Unit-1 (19 questions — Introduction, Linear Algebra, Stat. Learning, Eval, Optimization)

1. The dot product of vectors a = (1,2,3) and b = (4,0,-1) is:

- A) 3
- **B) 1**
- C) -3
- D) 11

2. Which matrix operation is NOT generally commutative?

- A) Addition
- B) Scalar multiplication
- **C) Matrix multiplication**
- D) Transpose

3. Empirical risk minimization minimizes:

- A) True risk
- **B) Training loss**
- C) VC dimension
- D) Model complexity

4. Which measure captures balance between precision and recall?

- A) Accuracy
- **B) F1-score**
- C) ROC AUC
- D) Mean squared error

5. K-fold cross-validation primarily helps to:

- A) Reduce model bias only
- **B) Estimate model generalization and reduce variance of estimate**
- C) Increase training set size
- D) Compute posterior probabilities

6. Inductive bias is:

- A) The randomness in dataset
- **B) Assumptions used by a learning algorithm to generalize**
- C) A type of regularization
- D) A visualization technique

7. Which is a convex optimization technique commonly used in ML?

- A) Newton’s method for nonconvex only
- B) Genetic algorithms
- **C) Gradient descent**
- D) Simulated annealing

8. Norm of vector v = (3,4) (Euclidean) is:

- A) 7
- **B) 5**
- C) √13
- D) 25

9. Which statement about hypothesis space H is true?

- **A) H is the set of models an algorithm can choose from**
- B) H refers only to linear models
- C) H is always finite
- D) H equals training data

10. Statistical learning theory primarily studies:

- A) Neural network architectures
- **B) Generalization of learning algorithms**
- C) Data visualization methods
- D) Hardware optimization

11. Which is NOT a type of learning?

- A) Supervised
- B) Unsupervised
- C) Reinforcement
- **D) Descriptive (as a standard ML category)**

12. Which quantity directly measures overfitting?

- A) Low training error and low test error
- B) High test error only
- **C) Low training error and high test error**
- D) High training error and low test error

13. The identity matrix I has which property?

- A) Determinant zero
- **B) I · A = A for any conformable matrix A**
- C) It is always singular
- D) Trace equals zero

14. Which is true for orthogonal vectors?

- A) They are linearly dependent
- **B) Their dot product is zero**
- C) They must have same magnitude
- D) They must be parallel

15. Mini-batch SGD:

- A) Uses entire dataset per update
- **B) Uses subsets (batches) for each gradient update**
- C) Always converges faster than full-batch GD in theory
- D) Is only used for convex problems

16. Which loss is commonly used for regression?

- **A) Mean Squared Error (MSE)**
- B) Cross-entropy loss
- C) Hinge loss
- D) KL divergence

17. Which is a property of convex functions?

- A) Every local minimum is not global
- **B) Every local minimum is global**
- C) They have multiple disconnected minima
- D) They are always linear

18. The rank of a matrix is:

- A) Number of columns only
- **B) Maximum number of linearly independent rows or columns**
- C) Minimum eigenvalue
- D) Determinant sign

19. Which optimization method adapts learning rate per parameter using moments?

- A) SGD with fixed lr
- B) RMSProp without momentum
- **C) Adam**
- D) Simulated annealing

Unit-2 (19 questions — Bayesian learning, Linear/Ridge/Lasso, PCA, PLS)

20. Maximum Likelihood (ML) estimate maximizes:

- A) Posterior probability
- **B) Likelihood P(Data | θ)**
- C) Prior P(θ)
- D) Evidence P(Data)

21. MAP estimate differs from ML by incorporating:

- A) Likelihood only
- **B) Prior over parameters**
- C) Test set performance
- D) VC dimension

22. A conjugate prior is chosen because:

- A) It minimizes loss
- **B) Posterior is same family as prior making analytic update easier**
- C) It is always noninformative
- D) It maximizes evidence

23. Ridge regression uses which penalty?

- A) L1
- **B) L2**
- C) Elastic net only
- D) No penalty

24. Lasso regression primarily:

- A) Increases variance
- **B) Encourages sparsity in weights (feature selection)**
- C) Is identical to Ridge
- D) Removes bias entirely

25. For a linear model y = Xw + ε, normal equations solve w by:

- A) Inverting X only
- **B) $(X^T X)w = X^T y$**
- C) Using eigenvectors only
- D) Gradient-free search

26. PCA principal components are eigenvectors of:

- A) Covariance matrix or correlation matrix
- **B) Covariance matrix (or standardized correlation matrix)**
- C) Gram matrix only
- D) Identity matrix

27. PCA reduces dimensionality by:

- A) Minimizing classification error directly
- **B) Preserving directions of maximum variance**
- C) Maximizing entropy
- D) Clustering data

28. Partial Least Squares (PLS) differs from PCA by:

- A) Ignoring y
- **B) Considering response y while extracting components**
- C) Using only covariance of X
- D) Being unsupervised

29. Which prior is conjugate to Gaussian likelihood with known variance?

- A) Beta prior
- **B) Gaussian prior**
- C) Dirichlet prior
- D) Uniform only

30. Which statement about bias and variance under L2 regularization is true?

- A) Both bias and variance increase
- **B) Regularization increases bias and usually reduces variance**
- C) Regularization decreases bias only
- D) Regularization only affects test set

31. In Bayesian estimation, the evidence term P(D) is used for:

- A) Computing MAP only
- **B) Model comparison (marginal likelihood)**
- C) Regularization strength
- D) Gradient descent

32. The design matrix X has size n×p (n samples, p features). If p >> n, which is true?

- A) $(X^T X)$ is always invertible
- **B) $(X^T X)$ may be singular; regularization is helpful**
- C) Overfitting cannot occur
- D) PCA is unnecessary

33. Ridge regression closed-form solution:

- A) $(X^T X)^{-1} X^T y$
- **B) $(X^T X + \lambda I)^{-1} X^T y$**
- C) $X^T X + \lambda y$
- D) $\lambda I X$

34. Lasso’s L1 penalty creates difficulty because:

- A) It is smooth everywhere
- **B) L1 is non-differentiable at zero causing analytic solution difficulty**
- C) It always increases coefficients
- D) It needs no hyperparameter

35. Which of these is an unsupervised dimensionality reduction?

- **A) PCA**
- B) Lasso regression
- C) Ridge regression
- D) Logistic regression

36. In Bayesian linear regression with Gaussian prior and likelihood, the posterior is:

- A) Uniform
- **B) Gaussian**
- C) Beta
- D) Poisson

37. Which method directly addresses multicollinearity?

- A) LDA
- **B) Ridge regression**
- C) KNN
- D) Naive Bayes

38. Selecting the number of PCA components often uses:

- A) Random choice
- **B) Explained variance threshold (e.g., 95%)**
- C) Number of classes only
- D) Cross-entropy

39. PLS is especially useful when:

- A) Features are independent
- **B) Predictors are many and highly collinear**
- C) No response variable exists
- D) Data is categorical only

Unit-3 (21 questions — Linear/Logistic/LDA/QDA/Perceptron/SVM/NN/Decision Trees/Naive Bayes)

40. Logistic regression outputs:

- A) Raw class label only
- **B) Probability (via sigmoid) for binary class**
- C) Cluster ID
- D) Distance to centroid

41. The loss function typically used for logistic regression is:

- A) MSE
- **B) Cross-entropy (logistic) loss**
- C) Hinge loss
- D) L2 loss

42. LDA assumes:

- A) Different covariance for each class
- **B) Equal covariance matrices across classes and Gaussian class-conditional distributions**
- C) Non-parametric distributions
- D) No distributional assumptions

43. QDA differs from LDA by:

- A) Assuming same means only
- **B) Allowing each class to have its own covariance matrix**
- C) Being linear always
- D) Not using priors

44. Perceptron learning converges if:

- A) Data is nonlinearly separable
- **B) Data is linearly separable**
- C) Learning rate is zero
- D) Data has missing values

45. SVM's objective is to:

- A) Minimize training error only
- **B) Maximize margin between classes subject to classification errors**
- C) Minimize variance only
- D) Maximize number of support vectors

46. The kernel trick enables SVMs to:

- A) Run faster always
- **B) Perform implicit mapping to higher-dimensional feature spaces without explicit transform**
- C) Avoid hyperparameters
- D) Be used only for regression

47. Which kernel is radial basis function (RBF)?

- A) Polynomial kernel
- B) Linear kernel
- **C) Gaussian kernel**
- D) Sigmoid kernel

48. In a neural network, backpropagation is used to:

- A) Randomly initialize weights
- **B) Compute gradients of loss w.r.t. weights using chain rule**
- C) Make predictions only
- D) Normalize inputs

49. A decision tree split criterion using information gain uses:

- A) Euclidean distance
- **B) Entropy reduction**
- C) L2 norm
- D) Variance inflation

50. Gini impurity and entropy are used in:

- A) SVM training
- **B) Decision tree node splitting**
- C) PCA component selection
- D) Kernel selection

51. Naive Bayes classifier assumes:

- A) Features are highly dependent
- **B) Conditional independence of features given class**
- C) No prior over classes
- D) Linear decision boundary always

52. Which activation function is non-linear and used widely in hidden layers?

- A) Linear
- B) Step
- **C) ReLU**
- D) Identity

53. Hinge loss is primarily associated with:

- A) Logistic regression
- **B) Support Vector Machines**
- C) K-means
- D) Naive Bayes

54. Softmax function is used for:

- A) Binary regression only
- **B) Multi-class probability outputs in neural networks**
- C) Decision tree pruning
- D) Spectral clustering

55. Early stopping in neural networks helps to:

- A) Increase model size
- **B) Prevent overfitting by stopping training using validation performance**
- C) Ensure perfect training accuracy
- D) Increase learning rate

56. A support vector is:

- A) Any training point in dataset
- **B) Training points that lie on or inside the margin and define the decision boundary**
- C) Test sample only
- D) A kernel function

57. Which method can handle both numerical and categorical features naturally?

- A) SVM without preprocessing
- **B) Decision trees**
- C) PCA
- D) Linear regression without encoding

58. Ensemble of decision trees using random feature subsets per split is called:

- A) AdaBoost
- **B) Random Forest**
- C) KNN ensemble
- D) Bagging with single tree

59. Naive Bayes is especially effective for:

- A) Continuous image pixels without preprocessing
- **B) Text classification (e.g., spam detection)**
- C) Reinforcement learning
- D) Spectral clustering

60. Cross-entropy loss for multi-class compares:

- A) Input features to weights
- **B) True one-hot labels to predicted probabilities**
- C) Variance across classes
- D) Cluster centroids

Unit-4 (12 questions — Hypothesis testing, Ensemble Methods)

61. Null hypothesis (H0) typically represents:

- A) Desired effect
- **B) No effect or status quo**
- C) Alternative claim
- D) Model complexity

62. A p-value is:

- A) Probability H0 is true given data
- **B) Probability of observing data (or more extreme) assuming H0 is true**
- C) Bayesian posterior probability
- D) Confidence interval

63. Type I error corresponds to:

- **A) Rejecting a true null hypothesis (false positive)**
- B) Failing to reject false null (false negative)
- C) Correct acceptance
- D) Power of test

64. Bagging primarily reduces:

- A) Bias only
- **B) Variance by averaging models trained on bootstrap samples**
- C) Complexity of base learners
- D) Need for cross-validation

65. Random Forest differs from bagged trees by:

- A) Using single tree only
- **B) Randomly selecting subset of features at each split**
- C) Not using bootstrap
- D) Using SVMs as base learners

66. AdaBoost adapts by:

- A) Changing model architecture only
- **B) Increasing weights on misclassified examples for next weak learner**
- C) Averaging predictions equally always
- D) Using bagging technique

67. Gradient boosting builds models by:

- A) Training independent models only
- **B) Sequentially fitting models to residuals (negative gradients)**
- C) Random subspace method
- D) PCA on outputs

68. Which ensemble is most robust to noisy labels?

- A) Simple averaging of two models
- **B) Bagging (e.g., Random Forest)**
- C) Single deep decision tree
- D) AdaBoost (often sensitive to noise)

69. Stacking ensemble uses:

- A) Only identical base learners
- **B) A meta-learner trained on base learner predictions**
- C) Bootstrap resampling only
- D) K-means clustering

70. In hypothesis testing, confidence interval width decreases when:

- A) Sample size decreases
- **B) Sample size increases**
- C) Variance increases only
- D) Significance level is lower (e.g., from 0.05 to 0.01)

71. Which boosting method fits additive models using gradient descent in function space?

- A) Bagging
- B) Random Forest
- **C) Gradient Boosting Machines (GBM)**
- D) KNN boosting

72. Which is true about ensemble methods?

- A) They always reduce bias only
- **B) Proper ensembles can reduce error by lowering variance and/or bias depending on method**
- C) Ensembles never improve performance
- D) Ensembles replace need for validation

Unit-5 (12 questions — Clustering: K-means, K-medoids, Density-based, Hierarchical, Spectral)

73. K-means algorithm minimizes:

- A) Sum of absolute distances to medoid
- **B) Sum of squared Euclidean distances to cluster centroids**
- C) Entropy of clusters
- D) Silhouette score

74. K-medoids differs from K-means by:

- A) Using centroids outside dataset always
- **B) Using actual data points (medoids) as centers and being more robust to outliers**
- C) Being faster for large n always
- D) Using PCA internally

75. DBSCAN identifies clusters by:

- A) Fixed number of clusters K
- **B) Density-connectivity using ε (eps) and MinPts**
- C) Using hierarchical merges only
- D) Spectral decomposition

76. Which clustering method can find arbitrary-shaped clusters and ignore noise?

- A) K-means
- B) K-medoids
- **C) DBSCAN**
- D) PCA

77. Agglomerative hierarchical clustering is:

- A) Top-down splitting
- **B) Bottom-up merging**
- C) Always flat clustering
- D) Density-based

78. Single linkage in hierarchical clustering tends to:

- A) Create compact round clusters
- **B) Produce chaining effect (sensitive to noise)**
- C) Require K pre-specified
- D) Use medoids

79. Spectral clustering uses:

- A) KNN classifier internally
- **B) Eigenvectors of graph Laplacian for embedding**
- C) Bayesian priors over clusters
- D) SVM kernel only

80. Silhouette score measures:

- A) Classification accuracy
- **B) How well samples are clustered (cohesion vs separation)**
- C) PCA reconstruction error
- D) Likelihood of GMM

81. Choosing K in K-means can use:

- A) Always K=2
- **B) Elbow method or silhouette analysis**
- C) Cross entropy minimization only
- D) Random selection

82. K-means++ initialization improves K-means by:

- A) Removing need for K
- **B) Spreading initial centroids to improve convergence and final cost**
- C) Using hierarchical initialization only
- D) Making K-means deterministic always

83. Which is an advantage of K-medoids over K-means?

- A) Faster for very large datasets
- **B) Robustness to outliers since medoids are actual observations**
- C) Uses squared distances only
- D) No need to choose K

84. In spectral clustering, the choice of similarity graph affects:

- A) Only runtime
- **B) Final cluster structure significantly**
- C) Whether PCA is used
- D) Choice of distance metric not necessary

Unit-6 (17 questions — EM, GMM, Intro RL, Bayesian Networks, Learning theory)

85. Expectation-Maximization (EM) alternates between:

- A) Random search and gradient step
- **B) E-step (estimate responsibilities) and M-step (maximize parameters)**
- C) Training and pruning
- D) PCA and clustering

86. EM is used to fit:

- A) SVMs only
- **B) Mixture models like GMMs**
- C) Decision trees only
- D) K-means always

87. In Gaussian Mixture Models (GMM), each component is:

- A) A Poisson distribution
- **B) A Gaussian (normal) distribution**
- C) A uniform distribution only
- D) A Beta distribution

88. Responsibilities in EM for GMM mean:

- A) Prior probabilities of components only
- **B) Posterior probabilities that a data point belongs to each component**
- C) Covariance matrices
- D) Learning rates

89. The log-likelihood in EM is guaranteed to:

- A) Decrease every iteration
- **B) Not decrease (non-decreasing) each iteration**
- C) Be constant always
- D) Reach global maximum always

90. In RL, the policy π maps:

- A) Rewards to states
- **B) States to actions (possibly probabilistic mapping)**
- C) Actions to rewards only
- D) States to transition probabilities

91. Q-learning is:

- A) A supervised learning algorithm
- **B) An off-policy temporal-difference RL algorithm that learns action-value function**
- C) A model-based planning technique only
- D) Equivalent to policy gradient

92. The Bellman equation relates:

- A) Only immediate rewards to actions
- **B) Value of a state to immediate reward plus discounted value of successor states**
- C) Clustering distances
- D) Bayes theorem

93. Bayesian networks represent:

- A) Undirected dependencies only
- **B) Directed acyclic graphs (DAG) encoding conditional independencies**
- C) Only linear correlations
- D) Time-series forecasting models only

94. A node in a Bayesian network is conditionally independent of its non-descendants given:

- A) Nothing
- **B) Its parents**
- C) Its children only
- D) The whole graph always

95. VC dimension measures:

- A) Test error directly
- **B) Capacity/complexity of hypothesis class (max shatterable points)**
- C) Training runtime
- D) Number of parameters only

96. Bias-variance tradeoff: high model complexity typically leads to:

- A) High bias, low variance
- **B) Low bias, high variance**
- C) Low bias, low variance always
- D) No change

97. In EM for GMM, which parameters are updated in M-step?

- A) Only responsibilities
- **B) Mixture weights, component means, and covariances**
- C) Number of clusters K
- D) Learning rate only

98. Policy gradient methods in RL optimize:

- A) Value function only
- **B) Policy directly by gradient ascent on expected return**
- C) Transition probabilities
- D) Only Q-values

99. D-separation in Bayesian networks is used to determine:

- A) Number of parents of a node
- **B) Conditional independencies between variables**
- C) Optimal policy in RL
- D) Network capacity

100. No Free Lunch theorem implies:

- A) One algorithm works best for all problems
- **B) No single learning algorithm is universally best over all possible problems**
- C) Deep nets always outperform others
- D) Bayesian methods are superior always
