Course Home
Announcements
Resources
Forums
Hangout
My Profile
Logout
Help
Feedback
Back to Portal

Select Language
English

powered by
mooKIT

Submission Deadline : 18/12/2020 19:10

End-semester Examination

Q.1 To converge, gradient descent must use the same learning rate in all dimensions of the optimization variable.

Max. score: 1; Neg. score: 0

[Imp. Note: If you wish to skip the question, you should do so immediately. Once an option is chosen (either true or false), you can not skip the question at a later stage.]

- false
- true

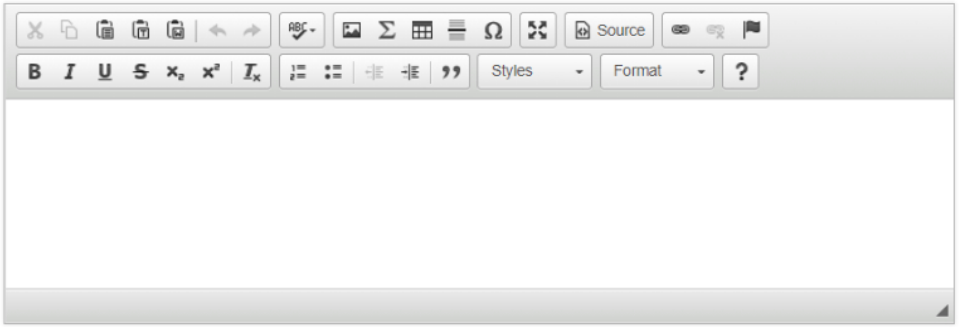
Q.2 Why might PCA not be a good choice to reduce data dimensionality if our end goal is to learn a classification model?

Max. score: 3; Neg. score: 0

This is a long answer type question. You can either upload a file or type your answer below.

UPLOAD A FILE

or



Q.3 Consider 4 training inputs {(0; 0); (1; 0); (0; 1); (1; 1)} with labels {+1;-1;-1; +1}. Which of the following models is capable of learning a separator with zero training error?

Max. score: 2; Neg. score: 0

- Perceptron
- Multi-layer Perceptron

Kernel SVM

Learning with Prototypes

Q.4 (For this question, you may either typeset your answer in LaTeX and upload the PDF, or write your solution on a piece of paper and upload the pictures of the solution, or directly write your solution in the provided text-box below which supports writing equations via LaTeX as well as inserting symbols. However, please use only one of these methods)

Consider a K class generative classification problem with Gaussian class-conditionals. We are given N labeled examples $\{(x_n, y_n)\}_{n=1}^{N}$ and M additional unlabeled inputs $\{x_n\}_{n=N+1}^{N+M}$. The goal is learn the parameters $\{(\pi_k, \mu_k, \Sigma_k)\}_{k=1}^{K}$ of this generative classification models.

(1) In words (max 2-5 sentences), briefly describe how you can use an Expectation Maximization (EM) algorithm to estimate the parameters of this model, and clearly mention what are the latent variables in this EM algorithm.

(2) Write down the (expected) complete data log-likelihood that EM will maximize for this problem. You need not show all the steps in the derivation. If some of the steps are familiar (e.g., seen in the lectures), you may directly write those with a brief explanation.

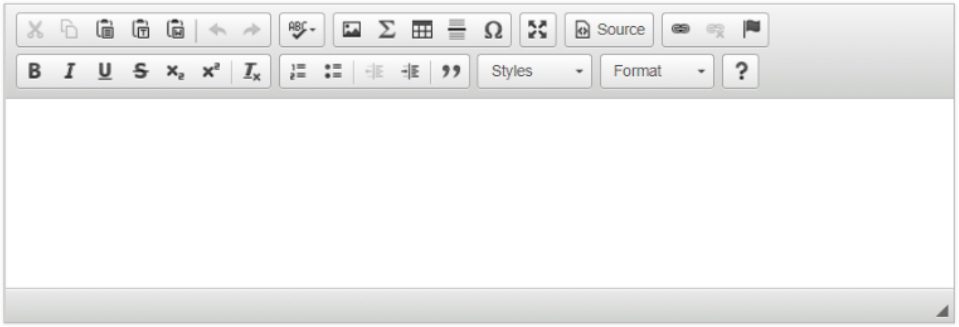
(3) Write down the complete EM algorithm and, for each of the steps of the EM algorithm, clearly write down the mathematical expressions for estimating the various unknowns of this model.

Max. score: 12; Neg. score: 0

This is a long answer type question. You can either upload a file or type your answer below.

UPLOAD A FILE

or



Q.5 Standard SVM is a discriminative model, i.e., it does not model the inputs.

Max. score: 1; Neg. score: 0

[Imp. Note: If you wish to skip the question, you should do so immediately. Once an option is chosen (either true or false), you can not skip the question at a later stage.]

false

false

true

Q.6 Which of the following is true about the Gaussian/RBF kernel function, with some reasonable value of the bandwidth parameter?

Max. score: 2; Neg. score: 0

- [ ] Its maximum possible value for a pair of inputs is infinity
- [ ] For two inputs located very very close, its value will be close to 0
- [ ] Its associated feature map is always infinite dimensional
- [ ] For two inputs located very very close, its value will be close to 1

Q.7 For a quadratic function, Newton's method will converge to the global optima in one step.

Max. score: 1; Neg. score: 0

[Imp. Note: If you wish to skip the question, you should do so immediately. Once an option is chosen (either true or false), you can not skip the question at a later stage.]

- false
- true

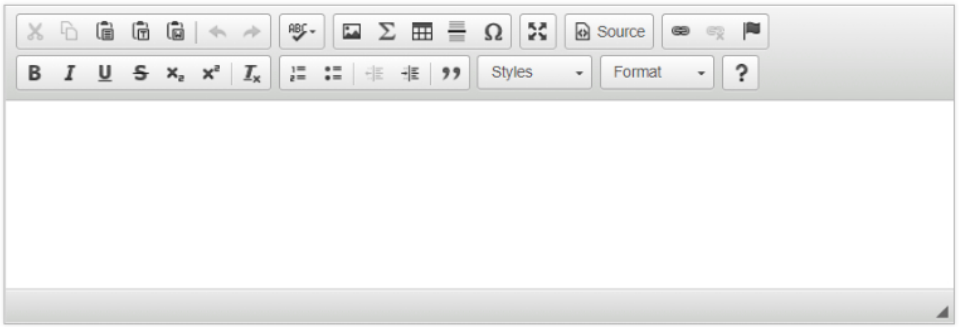
Q.8 Rank (with a brief justification) the following classification (assume binary) methods in terms of their speed at test time, with fastest first and slowest last: KNN, kernel SVM, LwP, decision tree (assume each node tests a single feature and the number of level is small), deep neural network (assuming the hidden layer computations take negligible time). If two methods are equally fast, you may say so.

Max. score: 3; Neg. score: 0

This is a long answer type question. You can either upload a file or type your answer below.

UPLOAD A FILE

or



Q.9 Which of the following tasks a multi-layer Perceptron (MLP) can be used for (assuming no other learning algorithm is used in addition to MLP)?

Max. score: 2; Neg. score: 0

- [ ] Nonlinear classification
- [ ] Nonlinear dimensionality reduction
- [ ] Nonlinear regression
- [ ] Nonlinear clustering

Q.10 Which of the following models cannot be learned without using gradient based methods? Answer this based only on the knowledge you have from what was covered in the course (not from some research paper that you may have read somewhere :))

Max. score: 2; Neg. score: 0

- [ ] Ridge regression
- [ ] Logistic regression
- [ ] Kernelized ridge regression
- [ ] Multi-layer perceptron

Q.11 PCA selects a subset of the features from the original features.

Max. score: 1; Neg. score: 0

[Imp. Note: If you wish to skip the question, you should do so immediately. Once an option is chosen (either true or false), you can not skip the question at a later stage.]

- ( ) false
- ( ) true

Q.12 For an RBF kernel $k(x_n, x_m) = \exp(-\gamma||x_n - x_m||^2)$, if the bandwidth parameter γ is set to 0, the kernel matrix (assuming we have N training inputs) becomes rank N.

Max. score: 1; Neg. score: 0

[Imp. Note: If you wish to skip the question, you should do so immediately. Once an option is chosen (either true or false), you can not skip the question at a later stage.]

- ( ) false
- ( ) true

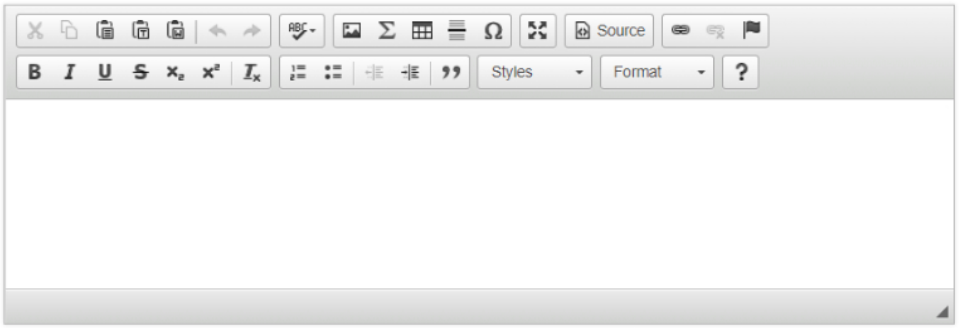
Q.13 Can we perform MLE for the parameters of a gaussian mixture model without using ALT-OPT or EM? If yes, how? If no, why not?

Max. score: 3; Neg. score: 0

This is a long answer type question. You can either upload a file or type your answer below.

UPLOAD A FILE

or



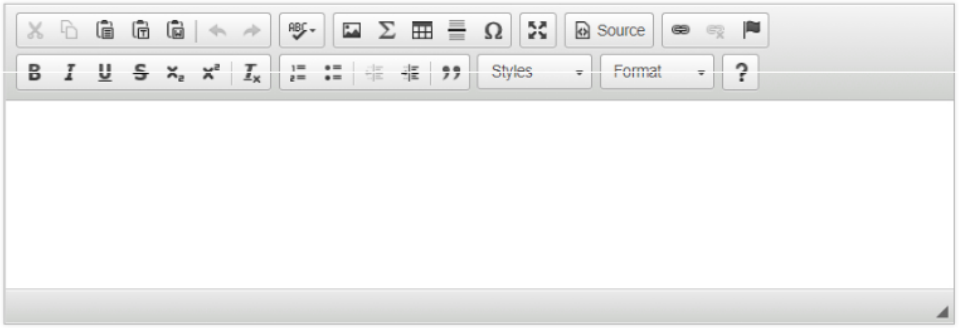
Q.14 After using the landmarks or random features approach to construct kernel based features, what is done next? Also, how is prediction made given a new test input?

Max. score: 3; Neg. score: 0

This is a long answer type question. You can either upload a file or type your answer below.

UPLOAD A FILE

or



Q.15 What is the difference between incomplete data log likelihood and complete data log likelihood? Also, does every model have these two types of likelihoods?

Max. score: 3; Neg. score: 0

This is a long answer type question. You can either upload a file or type your answer below.

UPLOAD A FILE

or

Source Styles Format ?

Q.16 Instead of a Gaussian prior which has a form $p(w) \propto \exp(-||w||^2)$, using a Laplace distribution $p(w) \propto \exp(-||w||_1)$ as the prior will be equivalent to using ℓ_0 regularization for the weight vector.

Max. score: 1; Neg. score: 0

[Imp. Note: If you wish to skip the question, you should do so immediately. Once an option is chosen (either true or false), you can not skip the question at a later stage.]

- false
- true

Q.17 GMM with ALT-OPT would not give a soft clustering.

Max. score: 1; Neg. score: 0

[Imp. Note: If you wish to skip the question, you should do so immediately. Once an option is chosen (either true or false), you can not skip the question at a later stage.]

- false
- true

Q.18 What is the relationship between generative classification with Gaussian class conditionals and a Gaussian mixture model? Answer only using words (max 2-3 sentences).

Max. score: 3; Neg. score: 0

This is a long answer type question. You can either upload a file or type your answer below.

UPLOAD A FILE

or

Source Styles Format ?

Q.19 Probabilistic PCA can be used for nonlinear dimensionality reduction.

Max. score: 1; Neg. score: 0

[Imp. Note: If you wish to skip the question, you should do so immediately. Once an option is chosen (either true or false), you can not skip the question at a later stage.]

- false
- true

Q.20 Which of these methods can be used for nonlinear dimensionality reduction?

Max. score: 2; Neg. score: 0

- PCA
- Locally Linear Embedding (LLE)
- PCA with features constructed with kernel based landmarks
- Matrix Factorization of the form $X \approx UV$ where U and V are matrices with a small number of columns and rows, respectively.

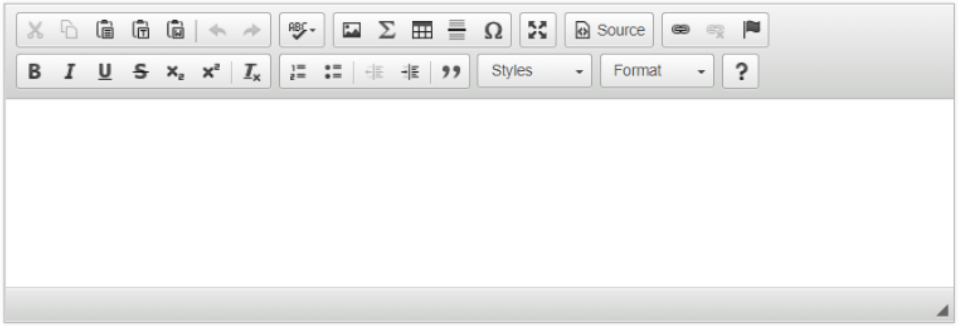
Q.21 In what ways, GMM with expectation maximization is better than a soft K-means clustering algorithm?

Max. score: 3; Neg. score: 0

This is a long answer type question. You can either upload a file or type your answer below.

UPLOAD A FILE

or



Q.22 Which of the following is true about data clustering?

Max. score: 2; Neg. score: 0

- It can be seen as a method for learning a new feature representation of the data
- There is only one way to cluster the inputs, i.e., only one clustering is possible for a given dataset
- It can be seen as a method to reduce the dimensionality if the number of clusters is some small value
- It expects the inputs to be given in form of vectors

Q.23 Which of the following is true about PCA?

Max. score: 2; Neg. score: 0

- It can be used to compress the number of inputs
- Its kernel variant (kernel PCA) is used if we want to increase the size of the feature representation
- It can be used to compress the feature representation of each input
- Its computational cost depends on the original dimension of the inputs

---

Q.24 The function $|\boldsymbol{a}\boldsymbol{x} + \boldsymbol{b}|$ of variable $\boldsymbol{x}$, assuming $\boldsymbol{a}$ and $\boldsymbol{b}$ are non-negative, is convex.

Max. score: 1; Neg. score: 0

[Imp. Note: If you wish to skip the question, you should do so immediately. Once an option is chosen (either true or false), you can not skip the question at a later stage.]

- false
- true

Q.25 A deep neural network is essentially equivalent to several linear models stacked on top of each other.

Max. score: 1; Neg. score: 0

[Imp. Note: If you wish to skip the question, you should do so immediately. Once an option is chosen (either true or false), you can not skip the question at a later stage.]

- false
- true

Q.26 Which is these is not a latent variable model?

Max. score: 2; Neg. score: 0

- [ ] GMM
- [ ] Probabilistic PCA
- [ ] Probabilistic Linear Regression
- [ ] Multilayer Perceptron

Q.27 Doing MLE based parameter estimation is equivalent to parameter estimation by minimizing a regularized loss function.

Max. score: 1; Neg. score: 0

[Imp. Note: If you wish to skip the question, you should do so immediately. Once an option is chosen (either true or false), you can not skip the question at a later stage.]

- ( ) false
- ( ) true

Q.28 Expectation-Maximization is insensitive to initialization.

Max. score: 1; Neg. score: 0

[Imp. Note: If you wish to skip the question, you should do so immediately. Once an option is chosen (either true or false), you can not skip the question at a later stage.]

- ( ) false
- ( ) true

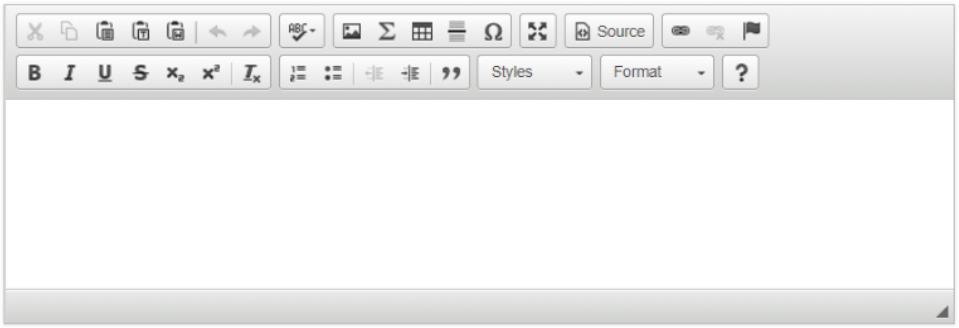
Q.29 Assuming 100 dimensional inputs, how many filters does an MLP with one hidden layer and 5 hidden units would learn, and what is the size/dimension of each filter?

Max. score: 3; Neg. score: 0

This is a long answer type question. You can either upload a file or type your answer below.

UPLOAD A FILE

or



Q.30 (For this question, you may either typeset your answer in LaTeX and upload the PDF, or write your solution on a piece of paper and upload the pictures of the solution, or directly write your solution in the provided text-box below which supports writing equations via LaTeX as well as inserting symbols. However, please use only one of these methods)

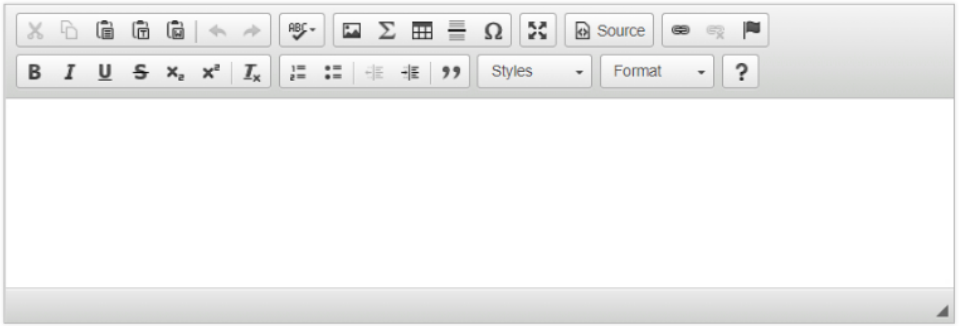
Assume you are given \(N\) inputs \(x_1,x_2,\ldots,x_N\) with each \(x_n \in \mathbb{R}^D\) and you wish to perform soft clustering for this data. Write down the expression for soft clustering as a matrix factorization problem and briefly describe how you would use an ALT-OPT scheme to solve this problem. For each sub-problem of this ALT-OPT scheme, (1) clearly specify the variable being optimized for, along with the constraints, if any; (2) briefly state how this sub-problem can be solved (you do not need to solve them in full detail; just a brief explanation is needed)

Max. score: 8; Neg. score: 0

This is a long answer type question. You can either upload a file or type your answer below.

UPLOAD A FILE

or



Q.31 Assuming we are using \(\ell_2\) regularization, increasing the regularization hyperparameter will tend to make the magnitude of the entries of the learned weight vector smaller.

Max. score: 1; Neg. score: 0

[Imp. Note: If you wish to skip the question, you should do so immediately. Once an option is chosen (either true or false), you can not skip the question at a later stage.]

- false
- true

Q.32 A convex function, at the point of non-differentiability, has a unique sub-gradient.

Max. score: 1; Neg. score: 0

[Imp. Note: If you wish to skip the question, you should do so immediately. Once an option is chosen (either true or false), you can not skip the question at a later stage.]

false

true

Q.33 It is in general easier to learn decision trees with discrete features than decision trees with real-valued features.

Max. score: 1; Neg. score: 0

[Imp. Note: If you wish to skip the question, you should do so immediately. Once an option is chosen (either true or false), you can not skip the question at a later stage.]

false

true

Q.34 Bag of words is a continuous representation for text data.

Max. score: 1; Neg. score: 0

[Imp. Note: If you wish to skip the question, you should do so immediately. Once an option is chosen (either true or false), you can not skip the question at a later stage.]

false

true

Q.35 For a scalar-valued regression problem with 10-dimensional features, the number of weights to be learned for a single hidden layer neural network with 3 hidden units will be

Max. score: 2; Neg. score: 0

10

33

11

34

Q.36 Which of these kernels will have a finite-dimensional feature mapping?

Max. score: 2; Neg. score: 0

Quadratic

Sum of linear kernel and Gaussian kernel

Gaussian

Sum of quadratic kernel and Gaussian kernel

Q.37 Mahalanobis distance is equivalent to a Euclidean distance that is being computed in a different feature space than the original feature space.

Max. score: 1; Neg. score: 0

[Imp. Note: If you wish to skip the question, you should do so immediately. Once an option is chosen (either true or false), you can not skip the question at a later stage.]

- false
- true

Q.38 PCA cannot be kernelized because we will need to perform an eigen-decomposition of very large sized covariance matrix (infinite x infinite in case of a Gaussian kernel)

Max. score: 1; Neg. score: 0

[Imp. Note: If you wish to skip the question, you should do so immediately. Once an option is chosen (either true or false), you can not skip the question at a later stage.]

- false
- true

Q.39 Which of the following learning problems a binary classification model can be used for (you are allowed to use several such binary classifiers)?

Max. score: 2; Neg. score: 0

- Multi-class classification
- One-class classification
- Multi-label classification
- Dimensionality reduction

Q.40 Which of these learning algorithms do not need labeled data?

Max. score: 2; Neg. score: 0

- Binary SVM
- K-means clustering
- One-class SVM
- Probabilistic PCA

Q.41 Which of these regularization schemes will produce sparse weights?

Max. score: 2; Neg. score: 0

- \(\ell_2\)
- \(\ell_0\)

- [ ] \(\ell_1\)
- [ ] Early stopping

Q.42 (For this question, you may either typeset your answer in LaTeX and upload the PDF, or write your solution on a piece of paper and upload the pictures of the solution, or directly write your solution in the provided text-box below which supports writing equations via LaTeX as well as inserting symbols. However, please use only one of these methods)

Assume we are given an \(N\times M\) matrix \(\mathbf{X}\) (assume each entry to be real-valued) and assume each entry \(X_{nm}\) of this matrix can be approximated by an inner product of two \(K\)-dimensional vectors \(a_n \in \mathbb{R}^K\) and \(b_m \in \mathbb{R}^K\), i.e., \(X_{nm} \approx a_n^\top b_m\). The goal is to estimate the unknowns \(a_n (n=1,2,\ldots,N)\) and \(b_m (m=1,2,\ldots,M)\).

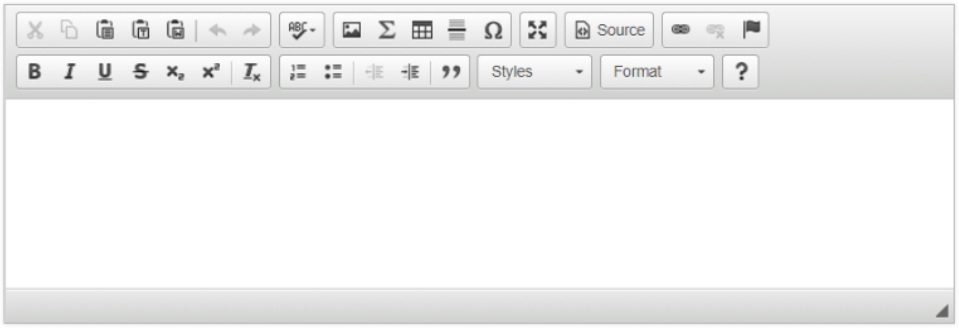
Show that one of the ways to estimate them is by treating the estimation of each \(a_n (n=1,2,\ldots,N)\) and each \(b_m (m=1,2,\ldots,M)\) as solving a linear regression problem (thus a total of \(N+M\) regression problems), and clearly mention what the feature matrix and response vector of each of these regression problems will be. Give a brief outline of the overall algorithm.

Also, which of these \(N+M\) regression problems can be solved in parallel (meaning which of these do not depend on others so that we could solve them in parallel)?

Max. score: 10; Neg. score: 0

This is a long answer type question. You can either upload a file or type your answer below.

or



Q.43 Which of the following is true about the Expectation Maximization (EM) algorithm?

Max. score: 2; Neg. score: 0

- [ ] It is a data clustering algorithm
- [ ] Each of its steps (E and M step) increases the value of the incomplete data log likelihood until it converges.
- [ ] It is a dimensionality reduction algorithm
- [ ] When applied to a Gaussian mixture model, it produces soft clustering for each input

When applied to a Gaussian mixture model, it produces soft clustering for each input

Q.44 For a decision tree learner, low entropy label distributions (after splitting at a node) imply large information gain.

Max. score: 1; Neg. score: 0

[Imp. Note: If you wish to skip the question, you should do so immediately. Once an option is chosen (either true or false), you can not skip the question at a later stage.]

- false
- true

Q.45 For a logistic regression model, assuming \(p\) to be the probability that the label is 1 conditioned on the inputs, the quantity \(\log \frac{p}{1-p}\) is a linear function of the inputs.

Max. score: 1; Neg. score: 0

[Imp. Note: If you wish to skip the question, you should do so immediately. Once an option is chosen (either true or false), you can not skip the question at a later stage.]

- false
- true

Q.46 two random variables X (with its possible values denoted by x) and Y (with its possible values denoted by y) . Which of the following is true?

Max. score: 2; Neg. score: 0

- \(\sum_x p(X=x|Y=y) = 1\)
- \(\sum_x p(X=x,Y=y) = 1\)
- \(\sum_y p(X=x|Y=y) = 1\)
- \(\sum_x \sum_y p(X=x,Y=y) = 1\)

Q.47 For a ridge regression model, a very very small value of the regularization hyperparameter would be expected to give:

Max. score: 2; Neg. score: 0

- Small training error, small test error
- Small training error, large test error
- Large training error, small test error
- Large training error, large test error

Q.48 Briefly describe how/in what ways the Lloyd's algorithm for solving the K-means

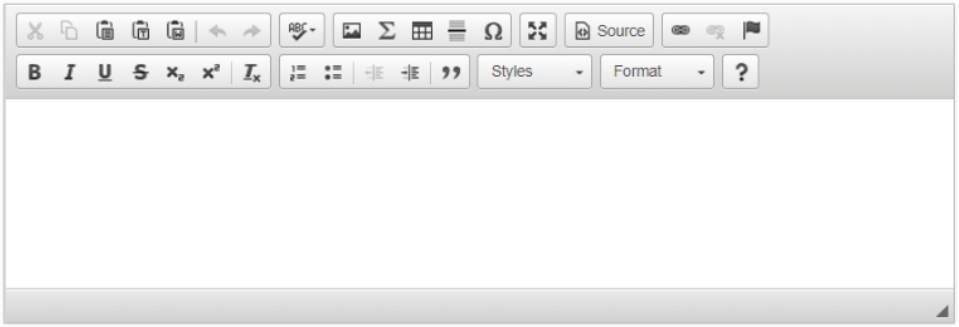
problem is analogous to the EM or ALT-OPT algorithms?

Max. score: 3; Neg. score: 0

This is a long answer type question. You can either upload a file or type your answer below.

UPLOAD A FILE

or



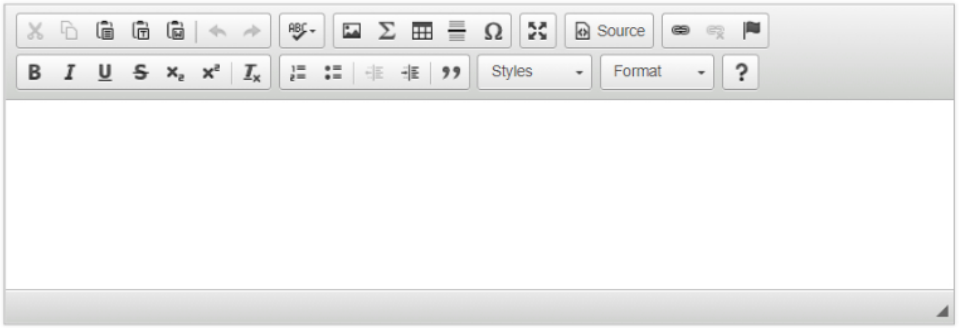
Q.49 What is the difference in the form of the cluster assignment vector \(z_n\) for the following types of clustering algorithms: hard clustering, soft clustering, overlapping clustering ?

Max. score: 3; Neg. score: 0

This is a long answer type question. You can either upload a file or type your answer below.

UPLOAD A FILE

or



Q.50 Which of the following is true about K-means++ clustering?

Max. score: 2; Neg. score: 0

- [ ] Once initialized, it typically converges faster than standard K-means with random initialization
- [ ] The cluster means coincide with K of the inputs in all iterations

The cluster means coincide with K of the inputs in all iterations

The initial cluster means coincide with K of the inputs

It encourages initial cluster means to be close to each other

---

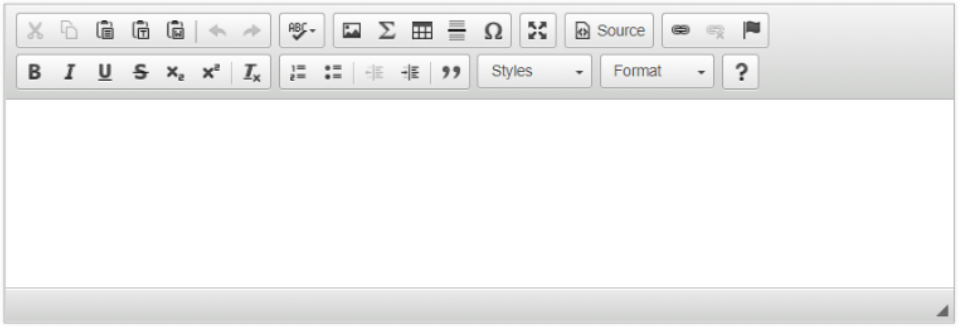
Q.51 Consider a neural network for scalar-valued regression. Assume the network has two hidden layers with connection weight matrices \(W_1\) and \(W_2\) and one output layer with weight vector \(v\). If there is no nonlinear activation after each hidden layer, can the model learn a nonlinear regression? Justify your answer using only text (using at most 2-5 sentences).

Max. score: 3; Neg. score: 0

This is a long answer type question. You can either upload a file or type your answer below.

UPLOAD A FILE

or



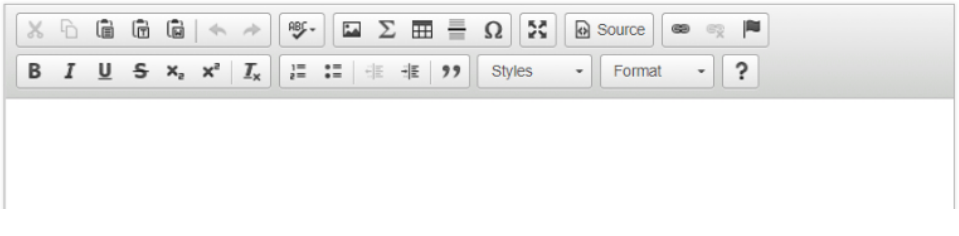
Q.52 The kernel ridge regression weight vector can be obtained as \(w = X^\top \alpha\) where \(X\) is \(N\times D\) feature matrix and \(\alpha = (K+\lambda I_N)^{-1}y\) where \(K\) is the \(N\times N\) kernel matrix and \(y\) is the \(N \times 1\) response vector. Although this model is typically used 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).

Max. score: 3; Neg. score: 0

This is a long answer type question. You can either upload a file or type your answer below.

UPLOAD A FILE

or



SAVE SUBMIT