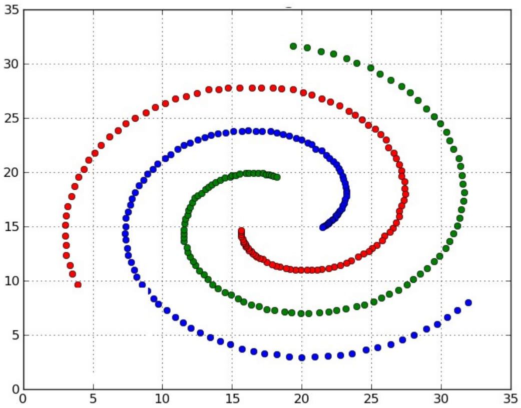
EE769: Introduction to Machine Learning
End-Semester Examination Questions
IIT Bombay, 2018.05.04 :: 09:30 to 12:30
May the fourth be with you!

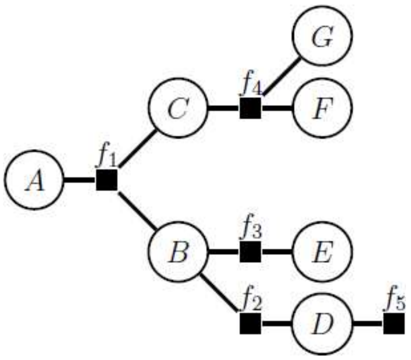
Instructions:

A) Begin each answer at the top of a new page/side of the answer sheet.
B) Max marks 41 (including bonus). Answer as many questions as you can.
C) Only correct and short answers will be given full marks.

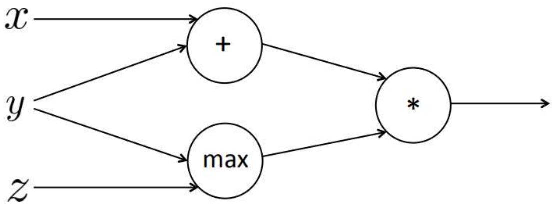
1. Answer the following short and easy questions:
   a. What is the underlying dimension of a helix (like the thread of a screw or a spring) as a manifold in 3-D space? [1]
   b. What is the number of parameters of an n-layer neural network with L_0 inputs, and $L_1, L_2, \dots, L_n$ neurons in the respective subsequent layers? Each neuron has a bias. [2]
   c. Theoretically, any nonlinear function can be used as an activation function in the hidden layers of a neural network such as $\tanh\,(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n + b)$. Is $\operatorname{sign}\,(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n + b)$ a good activation function? Give reason. [2]
   d. Is $\sin\,(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n + b)$ a good activation function? Give reason. [2]
   e. In DBSCAN, if there are too many clusters, then what all can you adjust and how? [2]
   f. In DBSCAN, to discover the clusters in the figure to the right, suggest a good hyper-parameter setting. [2]



2. Examine the factor graph given in the figure to the right.
   a. What is the message from *f₁* to *A* in terms of other messages? [2]
   b. What is the Markov blanket of *B*? [2]
   c. Write an expression for *p*(*E*|*B*=1) assuming binary variables. [2]



3. Design a neural network for binary classification such that the area inside the triangle given by the following points $\{(-5,0), (5,0), (0,10)\}$ in $\mathbb{R}^2$ is class 0, and outside is class 1. Assume that it has a single hidden layer with step function activations, and a single output neuron with step function activation also. [4]



4. Find the derivative of the output of the computation graph shown in the figure with respect to y using backpropagation, assuming $x = 1, y = 2, z = 0$, and $*$ is multiply. Hint: for multiplication $a * b$, derivative with respect to a gets multiplied by b, that is, it is $a'b$. If this is too hard to solve then replace $*$ with a − (minus) and solve the question with one mark penalty. [3]

5. Cross-entropy loss is given by the expression $-\{y\log f(\boldsymbol{x}) + (1-y)\log\left(1-f(\boldsymbol{x})\right)\}$ where desired output $y\epsilon\{0,1\}$ and estimated probability that $y=1$ is $f(x)\epsilon(0,1)$. Generalize this expression to 3-classes. Show that it has the lowest value for correct classification. What is its highest value and for what kind of output? Hint: you cannot use a scalar valued function f and nor can the desired output be a scalar y. You will need to define the desired output as a one-hot-bit vector, the estimated output also as a vector representing a probability mass function. [3]

6. (a) Assume that for a contiguous shape in $\mathbb{R}^2$ you are given an indicator function $inside(x,y)$ which when called in a program gives the value 1 if the point (x,y) lies inside the shape, and 0 if they lie outside the shape. You are asked to compute the approximate area of the shape using Metropolis algorithm. Write the pseudo code for computing the area of the shape using a proposal distribution around the current point that is a bivariate Gaussian with a fixed covariance matrix. Assume that the shape is a rectangle with height *h* and width *w.* [3]

   (b) For the previous part, how you will choose the hyper-parameters of the proposal distribution. Would you need any prior information about the shape? [2]

7. Convert a ridge regression objective function $J(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}\{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n - y_n\}^2 + \frac{\lambda}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}$ (where $\boldsymbol{w}$ is weight vector, $\boldsymbol{x}_n$ is a given input data vector, y_n is the desired scalar output, and $\lambda > 0$ is the weight for the L2 penalty) into its dual such that the solution is in terms of ${\boldsymbol{x}_n}^{\mathrm{T}}\boldsymbol{x}_n$ or in terms of kernel matrix $\boldsymbol{K}\epsilon\mathbb{R}^{N\times N} = \boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$, where $\boldsymbol{X}$ is the data matrix whose each row is an input instance ${\boldsymbol{x}_n}^{\mathrm{T}}$. Hint: try to express $\boldsymbol{w}$ as $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{a}$ and get $J(\boldsymbol{a})$ instead. [3]

8. In the figure the dashed lines represent a Cartesian grid, and not any data points.
   a. Sketch the approximate cluster *boundaries* for k-means with k={3,4,5}. [3]
   b. Sketch the graph of Davies-Bouldin index given by $\frac{1}{k}\sum_{i=1}^{k}\max_{i\neq j}\left(\frac{\sigma_i+\sigma_j}{\|\mu_i-\mu_j\|}\right)$, for k={3,4,5}, where σ_i is the average distance of points in cluster i to its centroid, μ_i is the centroid of cluster i, and ||.|| represents Euclidean distance. Exact values are not important, only the shape of the graph is needed. [2]