CS771A: Introduction to Machine Learning
Users Online : 190

Quiz 1

Q.1 Which of these statements about gradient descent (GD) are true?

- [ ] If run for sufficiently long, it is guaranteed to find the global minima
- [ ] When optimizing vector-valued variables, it optimizes one element of the vector at a time.
- [x] Every step of GD moves in the opposite direction to the current gradient.
- [x] It is sensitive to initialization

Q.2 Increasing the extent of regularization (e.g., the value of regularization hyperparameter) may not necessarily increase the validation set accuracy

- [ ] false
- [x] true

Q.3 Which of these regularization methods promote sparse solutions?

- [x] L1
- [ ] early stopping
- [x] L0
- [ ] L2

Q.4 L1 norm distance between two vectors a = [3,5,1] and b = [5,2,5] is

- [ ] -3
- [ ] 5
- [x] 9
- [ ] 6

Q.5 For convex functions, Newtons method has the same per-iteration time-cost as gradient descent.

- [x] false

CS771A: Introduction to Machine Learning
Users Online : 190

Q.6 Which of these are convex functions?

- [x] 2*x + 3
- [x] x*x*x
- [x] x*x
- [x] -2*x + 3

Q.7 Assuming binary classification problem with N training examples (assuming we have examples from both classes) and using Euclidean distance, LwP will learn the same decision boundary as KNN if (check all that apply)

- [ ] N is very large
- [ ] Such a thing will never happen
- [x] N=2
- [x] K=1

Q.8 A decision tree can be used for

- [ ] Clustering
- [ ] Dimensionality reduction
- [x] Regression
- [x] Classification

Q.9 At test time, a decision tree with a single decision node (with a single feature's value based or an LwP based splitting criterion) will be faster than a one-nearest neighbors method

- ( ) false
- (x) true

Q.10 Decision trees cannot be used with real-valued features

- (x) false
- ( ) true

CS771A: Introduction to Machine Learning
Users Online : 190

- [ ] Regression
- [x] Multi-class classification
- [x] Binary classification

Q.12 Assuming binary classification and each input to be 10 dimensional, the **minimum** number of parameters (in terms of the number of scalar values) to store an LwP model will be:

(Note: Do not assume that the input features have been transformed/augmented (each will be 10 dimensional)_

- [ ] 10
- [x] 11
- [ ] 22
- [ ] 20

Q.13 For a linear regression model (ignoring the bias term) with N training examples having D features each, the model size will be (in terms of the total number of scalars)

- [ ] N
- [ ] constant (independent of N and D)
- [x] D
- [ ] N*D

Q.14 Which of the following is true about the absolute value function f(x) = |x|

- [x] It has infinite many subgradients in its subdifferential set at x = 0
- [x] It is a convex function
- [x] It is differentiable everywhere except x = 0
- [ ] It has a very large but finite number of subgradients in its subdifferential set at point x = 0

Q.15 For regression with decision trees, information gain can't be used as a splitting criterion but gini index can be used

CS771A: Introduction to Machine Learning
Users Online : 190

Q.16 For which of these models, the test time cost (time it takes to make a prediction for a test example) will increase if we increase the training set size?

- [ ] LwP
- [ ] Ridge regression
- [ ] Decision tree (assuming a constant prediction rule at the leaf nodes)
- [x] Nearest neighbors

Q.17 For unconstrained problems, gradient descent and projected/proximal gradient descent will give the same solution.

- [ ] false
- [x] true

Q.18 LwP with Mahalanobis distance can learn nonlinear decision boundaries

- [x] false
- [ ] true

Q.19 A linear regression model with L1 norm regularizer will not have a closed form expression for the optimal weight vector.

- [ ] false
- [x] true

Q.20 Which of the following regression loss functions are differentiable everywhere?

- [ ] epsilon-insensitive loss
- [ ] Huber loss
- [x] squared loss
- [ ] absolute loss

Score: 14

CS771A: Introduction to Machine Learning
Users Online : 190