COMP-551-001 Applied Machine Learning
Self-Assessment Questions

Sarath Chandar

Lecture-01

1. In the class, we have seen that both classification and regression are supervised learning problems. What is the differece between these two supervised learning problems?
2. Give five real life applications for classification and five real life applications for regression.
3. When would you say that a particular model is a linear model?
4. What is the significance of w_0 (bias) in the linear model: $y = w_0 + w_1 x$? Why not use a model like $y = w_1 x$?

Lecture-02

1. What is overfitting? How can we find if a model is overfitting to a particular dataset?
2. Suggest at least 3 approaches to solve the overfitting problem.
3. What is the hyperparameter of a model? How is it different from the parameters of the model? How can we choose these hyperparameters?
4. While doing model selection, we choose the best hyperparameter based on the validation set performance. What will happen if we choose the best hyperparameter based on training set performance? What will happen if we choose the best hyperparameter based on the test set performance? Do we really need a separate validation set?
5. What are the hyper-parameters of the linear regression model? What are the hyper-parameters of the k-NN classifier?

Lecture-03

1. Compare and contrast least squares approach and nearest neighbor approach in terms of bias and variance.
2. In the class, we have seen that if we use squared error loss, then the expected prediction error is minimized by the conditional mean. Explain how nearest neighbor approach and least squares approach are trying to approximate this conditional mean.
3. What is Bayes rate? We have seen in the class that Bayes rate is the best possible performance any classifier can achieve. What does a classifier require in order to achieve this optimal error rate?

Lecture-04

1. What is the advantage of using non-linear basis functions with a linear model like linear regression?
2. What is the pseudo-inverse of a matrix? How is it different from the inverse of a matrix? When will the psuedo-inverse and inverse be equivalent?
3. Explain the geometrical interpretation of least squares approach.
4. When can one resort to gradient descent to minimize the objective function?
5. What happens when the step size is too large in gradient descent? What happens when the step size is too small?
6. What is the difference between gradient descent and stochastic gradient descent?
7. Gradient descent can always find the global minimum. True or False? If false, is there any scenario when it is guarenteed to find the global minumum?

Lecture-05

1. What is inductive bias? What is the inductive bias of linear regression and nearest neighbor algorithms?
2. A hypothesis which minimizes the empirical risk is also guarenteed to minimize the true risk. True or False?
3. Define bias and variance. Explain the bias-variance tradeoff.
4. Define Occam's razor.
5. Adding regularization controls overfitting. True or False?
6. Can we use L1 regularization and L2 regularization for feature selection? If so, explain how will you do that. Will there be any difference in the feature selection procedure based on whether the reuglarizer is L1 or L2 regurlarizer?
7. L1 regularization prefers sparse models. Justify.
8. Compare the geometrical views of L1 regularization and L2 regularization and argue why L1 regularizer sets more weights to zero than L2 regularizer.

Lecture-06

1. What are the three approaches to solving classification problem? Sort them in ascending order of procedure complexity.
2. Why are generative models called as *generative* models?
3. What are linearly separable problems? Give cartoon examples for linearly separable 2-class problem and not linearly separable 2-class problem.
4. In a linear discriminant model, decision surface is perpendicular to the weight vector. Prove.
5. Explain the difference between one-vs-rest classifier and one-vs-one classifier.
6. Explain various ways of solving multi-class classification problem using discriminant functions.

Lecture-07

1. Least-squares solution lacks robustness to outliers when used for classification. Justify.
2. List down the applications of PCA.
3. Why should we constrain the norm of the projection vector in PCA to 1?
4. PCA and LDA project data from one space to another space. How can we use such algorithms for classification? Which projection will be more helpful to design a classifier and why?
5. Which one of the following projection algorithms is supervised? PCA or LDA?

Lecture-08

1. What is the difference between a linear model and a generalized linear model?
2. In GDA, covariance matrix of all the class conditional densities are shared. How is this affecting the decision boundary?
3. Explain the i.i.d assumption.
4. What is the difference between GDA and QDA?
5. Define confusion matrix. How will an ideal matrix look like?
6. Define precision and recall. List two applications where precision is more important and two applications where recall is more important.
7. Explain the tradeoff between precision and recall.
8. Define F1-measure. What is the advantage of using F1-measure as an evaluation metric?

Lecture-09

1. Compare GDA and QDA in terms of parameter complexity.
2. Explain the naive Bayes assumption.
3. Gaussian Naive Bayes has linear decision boundary. True or False?
4. What is Laplace smoothing? Why do we need to smooth our Naive Bayes estimates?
5. Laplace smoothing is a biased smoothing. Justify.
6. What are the advantages of discriminative approach over generative approach for classification?
7. Explain the relationship between maximum likelihood and least squares.

Lecture-10

1. Is there a closed form solution for logistic regression? If not, why?
2. What is the difference between gradient descent and Newton-Raphson method?
3. Derive Newton-Raphson update rule for least squares problem.
4. Explain the geometric view of gradient descent and Newton-Raphson method.

5. Prove that in the absence of regularization, the maximum likelihood training for logistic regression can exhibit severe overfitting for datasets that are linearly separable.
6. What are the advantages of generative models over discriminative models?
7. What are the advantages of discriminative model over generative model?
8. What are the advantages of discriminative models over discriminant based models?
9. What is perceptron error criteria?
10. What are the issues with the perceptron algorithm?

Lecture-11

1. Define margin of a decision boundary.
2. What are active constraints and inactive constraints in a constrained optimization problem?
3. Argue that there will be at least one active constraint in the max-margin problem and that there will be at least two active constraints when the margin is maximized.
4. In SVMs, when will you solve the primal problem and when will you solve the dual problem?
5. What is the error function that max-margin classifier is trying to minimize?
6. What is the motivation for introducing slack variables in the max-margin optimization problem?

Lecture-12

1. Compare the characteristics of squarred error, cross-entropy loss, and hinge loss. How are they trying to approximate the misclassification error? Would you prefer one loss over the other? If so, why?
2. What are the differences between parametric methods and non-parametric methods? Give examples for both methods.
3. Why is K-NN called a lazy classifier?
4. What is the advantage of using basis functions?
5. What are the various ways of splitting continuous valued attributes in a decision tree?
6. What are the different measures of node impurity?
7. What is GINI index? How would you compute GINI index for a continuous valued attribute?
8. What is the disadvantage of using entropy as a criteria for attribute test selection in a decision tree? How can you overcome it?
9. When can a decision tree overfit? What are the various ways to avoid overfitting in decision trees?
10. Explain reduced error pruning and rule post pruning. What are the advantages of rule post pruning over reduced error pruning?
11. What are the advantages of a decision tree classifier?

Lecture-13

1. What is a mixture of expert?
2. Explain the procedure for bootstrapping datasets. Why is it useful?
3. Prove that bagging reduces variance when we bag classifiers whose errors are uncorrelated.
4. What is the difference between bagging decision trees and random forests? Which one would you prefer and why?
5. What are the differences between bagging and boosting?
6. What is the error function that AdaBoost is trying to minimize?

Lecture-14

1. What is the optimal solution for exponential error function? How is AdaBoost trying to approximate it?
2. What is k-fold cross validation? What are the advantages and disadvantages of doing k-fold cross validation?
3. Explain stacking.
4. What is representation learning?
5. Can we add multiple linear layers in a neural network? If not, why?
6. What type of output activation function will you use for the following scenarios:
    (a) When the output is a regression target.
    (b) When the output is a probability value.
    (c) When the output is a probability distribution.
7. State universal approximation theorem.
8. What is the advantage of adding more hidden layers in a neural network?

Lecture-15

1. Backpropagation is an efficient way of implementing chain rule by using dynamic programming. True or False?
2. Explain the feed-forward stage and backpropagation stage in training a neural network.
3. Explain how to compute gradient for the following cases using B-diagrams:
    (a) function composition.
    (b) function addition.
    (c) weighted edges.
4. Backpropagation algorithm computes the derivative of the network function F w.r.t the input x correctly. Prove.
5. Derive backpropagation for a multi-layered feed-forward network with
    (a) sigmoidal hidden layers and softmax output layer.
    (b) tanh hidden layers and sigmoidal output layer.

Lecture-16

1. What is vanishing gradient problem? By using chain rule, explain why gradients shrink as the depth of the network increases.

2. What is an auto-encoder? How can we use auto-encoder for dimensionality reduction? How is it related to PCA?

3. Explain greedy layer-wise training procedure for training deep neural networks. How is it solving the vanishing gradient problem?

4. Define ReLU activation function. What is the issue with ReLU activation? Explain how leaky ReLU solves that issue?

5. Explain Batch Normalization procedure. How is it helpful in solving vanishing gradient problem? How will you use batch normalization during prediction time when you have only a single test instance?

Lecture-17

1. Explain the advantage of CNNs over feed-forward neural networks for image data.

2. What are the hyper-parameters of a convolutional layer? Explain the effect of each hyperparameter in terms of parameter complexity (how it affects the number of parameters required) and output complexity (how it affects the output volume).

3. What is zero-padding? Why should we do zero-padding?

4. What is pooling? What is the advantage of using a pooling layer? Can we substitute a pooling layer with a convolutional layer to get the same effect? If so, how?

5. What is a sequential problem? Give at least 3 examples for sequential problem.

6. Explain how can use use a feed-forward network to solve a sequential problem. Also explain the limitations of such an approach to solve the sequential problem.

Lecture-18

1. Explain how to unroll RNNs over time. What is the advantage of unrolling an RNN?

2. Explain back-propagation through time (BPTT). What is the advantage of doing truncated BPTT over BPTT?

3. Explain exploding gradient and vanishing gradient issues in training an RNN.

4. Describe the LSTM architecture. Explain how it solves the vanishing gradient problem.

5. What is the difference between the cell state and the hidden state in an LSTM?

6. Compare the advantages and disadvantage of

   (a) batch gradient descent.

   (b) stochastic gradient descent.

   (c) mini-batch gradient descent.

7. Explain gradient descent with momentum. Why does momentum help in faster convergence?

8. What is the difference between regular momentum method and Nesterov's accelerated gradient method?

9. Explain Adagrad method. What are the issues in using Adagrad for training and how will you resolve them?

Lecture-19

1. What is the difference between frequentist approach and Bayesian approach for machine learning?
2. What is the difference between an estimation problem and a prediction problem?
3. Explain how will you estimate the parameters when you use
   (a) Maximum Likelihood Estimation (MLE).
   (b) Maximum A Posteriori (MAP) Estimation.
   (c) Bayesian Estimation.
4. What are the advantages of using prior over parameters?
5. Under what circumstances would MAP reduce to MLE?
6. What is a conjugate prior? What are the advantages of using a conjugate prior.
7. Prove that Beta distribution is a conjugate prior for Bernoulli distribution.
8. What are the advantages of using Bayesian estimation?

Lecture-20

1. What are the disadvantages of doing cross-validation for choosing model complexity?
2. Prove that MAP estimate for linear regression, when using Gaussian prior over the parameters, is equivalent to the ridge regression solution.
3. What is the limitation of Bayesian Linear Regression while using Gaussian basis functions?
4. The localization property of the equivalent kernel emerges because of the Gaussian basis function. True or False?

Lecture-21

1. What is the difference between parametric methods and memory-based methods?
2. Explain kernel trick.
3. Derive the dual representation for ridge regression. Show that the solution for the dual problem uses entire training data to make prediction for new data point by using kernels.
4. What is the advantage of using non-linear kernels over using non-linear basis functions?
5. What is a Gaussian process?
6. Derive Gaussian process for regression.

Lecture-22

1. Define clustering. How is it different from supervised learning problems like classification?
2. List at least 3 applications of clustering.
3. What is the difference between hierarchical clustering and partitional clustering?
4. K-Means clustering is a partitional clustering algorithm. True or false?

5. What are the hyper-parameters in K-Means clustering algorithm? Do we need a separate validation set to select the best hyper-parameters? If not, why?
6. What are the limitations of K-Means algorithm?
7. Explain K-Means++. What limitation of K-Means is K-Means++ trying to solve?
8. Explain how can you perform hierarchical clustering by using bisecting K-means algorithm?
9. What are the hyper-parameters of DBSCAN algorithm?
10. What is the difference between K-Means and DBSCAN?
11. Explain at least three evaluation metrics for clustering.
12. Define cluster cohesion and cluster separation.