufficient for θ if the density $p(D|\theta)$ can be written as the product $P(D|\theta) = g(s, \theta)h(D)$, for some function of $h()$ and $g()$. (2)

(c) Suppose we have a binary classification problem with d-dimensional features with Gaussian Class-conditional densities. Find out a sufficient statistic for $P(D|\theta)$ and express a Bayes's classifier in terms of the obtained sufficient statistic (2).

2. **Nearest Neighbors and Linear Models**

(a) Suppose we are designing a 1-nearest neighbour based classifier with n datapoints. Show that this rule divides the feature space into Dirichlet tessellation: The partitioning of the space into **convex polygons** such that each polygon contains exactly one generating point and every point in a given polygon is closer to its generating point than to any other. Hint: Convexity implies, that for any two points $\mathbf{x}1$ and $\mathbf{x}2$ in a polygon, all points on the line linking $\mathbf{x}1$ and $\mathbf{x}2$ must also lie in the same polygon. (2.5)

(b) Suppose we have a binary classification problem in d-dimensions and we are interested in a linear classifier of the form $h(\mathbf{x}) = \mathbf{w}^t\mathbf{x}$. The task in this problem is to find the optimal $\mathbf{w}$, where the optimality is defined as that $\mathbf{w}$, which ~~minimizes~~ max the squared of the difference between the per-class means of the projected data ($\mathbf{w}^t\mathbf{x}$) is the projection). The objective also involves a scaling (multiplication factor) given by the inverse of the sum of the per-class variances of the projected data. Find the expression for such a direction $\mathbf{w}$. (2.5)

3. **SVMs and Kernels**

(a) Show that, irrespective of the dimensionality of the data space, a data set consisting of just two data points, one from each class, is sufficient to determine the location of the maximum-margin hyperplane. (1)

(b) Suppose we have a regression problem with $\mathbf{x_i} \in \mathcal{R}^d$ and $\mathbf{y_i} \in \mathcal{R}$. Let us define a loss function for regression as the epsilon sensitive loss:

$$L_\epsilon(h(\mathbf{x}), \mathbf{y}) = |\mathbf{y} - h(\mathbf{x})|_\epsilon = \max(0, |\mathbf{y} - h(\mathbf{x})| - \epsilon)$$

Here $\mathbf{x}$ is the input, $\mathbf{y}$ is the output, and h is the hypothesis function. Using this notation, let us define a risk function as

$$\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} L_\epsilon(h(\mathbf{x}_i), \mathbf{y}_i)$$

where $h(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$, and $C, \epsilon > 0$ are parameters. With this rewrite this problem as a quadratic problem (i.e. quadratic objective with linear constraints), by introducing appropriate slack variables. Write down the corresponding Lagrangian (3)

(c) Suppose we design a 1 nearest neighbor algorithm in a projected data space $\mathbf{z} = \Phi(\mathbf{x}$ embedded with a Kernel $K(\mathbf{x}_i, \mathbf{x}_j)$. Show that this algorithm does not need to know the projection $\Phi(\mathbf{x}$ as long as the Kernel can be computed (1)

4. **Neural Networks and Gradient Descent**

(a) Suppose $\mathbf{R}(\mathbf{w})$ denotes a scalar-valued Risk function of a vector-valued parameter denoted by $\mathbf{w} \in \mathcal{R}^m$. Derive a gradient-based iterative algorithm to minimize $\mathbf{R}(\mathbf{w})$ w.r.t $\mathbf{w}$ and argue (mathematically) its correctness. (2)

(b) Suppose the above optimization problem is constrained such that the 2-norm of the parameter vector is upper bounded by a constant. Rederive the iterative parameter update equation and argue why it can be called 'parameter decay' iterates (1)

(c) Derive an expression for the dimensions of the l^{th} layer activation in a CNN in terms of the (a) dimensions of the $(l-1)^{th}$ activations, (b) size of the weight matrix, (c) stride size and (iv) ~~pooling~~ padding (1).

(d) Suppose we have 4 datapoints with binary features as follows: $\{(0,0),(0,1),(1,0),(1,1)\}$ with the respective labels as $\{0,1,1,0\}$. Verify whether a 3-layer MLP with all linear layers (without any non-linear activations) can learn to separate the data.

5. **Boosting**

(a) Suppose we are interested in building an ensemble of classifiers on a binary classification problem, iteratively as follows: $H_{t+1} = H_t + \alpha h_{t+1}$ where $h_{t+1} = \text{argmin}_h \hat{R}(H_t + \alpha h)$ where $\hat{R}$ is empirical risk. With the usual notations, derive a gradient-based iterative algorithm to find h_{t+1}. Hint: Consider gradient descent in the function space (2)

(b) Suppose we have an exponential loss function $L(\mathbf{y}, h(\mathbf{x})) = \exp\big(-\mathbf{y}h(\mathbf{x})\big)$. With this loss function, show that the solution to the optimization problem of finding the next classifier reduces to minimizing the weighted classification error of the training samples (2).

(c) Derive an update equation for the weight distribution over the training samples, as obtained by the above iterative algorithm (1)

6. **Unsupervised Learning**

(a) Let $\{\mathbf{x_i}\}_{i=1}^N$, $x_i \in R^d$ be the data given. The objective is to cluster these data points into k clusters $(C_1, C_2, ..., C_k)$ with their centroids at $(\mu_1, \mu_2, ..., \mu_k)$. But unlike k-means clustering, we don't want to assign each data point to an exclusive cluster. Instead, we want to model the probability of a data point x_i, being in cluster C_j as $P(C_j|X_i) = \gamma_{ij}$. This can be achieved by solving the following optimization problem:

$$\gamma_{ij}^*,\ \mu_j^* = \underset{\gamma_{ij},\ \mu_j}{\text{argmin}} \quad \sum_{i=1}^{N}\sum_{j=1}^{k} \gamma_{ij}^2 ||x_i - \mu_j||^2$$

$$\text{s.t} \quad \sum_{j=1}^{k} \gamma_{ij} = 1, \quad \gamma_{ij} \geq 0, \quad i = 1, 2, \cdots, N, j = 1, 2, \cdots, k$$

With this, answer the following questions:

(b) Change the above to an unconstrained optimization problem using Lagrange multipliers. (1)
Hint: You can ignore the positivity constraint and check if it is satisfied after solving the problem.

(c) Solve for μ_j^*, γ_{ij}^* and the lagrangian multipliers. (3)

(d) Derive an iterative algorithm to get optimal μ_j, γ_{ij}. (1)
Hint: The algorithm will look like a distant cousin of k-means.

(e) **Bonus:** How will you quantize the γ_{ij}'s so that it turns out to be the well-known k-means clustering algorithm? (1)