EE514 & CS535 - Machine Learning

Final Examination Spring 2021

Part - 1 (60 pts)

Problem 1. Select ALL (upto three) correct choices. Every incorrect answer would earn a penalty of 1 point but the total marks of any multiple choice question will not be less than zero.

(1) Our regression model is giving large variance error for different training data-sets. Which of the following would you prefer to use to reduce the variance?

(a) Add more data points to each training data-set.

(b) Change the optimization algorithm that is minimizing the convex loss function.

(c) Use model with lower complexity.

(d) Reduce the noise in each training data-set.

(2) We have trained a binary classifier model for linearly separable classes. We obtain a new training data-point that is very far from the decision boundary. We add this new point in the training data-set and retrain the model using weights of the previously trained model as an initial value. For which of the following classifiers, the decision boundary will not change during the retraining?

(a) Perceptron classifier

(b) Hard SVM classifier

(c) Logistic regression classifier

(d) None of the above

(3) Which of the following kernels is/are appropriate for a binary classification problem with $d > n$, that is, we have more number of features than the number of data-points? Hint: We do not prefer to further increase the number of features?

(a) Quadratic kernel

(b) Linear kernel

(c) Polynomial kernel

(d) Radial basis function kernel

(4) While using neural network for classification, we can define the following loss functions:
A: 1/0 error
B:Squared-error
C: cross-entropy error Which of the following is/are true?

(a) We can use any of the loss function.

(b) (A) is preferred to (B) and (C) as the problem under consideration is classification.

(c) A should not be used because A is non-differentiable

(d) C is preferred to B

(5) While carrying out polynomial regression, we are getting high error for both the training and testing data-sets. Which of the following might be used to reduce the error?

(a) Decrease the number of training data points as model is under-fitting

(b) Increase the number of training data points as model is over-fitting

(c) Increase the order of the polynomial

(d) Use less noisy training data

(6) Which of the following is/are true about hard SVM classifier $\mathbf{w}^T\mathbf{x} - \theta = 0$ for $\mathbf{x} \in \mathbf{R}^d$?

(a) Increase or decreasing θ changes the distance of the decision boundary from the origin.

(b) The objective function in the optimization problem we solve to find the weights $\mathbf{w}, \theta$ is convex.

(c) SVM model defines a hyper-plane passing through origin in the $d+1$ dimensional space separating the classes.

(d) Classification margin is $\frac{\theta}{\|\mathbf{w}\|}$

(7) For a binary classification problem with linearly separable classes, which of the following classifiers can enable linear separation of the classes with zero training error assuming training data does not suffer from noise or contain outliers?

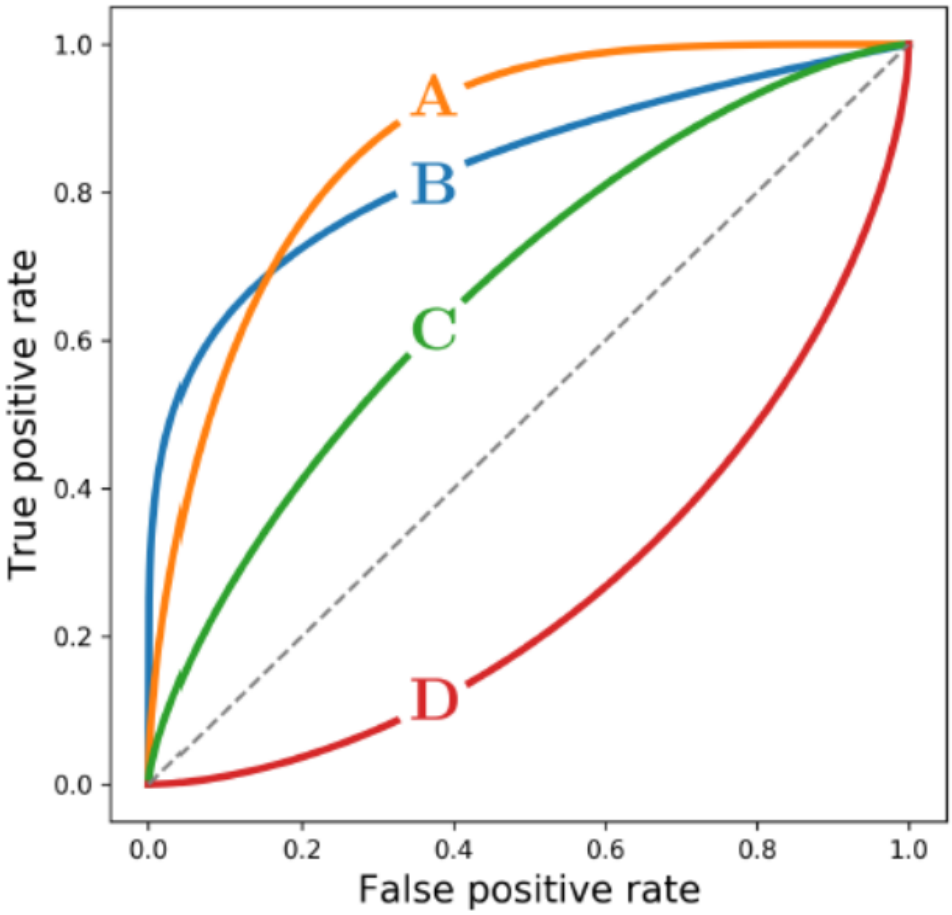
(a) Perceptron classifier

(b) SVM classifier

(c) Logistic regression classifier

(d) None of the above

(8) Given ROC curves (see plot below) for four different binary classifiers, choose the correct statement(s).



(a) Classifier D is the best classifier

(b) Classifier C is worse than the random guess

(c) Classifier B is better than Classifier C

(d) Classifier B is better than Classifier A

(9) Choose the correct statement(s) about hard SVM vs soft SVM for linearly separable training data.

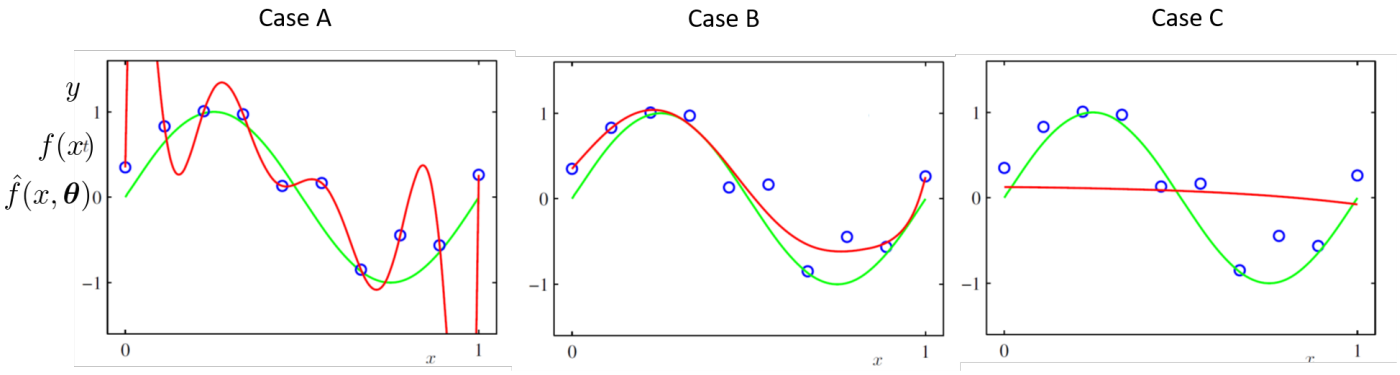
(a) Hard SVM can find a decision boundary with zero loss function, that is, correct classification of every data-point in the training data.

(b) Soft SVM can find a decision boundary with zero loss function, that is, correct classification of every data-point in the training data.

(c) We always have a data-point on either side of the decision boundary of Hard SVM at a distance of half of the classification margin.

(d) We always have a data-point on either side of the decision boundary of Soft SVM at a distance of half of the classification margin.

(10) With reference to the figure shown, we use ridge regression to learn the function $f(x)$ (green) as $\hat{f}(x, \boldsymbol{\theta})$ (red) using the data-points (blue) for different values of regularization parameter, that is, for $\lambda = 0$, 1 and e^{-18}. We want you to associate the value of λ for each case.



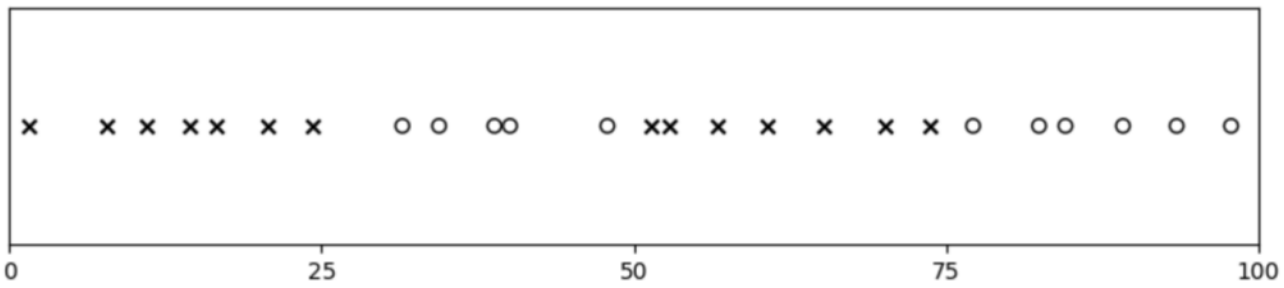
(a) Case A: $\lambda = 0$, Case B: $\lambda = 1$, Case C: e^{-18}

(b) Case A: $\lambda = 0$, Case C: $\lambda = 1$, Case B: e^{-18}

(c) Case B: $\lambda = 0$, Case C: $\lambda = 1$, Case A: e^{-18}

(d) Case C: $\lambda = 0$, Case B: $\lambda = 1$, Case A: e^{-18}

(11) Consider the single dimension classification problem given in figure below. Clearly, the data is not linearly separable. However we can convert it into a linearly separable problem by introducing new feature(s). Which of the following operations can help is in making the data linearly separable?



(a) Add another feature that is -1 for $x > 50$ and 1 for $x \leq 50$

(b) Add a quadratic feature, that is, x^2

(c) Both (a) and (b)

(d) We cannot make the data linearly separable as it is not centered around origin

(12) The distance of a point $\mathbf{x}_i$ from the hyper-plane $\mathbf{w}^T\mathbf{x} = 0$ is given by

(a) $\mathbf{w}^T\mathbf{x}_i$

(b) $\frac{\mathbf{w}^T\mathbf{x}_i}{\|\mathbf{w}\|}$

(c) $\frac{\mathbf{w}^T\mathbf{x}_i}{\|\mathbf{w}\|^2}$

(d) $\sqrt{\frac{\mathbf{w}^T\mathbf{x}_i}{\|\mathbf{w}\|^2}}$

(13) Bias-variance trade-off is controlled or explained by

(a) Regularization coefficient λ in L_2 regression

(b) The learning rate α in the gradient descent

(c) The polynomial degree in polynomial regression

(d) The choice of activation function in neural network

(14) We use non-linear activation functions, instead of linear activation functions, in neural networks to

(a) model non-linear decision boundaries

(b) enable faster computation of gradients and forward pass

(c) restrict the output between 0 and 1

(d) None of the above

(15) We predefine number of clusters in hierarchical clustering.

(a) True

(b) False

(16) While carrying out polynomial regression, which of the following errors you expect to decrease then increase with the degree of polynomial?

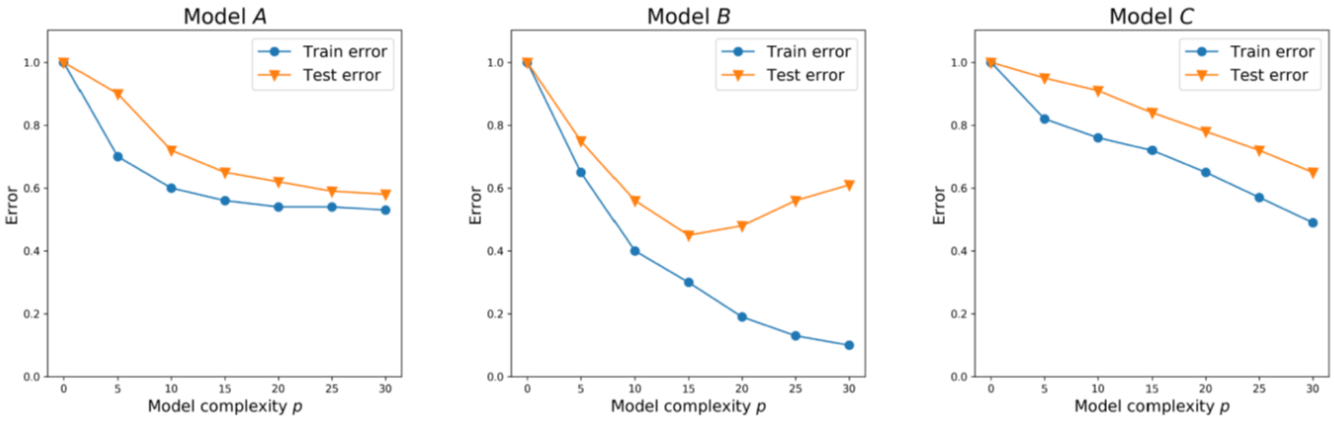
(a) Testing error

(b) Training error

(c) Validation error

(d)

(17) We have plotted below the training and testing error for three models A, B and C against the model complexity in the figure below. Which of the following indicate under-fitting for all values of model complexity p?



(a) A

(b) B

(c) C

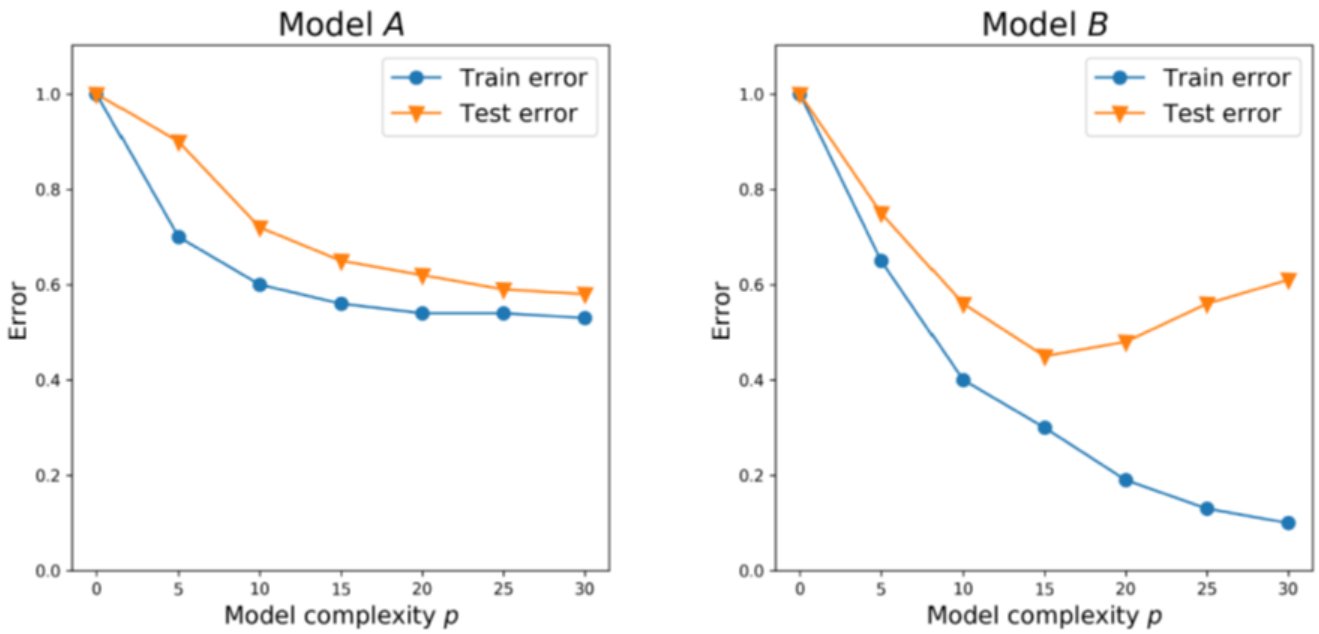
(d) none of the above

(18) We do not need to predefine the number of clusters in K-means clustering

(a) True

(b) False

(19) We have plotted below the training and testing error for two models A and B against the model complexity in the figure below. Which of the following indicate over-fitting?



(a) $p = 20$ for model A
(b) $p = 20$ for model B
(c) $p = 30$ for model A
(d) $p = 30$ for model B

(20) Neural network without hidden layers and sigmoid as an activation function at the output layer is equivalent to perceptron classifier.

(a) True

(b) False

(21) Neural network without hidden layers and sigmoid as an activation function at the output layer is equivalent to logistic regression classifier.

(a) True

(b) False

(22) We are using stochastic gradient descent to minimize the loss function $\mathcal{L}(\mathbf{w})$, where $\mathbf{w}$ denotes the model parameters. If α denotes the learning rate and $\mathcal{L}_i(\mathbf{w})$ denotes the loss function for i-th training input, we carry out the following update in each iteration:

(a) $\mathbf{w} \leftarrow \mathbf{w} - \alpha \nabla \sum_{i=1}^{n} \mathcal{L}_i(\mathbf{w})$, where n is the number of training data points

(b) $\mathbf{w} \leftarrow \mathbf{w} - \alpha \nabla \mathcal{L}_i(\mathbf{w})$

(c) $\mathbf{w} \leftarrow \mathbf{w} + \alpha \nabla \mathcal{L}_i(\mathbf{w})$

(d) $\mathbf{w} \leftarrow \mathbf{w} + \alpha \nabla \sum_{i=1}^{n} \mathcal{L}_i(\mathbf{w})$, where n is the number of training data points

(23) We can represent XOR function using a neural network without any hidden layer.

(a) True

(b) False

(24) In single link agglomerative clustering, we compute the distance between the cluster centroids for grouping of the clusters.

(a) True

(b) False

(25) Increasing C in soft SVM, that is, giving more importance to the misclassifications during minimization of the objective function will

(a) increase the classification margin

(b) increase the norm of $\mathbf{w}$

(c) increase the number of points inside the classification margin

(d) decrease the classification margin

(26) Choose the correct statement(s) about the convexity of loss functions.

(a) Log-loss is a convex function

(b) Squared error in linear regression is a convex function

(c) Hinge loss is a convex function

(d) 1/0 loss is a convex function

(27) Increasing C in soft SVM, that is, giving more importance to the misclassifications during minimization of the objective function will

(a) increase the training error

(b) decrease the training error

(c) increase the classification margin

(d) decrease the classification margin

(28) We interpreted linear regression as ML estimation using the model $y_i = f(\mathbf{x}_i) + n_i$. We made the following assumptions

(a) n_i follows Gaussian distribution

(b) the mean is same for all y_i

(c) the variance is same for all y_i

(d) the mean is same for all n_i

(29) Following is/are the practical issues with the use of sigmoid function as an activation function

(a) Derivative of the sigmoid becomes zero when pre-activation output reaches maximum or minimum value

(b) Sigmoid is a convex function

(c) Sigmoid output is non=zero centered

(d) Computationally expensive

(30) Which of the following classifier models can be used if the classes are not linearly separable?

(a) Perceptron classifier

(b) Neural network with linear activation function used for every nueron

(c) Soft SVM classifier

(d) Hard SVM classifier

(31) Which of the following is/are true about generative models?

(a) Perceptron classifier is the simplest generative model

(b) Perceptron classifier is a generative model

(c) We can use generative model for classification

(d) Generative models model the joint probability of the input and output

(32) In soft SVM we introduce slack variable ξ_i associated with each input $\mathbf{x}_i$. Slack variable ξ_i is non-zero for

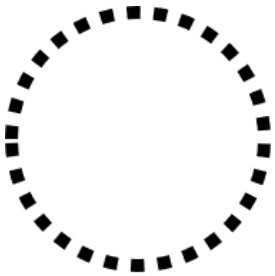
(a) misclassified $\mathbf{x}_i$

(b) correctly classified $\mathbf{x}_i$

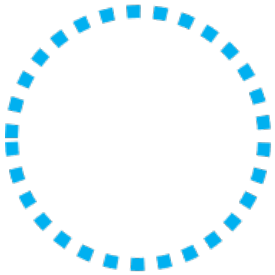
(c) $\mathbf{x}_i$ inside the margin

(d) $\mathbf{x}_i$ outside the margin

(33) With reference to the figure shown, we wish to cluster the given data (without labels) in the form of two clusters indicated in (b). Which of the following clustering algorithms can be employed to obtain the desired clusters.



(a) Data without labels



(b) Desired Clusters

(a) K-means for $K = 2$

(b) Agglomerative clustering with complete linkage and using Euclidean distance

(c) Agglomerative clustering with single linkage and using Euclidean distance

(d) Agglomerative clustering with complete linkage and using Manhattan distance

(34) Which of the following are true for ReLu vs sigmoid as activation functions?

(a) Sigmoid is more computationally expensive

(b) Both ReLu and sigmoid are monotonically non-decreasing

(c) Derivative of sigmoid is quadratic, that is, it requires the computation of square of sigmoid

(d) ReLu does not suffer from vanishing gradient problem

(35) In logistic regression, we

(a) minimize log-loss (cross entropy)

(b) obtain an estimate of the posterior probabilities at the output

(c) have a generative model

(d) represent the log-odds as a linear function

(36) For 2 inputs (only two features), 10 outputs classification problem, we use a neural network with one hidden layer of 20 neurons. Choose the correct statement(s).

(a) We use softmax function as an activation function at the output layer

(b) The use of linear function as an activation function at the hidden layer is an appropriate choice

(c) There are 270 parameters (weights+biases) defining a neural network

(d) There are 240 parameters (weights+biases) defining a neural network

(37) Which of the following statements is/are true for hierarchical clustering?

(a) Complete linkage is less sensitive to outliers than single linkage

(b) We can only use Euclidean distance for computing distance between clusters

(c) Nested clusters can also be visualized using dendrogram

(d) The number of clusters is a hyper parameter