CS771 Final Exam: Answer Key for Subjective Questions

Short answer questions (3 marks)

Q) Can we perform MLE for the parameters of a gaussian mixture model without using ALT-OPT or EM? If yes, how? If no, why not?
A) Yes, we can take the gradients of the incomplete data log likelihood and perform gradient descent (though EM or ALT-OPT is better due to cleaner and closed form updates in each iteration)
Rubric: If says yes and mentions gradient descent in the answer then give 3, otherwise 0.

Q) What is the difference between incomplete data log likelihood and complete data log likelihood? Also, does every model have these two types of likelihoods?
A) In latent variable models, ILL is the likelihood (probability of data given parameters) obtained after marginalizing the latent variables (so they are not present in the expression) whereas CLL is the joint probability distribution of data and latent variables given parameters. Not all probabilistic models have both types -- only LVMs will have both (e.g., logistic regression, despite being a probabilistic mode, doesn't have such separate notions).
Rubric: 2 marks for mentioning the difference correctly. 1 more mark for correctly answering whether every model has both types of likelihood.

Q)Consider a neural network for scalar-valued regression. Assume the network has two hidden layers with connection weight matrices \(W_1\) and \(W_2\) and one output layer with weight vector \(v\). If there is no nonlinear activation after each hidden layer, can the model learn a nonlinear regression? Justify your answer using only text (using at most 2-5 sentences).
A) No, the model will effectively still be a linear regression model with weight vector being W1*W2*v because, without the nonlinear activations, the output will simply be v*W2*(W1*x)
Rubric: Answers no and correctly justified it. Just saying no activations hence nonlinear will not be sufficient -- at least some reasoning such as successive linear transforms being still linear (in words or in equations) is necessary

Q) Rank (with a brief justification) the following classification (assume binary) methods in terms of their speed at test time, with fastest first and slowest last: KNN, kernel SVM, LwP, decision tree (assume each node tests a single feature and the number of level is small), deep neural network (assuming the hidden layer computations take negligible time). If two methods are equally fast, you may say so.
A) DT, LwP, deep neural net, kernel SVM, KNN. Note that the deep neural net will be faster than kernel methods since it does not need to store training inputs at test time. To predict the label, we simply pass the input through the network, and then after the last layer, it is just like a linear model.
Rubric: 3 marks for fully correct ordering; 2 marks if DT first and KNN last and some minor error in ordering the rest

Q) Assuming 100 dimensional inputs, how many filters does an MLP with one hidden layer and 5 hidden units would learn, and what is the size/dimension of each filter?
A) 5 filters, each 100 dimensional
Rubric: 3 marks for correct answer.

Q) What is the difference in the form of the cluster assignment vector z_n for the following types of clustering algorithms: hard clustering, soft clustering, overlapping clustering ?
A) z_n is one-hot for hard clustering, a probability vector for soft clustering, and a binary vector for overlapping clustering
Rubric: One mark each for answering each correctly.

Q) Why might PCA not be a good choice to reduce data dimensionality if our end goal is to learn a classification model?
A) Because the maximum variance directions along which PCA projects the data may not be the same as the directions along which data is well-separated.
Rubric: 3 marks for correct explanation, and 0 otherwise (nothing in between)

Q) Briefly describe how/in what ways the Lloyd's algorithm for solving the K-means problem is analogous to the EM or ALT-OPT algorithms?
A) The input-to-cluster assignment step in Lloyd's algorithm is like the step in EM where the posterior probability of an input going to each cluster is computed (or in ALT-OPT where we simply assign it to the cluster with largest posterior probability). The mean computation in Lloyd's algorithm is like the step in EM (and ALT-OPT) where the parameters are updated given the latent variables (cluster ids in this case).
Rubric: 3 marks for correct explanation, and 0 otherwise (nothing in between)

Q) In what ways, GMM with expectation maximization is better than a soft K-means clustering algorithm?
A) Although EM for GMM also gives the cluster membership probabilities like soft K-means, it computes the posterior probability of the point being assigned to a cluster and this probability depends on (1) how many points belong to that cluster; and (2) the shape of the cluster. Thus EM for GMM uses extra information in the probability computation.
Rubric: 3 marks if mentioned both how many points in each cluster and the share of cluster. If only shape is mentioned, give 2 marks. Otherwise give 0.

Q) After using the landmarks or random features approach to construct kernel based features, what is done next? Also, how is prediction made given a new test input?
A) We train a linear model (say with weight vector) using the extracted features, At test time, we first extract the landmark or random features based features for the test input and apply the linear model as w'\psi(x).
Rubric: 1.5 marks for the first step correct. 1.5 marks more for second step correct

Q) The kernel ridge regression weight vector can be obtained as $w = X^\top \alpha$ where X is $N\times D$ feature matrix and $\alpha = (K+\lambda I_N)^{-1}y$ where K is the

$(N\times N)$ kernel matrix and (y) is the $(N \times 1)$ response vector. Although this model is typically used for nonlinear regression, is there any benefit of using this solution if learning a linear ridge regression model? Provide a justification using only words (you may use some symbols if needed).
A) Yes, this approach requires inverting an NxN matrix whereas the standard closed form solution for ridge regression requires inverting a DxD matrix. If N is smaller than D then the kernel ridge regression approach will be more efficient even for linear ridge regression (so basically, it means we will run kernel ridge regression with a linear kernel).
Rubric: 3 marks for the correct answer (mentions expensive inversion of DxD matrix); 0 otherwise

Q) What is the relationship between generative classification with Gaussian class conditionals and a Gaussian mixture model? Answer only using words (max 2-3 sentences).
A) GMM is the same as generative classification but with the class labels being unknown (and taking the form of cluster ids which are treated as latent variables).
Rubric: 3 marks for the correct answer; 0 otherwise

Long answer questions

Note: The solutions below are only to give you a brief idea.

Q) Soft-clustering objective and optimization problem
A) The objective will be of the form ||X-Z\mu||^2. The subproblems will be (1) with \mu fixed, solving \argmin_Z ||X-Z\mu||^2 or \argmin_{z_n}||x_n - \mu z_n||^2 with non-neg and sums to 1 constraints on each z_n; and (2) with Z fixed, solving \argmin_\mu ||X - Z\mu||^2 which is an unconstrained problem.
Rubric (roughly; exact breakup of the marking scheme also depends on the overall solution): 2 marks for writing the expression for overall optimization problem
3 marks for writing the optimization subproblem for \mu
3 marks for writing the optimization subproblem for Z (1 mark if constraint for z_n is missing)

Q) Generative classification with labeled and unlabeled data
A) This is a combination of generative classification and GMM. In the E step, we make soft guesses of the labeled of the unlabeled examples by computing their posterior expectation, just like we do in GMM, and in the M step, we use both labeled examples (with true labels) and unlabeled examples (with their guessed soft labels) to estimate the parameters. Skipping the equations here but they will be similar to what we saw for generative classification and GMM in the lectures.
Rubric (roughly; exact breakup of the marking scheme also depends on the overall solution): 3 marks for the description (in words) of how to solve the problem3 marks for expected CLL expression6 marks for the EM algo with update equations for all latent variables and parameters (if only gave basic steps without update equations then only 2 marks)

Q) Matrix factorization

A) To solve for a_n, treat the b_1,b_2,...,b_M as input feature vectors (equivalent to thinking of B as the MxK feature matrix with each b_m as one of the rows) and X_{n1}. X_{n2},...,X__{nM} as the respective outputs (equivalent to arranging them as X_n -- an Mx1 vector of responses) and learn a regression model with a_n as the unknown weight that maps B to X_n.

To solve for b_m, treat the a_1,a_2,...,a_N as input feature vectors (equivalent to thinking of A as the NxK feature matrix with each a_n as one of the rows) and X_{1m}. X_{2m},...,X__{Nm} as the respective outputs (equivalent to arranging them as X_m -- an Nx1 vector of responses) and learn a regression model with b_n as the unknown weight that maps A to X_m.

Due to the independence of the subproblems involving a_n's, all a_n's can be solved for in parallel, with B fixed at its current value. Likewise, all b_m's can be solved for in parallel with A fixed at its current value.

Rubric (roughly; exact breakup of the marking scheme also depends on the overall solution): 3 marks: Correctly stating how to estimate a_n (give only 1 mark if feature matrix or response vector not correctly specified)

3 marks: Correctly stating how to estimate b_m (give only 1 mark if feature matrix or response vector not correctly specified)

2 marks: For sketching the algorithm

2 marks: For stating correctly which problems can be solved in parallel and why.

Note: Showing a direct update for all a_n (as a matrix A) and all b_m (as a matrix B) using gradient methods or even in closed form is not the correct way since the problem asked you to show for each a_n and each b_m.