ELL 409/784 Major 2020

Profs. Jayadeva and Prathosh

January 12, 2021

1. Answer all questions below:

   (a) Show that for a binary classification problem, the risk under a zero-one loss is the probability of mis-classification. (1)

   (b) Design a Bayes' classifier with Gaussian class conditional densities with different means and variances and show that it effectively is a polynomial discriminant function in the feature space. What is the degree of the discriminant polynomial? (1)

   (c) Let $D = \{X_1, X_2, ..., X_n\}$ denote the data. A statistic Y defined as any fixed function of data $Y = f\{X_1, X_2, ..., X_n\}$ which can be used to estimate a parameter θ. A statistic is called sufficient with respect to a parameter θ, if the conditional density of the data conditioned on the statistic does not depend on the parameter θ. Show that, sample sum, $Y = \sum_i X_i$ is a sufficient statistic with respect to p if $X_i \sim IID$(Bernoulli) and p represents the success probability of the random variable X. (2)

   (d) Establish the relation between the KL Divergence and Maximum Likelihood estimation (0.5).

   (e) Show that the error of a classifier obtained by adding K classifiers is exponentially less than the errors of the individual classifier (1).

   (f) In a GMM, what does the latent variable represent? Is K-means related to a GMM? If so, how? (0.5)

   (g) Write out the expression for the density estimate with a Parzen window estimator with Gaussian kernels. Interpret all the terms and relate it to a histogram. (1)

   (h) Suppose while learning a linear model, all the weights are restricted to lie within a convex hull in the 3rd quadrant of the Euclidean feature space. How does this effect Bias and Variance of the model and why? (1)

   (i) Suppose we have a data from a 5-th degree polynomial with noise. Can one learn the underlying polynomial using linear regression? If so how? (0.5)

   (j) Consider a 2 layer (with a single hidden layer) fully-connected NN with M hidden units with hyperbolic tan activations without bias terms - What happens to the output of a hidden neuron if signs of the weights feeding

into it are flipped? Can one generate more NNs with the same input-output relationship as the original one, just by such sign-flipping operations on weights of a set of neurons? If so, how many such equivalent NNs can be generated? (1)

(k) What are the three main components in a CNN that is different from an MLP (0.5)

2. Answer the following:

(a) In the empirical feature space, the weight vector w may be written as $w = \sum_{j=1}^{M} \alpha_j \phi(x^j)$, where symbols have their usual meanings. Which of the following is true ?
(a)$0 \leq \alpha_j \leq C$ (b)$0 \leq \alpha_j \leq \infty$ (c)$-\infty \leq \alpha_j \leq C$ (d)$-\infty \leq \alpha_j \leq \infty$

(1 mark)

(b) A set of 4 samples in 2D at the corners of a unit square is transformed to a higher dimensional space. The bottom left corner of the square is at the origin. The kernel used is the RBF kernel. The image vectors are denoted by $\phi(x^j)$. Find the maximum norm of any of the image vector (1 mark).

(c) A set of 10 samples lie on the line $x_1 + 2x_2 + 3 = 0$. Find the principal component of the samples (1 mark)

(d) A linear hyperplane classifier is trained on a set of 10,000 images of 1 megapixel resolution without employing any feature reduction methodology. What is the approximate VC dimension of the classifier (1 mark).

(e) Find the number of linear dichotomies possible with 6 samples in 2D (1 mark).

(f) Two hyperplane classifiers are trained with exactly the same training set and show the same error on the training set. Classifier A uses 10 features while classifier B uses 12. Relate the test errors of both the classifiers (1 mark).

(g) An SVM is used to solve a binary classification problem with a kernel function $K(p, q) \equiv \phi(p)^T \phi(q)$. If the Lagrange multipliers are denoted by λ_i, i = 1, 2, ..., M, and the class labels by y_i, and if the samples are linearly separable, determine: The margin $\|w\|$ in the image space. The separating hyperplane is given by $w^T \phi(x) + b = 0$. Also determine the mean of the image vectors $\phi(x^i)$, i = 1. 2, ..., M.

(1 + 1 marks)

(h) A modified SVM formulation is given by

$$\text{Minimize}\, 0.5(\|w\|^2 + Ab^2) + \frac{C}{2} \sum_{i=1}^{M} q_i$$

subject to the constraints

$$y_i \left(w^T x^i + b \right) + q_i \geq 1$$

$$q_i \geq 0$$

where A is a constant. Determine the Lagrangian.
Write the K.K.T. conditions.
Determine the dual formulation.

(1 + 1 + 1 marks)

3. Answer the Following:

   (a) Finding the principal components of M samples in n dimensions requires determining the eigenvectors of a $n \times n$ covariance matrix XX^T. This is prohibitive in high dimensions, where $n >> M$. Instead, it is possible to determine the eigenvectors of the matrix X^TX and use them to determine the eigenvectors of XX^T. Can you explain how ?

   (3 marks)

   (b) Given a sample x in n dimensions, $\hat{x} = (w^Tx)w$ is projection of x in the direction of a vector w. Show that the choice of w that minimizes the difference between $\hat{x}$ and x is the principal component.

   (3 marks)

   (c) Can a 2-hidden layer neural network with only linear activations be considered equivalent to a single layer neural network. Can this solve the above XOR problem (1)?

   (d) Verify by constructing an example network whether or not a layered neural network with a single hidden node can solve the XOR problem if it has direct connections from the input nodes to hidden nodes as well as output nodes, in addition to connection between hidden and output node. (1)

   (e) Suppose I am interested in automatically detecting the frames when a professor is erasing the black board in NPTEL videos. How do I formulate this problem from an ML perspective and which algorithm(s) do you recommend to solve this problem. Justify.(1)