CS771A: Introduction to Machine Learning
Users Online : 21

Submission Deadline : 15/10/2020 18:00

CS771 Midsem (component 1)

Q.1 Early stopping produces sparse solutions of the weight vector, similar to L1 or L0 regularization.

Please be careful about this selection. Once a selection is made, it cannot be reversed.

- false
- true

Q.2 Maximum Likelihood Estimate (MLE) is obtained by finding the value of the parameter that maximizes its probability (or probability density) given data.

Please be careful about this selection. Once a selection is made, it cannot be reversed.

- false
- true

Q.3 A model that generalizes well on test data should ideally have zero training error.

Please be careful about this selection. Once a selection is made, it cannot be reversed.

- false
- true

Q.4 KNN with K=1 can only learn a linear decision boundary.

Please be careful about this selection. Once a selection is made, it cannot be reversed.

- false
- true

Q.5 Absolute value loss is more robust against outliers than squared loss.

Please be careful about this selection. Once a selection is made, it cannot be reversed.

- false

CS771A: Introduction to Machine Learning
Users Online : 21

Please be careful about this selection. Once a selection is made, it cannot be reversed.

false

true

Q.7 Which of these facts are true for generative classification with Gaussian class conditionals (check all that apply)?

The prediction at test time takes into account the probability density of the test point under each class.

The prediction at test time takes into account the fraction of training examples from each class

The prediction at test time takes into account the shape of each class

If using a uniform class prior, the prediction at test time does not take into account the shape of each class.

Q.8 In multi-label learning or tagging, a data point can simultaneously belong to more than one class.

Please be careful about this selection. Once a selection is made, it cannot be reversed.

false

true

Q.9 In general, a function's optima can never be on a saddle point.

Please be careful about this selection. Once a selection is made, it cannot be reversed.

false

true

Q.10 When doing classification using KNN, is it possible to compute the conditional PMF p(y|x), i.e., the probability of a label given the input? If yes, how? If no, why not? Assume y is binary. Keep your answer short (max 2-3 sentences).

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

UPLOAD A FILE

or

CS771A: Introduction to Machine Learning
Users Online : 21

Q.11 Unlike supervised learning, unsupervised learning does not have the notion of a test set on which the learned model is applied.

Please be careful about this selection. Once a selection is made, it cannot be reversed.

- false
- true

Q.12 Which of these is guaranteed to achieve zero training error if the data is linearly separable (i.e., there exists a hyperplane that can perfectly separate the training data)?

- Decision tree with 3 levels
- Naive Bayes generative classification
- Perceptron
- Learning with Prototypes (LwP)

Q.13 Decision Tree can be used even when some of the features are discrete-valued and some are continuous-valued.

Please be careful about this selection. Once a selection is made, it cannot be reversed.

- false
- true

Q.14 For a 3-class classification problem, consider two models: LwP and KNN with K=3. Which of these will be faster in general at test time and why? Keep your explanation brief (at most 2-5 sentences) and avoid using equations as they are not necessary to answer the question.

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

CS771A: Introduction to Machine Learning
Users Online : 21

Source
Styles
Format

Q.15 Generative classification with Gaussian class-conditionals having equal covariance matrices, and uniform class priors, is equivalent to the standard LwP classifier.

Please be careful about this selection. Once a selection is made, it cannot be reversed.

false

true

Q.16 What is the total number of parameters in a generative 3-class classification model with Gaussian class conditionals if (1) the class prior is uniform, (2) The number of features equals the last two digits of your roll number, and (3) The covariance matrix is diagonal (not spherical) ?

(note: You should count each scalar-valued parameter as one parameter. So if there is a parameter which is a 10-dim vector, it should be counted as 10 parameters)

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

CS771A: Introduction to Machine Learning
Users Online : 21

Alternating Optimization

Lagrange's multipliers method

Projected gradient

Newton's method

Q.18 Briefly (at most 2-3 sentences) explain what is a nonparametric model. Give an example of such a model from among the models we have seen so far in the course.

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

UPLOAD A FILE

or

Q.19 Which of the following is true about the predictive distribution for probabilistic linear regression (assuming Gaussian likelihood and Gaussian prior)?

If using fully Bayesian approach, the predictive distribution has different variances for different test inputs

Assuming known hyperparameters for the likelihood and prior terms, the predictive distribution will NOT be Gaussian

If using fully Bayesian approach, the predictive distribution has the same variance for all test inputs

If using MAP, the predictive distribution has the same variance for all test inputs

If using MLE, the predictive distribution has the same variance for all test inputs

CS771A: Introduction to Machine Learning
Users Online : 21

Please be careful about this selection. Once a selection is made, it cannot be reversed.

false

true

Q.21 Which of the following are true about Decision Trees (DT)?

For the same training set size, DT will be in general faster than nearest neighbors

The leaf nodes of a DT classifier are pure, i.e., they must contain all example having the same label.

Entropy/information gain can be used for deciding the splits for DT classification as well as DT regression.

Increasing the number of levels (that perform more tests) in a DT always improves the generalization performance/

Q.22 Per iteration cost of Newton's method is higher than that of gradient descent.

Please be careful about this selection. Once a selection is made, it cannot be reversed.

false

true

Q.23 For a probabilistic linear regression model with Gaussian likelihood and zero-mean Gaussian prior on the weights, and assuming all hyperparameters to be fixed

MLE is equivalent to ridge regression

MAP estimation corresponds to imposing an L2 regularization on the parameters

The posterior distribution over weights is a Gaussian

The predictive distribution is a Gaussian

Q.24 Consider two random variables $X \sim N(\mu_1,\sigma_1^2)$ and $Y \sim N(\mu_2,\sigma_2^2)$. Define another random variable $Z = X + Y$

$\mathbb{E}[Z] = \mu_1 + \mu_2$ only when X and Y are independent

$\mathbb{E}[Z] = \mu_1 + \mu_2$ regardless of whether X and Y are independent or not

CS771A: Introduction to Machine Learning
Users Online : 21

Q.25 The cost of calculating the gradient of the loss function depends on the size of the training set.

Please be careful about this selection. Once a selection is made, it cannot be reversed.

- false
- true

Q.26 When using a uniform prior distribution, the MAP solution is equivalent to the MLE solution.

Please be careful about this selection. Once a selection is made, it cannot be reversed.

- false
- true

Q.27 For logistic regression with Gaussian prior (assuming fixed hyperparameters for the prior), the posterior will not be Gaussian

Please be careful about this selection. Once a selection is made, it cannot be reversed.

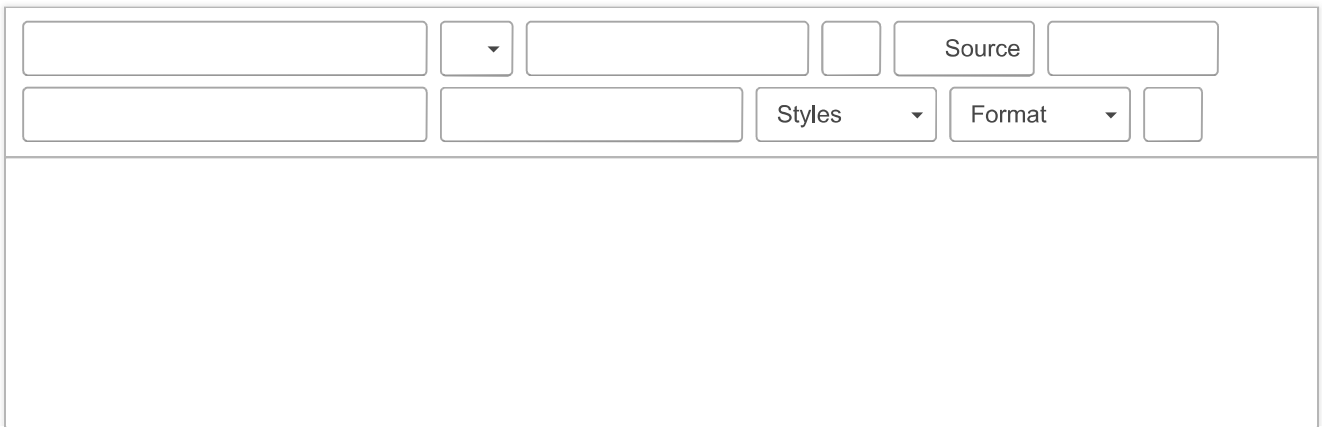
- false
- true

Q.28 LwP struggles when the class shapes are not ideal (e.g., when they are not spherical or equi-spread). Briefly explain (at most 2-3 sentences) how an approach that models each class by a Gaussian can help handle such cases?

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

UPLOAD A FILE

or



CS771A: Introduction to Machine Learning
Users Online : 21

Q.29 To model the bias of a coin, why using a Gaussian as a prior distribution would not be appropriate? Provide a brief answer (at most 1-2 sentences).

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.30 What is the difference between a plug-in predictive distribution and a posterior predictive distribution? Why (and in which situations) would you expect the latter to give better predictions?

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.31 The accuracy of a model on training set is not indicative of the accuracy it might have on

CS771A: Introduction to Machine Learning
Users Online : 21

false

true

Q.32 Briefly explain (at most 2-3 sentences) the significant/meaning/interpretation of hyperparameters of a prior distribution. You may do it using an example of a prior.

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

UPLOAD A FILE

or

Q.33 Which of the following loss functions are convex?

Squared loss for regression

\(\epsilon\)-insensitive loss for regression

0-1 loss for classification

Hinge loss for classification

Q.34 When using a Laplace approximation of the posterior for the logistic regression model, is the predictive distribution analytically computable? If yes, why? If no, why not? Your answer should be in words (no equations) and in at most 2-3 sentences.

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

UPLOAD A FILE

or

CS771A: Introduction to Machine Learning
Users Online : 21

Q.35 When doing gradient ascent for finding the maxima of a function, each iteration takes a step towards the direction of the gradient.

Please be careful about this selection. Once a selection is made, it cannot be reversed.

- false
- true

Q.36 Select all options below that apply for ridge regression.

- It uses absolute loss
- It uses L2 regularizer
- It has a closed form solution
- It is more prone to overfitting that least squares regression

Q.37 A generative classification model must use class prior information.

Please be careful about this selection. Once a selection is made, it cannot be reversed.

- false
- true

Q.38 For a Learning with Prototypes (LwP) classifier, using more training examples per class is, in general, expected to improve accuracy at test time (assuming that the classes are reasonably simple shaped). Briefly answer (max 1-2 sentences) as to why would that be the case.

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

UPLOAD A FILE

CS771A: Introduction to Machine Learning
Users Online : 21

Q.39 Which of these can be used for regression problems?

Perceptron

Decision Trees

KNN

Learning with Prototypes

Q.40 Co-ordinate descent and alternating optimization can be thought of as being equivalent methods. Briefly state (max 2-3 sentences) why in the text box below.

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.41 Logistic regression (with no changes to the input features) learns nonlinear decision boundaries.

CS771A: Introduction to Machine Learning
Users Online : 21

true

Q.42 Gradient descent can be sensitive to the learning rate. Briefly explain (at most 2-5 sentences; avoid equations) how adaptive gradient methods help and give one or two examples of such methods?

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.43 Which of these are discriminative models (i.e., models that do not model the input but only the output)?

- Naive Bayes
- Softmax Regression
- Linear Regression
- Perceptron

Q.44 Which of the following are reasonable ways to assess convergence of a gradient descent algorithm?

- The loss function has a negligible change between successive iterations
- The parameter values have a negligible change between successive iterations
- Gradients become close to zero
- The gradients have a negligible change between successive iterations

CS771A: Introduction to Machine Learning

Users Online : 21