🧠 Machine Learning Lab (417525) Viva Preparation Guide (Non-Tabular)

This document summarizes the core theory, practical tasks, and high-probability oral questions for the 6 Machine Learning Practicals, structured for quick revision.

💻 Practical 1: PCA for Dimensionality Reduction

Questions & Answers

- **What is PCA and its primary purpose?**
  - PCA (Principal Component Analysis) is an **unsupervised dimensionality reduction** technique that finds a new basis (Principal Components) capturing the **maximum variance** in the data.
- **What is the significance of the covariance matrix?**
  - It is used to calculate the **eigenvalues and eigenvectors**. The eigenvectors represent the directions (components) of maximum variance.
- **Why is Feature Scaling necessary before PCA?**
  - It ensures that features with larger variance do not unfairly **dominate** the calculation of the principal components.
- **How is PCA usually visualized?**
  - Using a scatter plot showing the data projected onto the **first two Principal Components** ($PC1$ and $PC2$).

🚕 Practical 2: Regression (Linear, Lasso, Ridge)

Questions & Answers

- **What is overfitting, and how does regularization help?**
  - Overfitting occurs when a model learns the training data and noise too well. Regularization (Lasso/Ridge) adds a **penalty** to the cost function, **shrinking the weights** and simplifying the model.
- **What is the difference between Lasso (L1) and Ridge (L2)?**
  - **Lasso (L1)** adds an absolute value penalty, forcing some coefficients to **exactly zero (feature selection)**. **Ridge (L2)** adds a squared penalty, forcing weights close to zero but rarely zero.
- **Name the three primary metrics for regression evaluation.**
  - **MAE** (Mean Absolute Error), **MSE** (Mean Squared Error), and **RMSE** (Root Mean Squared Error).

- **What is the role of the Cost Function?**
    - It measures the **error** (or "cost") of a model's prediction. The goal of training is to **minimize** this function using techniques like gradient descent.

✋ Practical 3: Support Vector Machines (SVM)

Questions & Answers

- **What is the main objective of SVM?**
    - To find the optimal **Hyperplane** that separates the data into classes with the **maximum possible margin**.
- **Define Support Vectors and the Margin.**
    - **Support Vectors** are the data points closest to the hyperplane that determine the boundary's position. The **Margin** is the distance between the hyperplane and these support vectors.
- **What is the Kernel Trick?**
    - A method used for **Non-Linear SVM** that implicitly maps input data into a higher-dimensional space where a linear separation (hyperplane) can be found.
- **What is a Hyperplane?**
    - The decision boundary that separates the different classes in the feature space.

🌸 Practical 4: K-Means Clustering & Elbow Method

Questions & Answers

- **What type of learning is K-Means, and how does it work?**
    - It is **unsupervised learning**. It partitions data into K clusters by iteratively assigning points to the nearest **centroid** and recalculating the centroid position.
- **What is the purpose of the Elbow Method?**
    - It is a heuristic used to determine the **optimal number of clusters (K)**.
- **What is WCSS in the Elbow Method?**
    - **Within-Cluster Sum of Squares**. It measures the sum of squared distances between points and their assigned centroids. The "elbow" point marks where adding more clusters yields **diminishing returns** in WCSS reduction.
- **What is the difference between Classification and Clustering?**
    - **Classification** is **supervised** (predicting a label) while **Clustering** is **unsupervised** (grouping similar data points without prior labels).

🌲 Practical 5: Ensemble Learning (Random Forest)

Questions & Answers

- **Why use Random Forest over a single Decision Tree?**
  - It is more **robust**, provides better **accuracy**, and significantly **reduces overfitting** by averaging the predictions of multiple individual trees.
- **Explain Bootstrap Aggregation (Bagging).**
  - It's the process of sampling the training data **with replacement** to create diverse subsets. Each tree in the forest is trained on a different subset.
- **How is Feature Randomness achieved?**
  - At each node split, the algorithm only considers a **random subset of the available features**, ensuring the trees are decorrelated.
- **How is the final prediction made in Random Forest?**
  - For classification, the final output is the **mode** (majority vote) of the predictions from all the individual trees.

🤖 Practical 6: Reinforcement Learning (Q-Learning)

Questions & Answers

- **How does RL differ from supervised/unsupervised learning?**
  - RL learns by interacting with an environment to maximize a cumulative **reward** signal, relying on **trial and error** and **delayed feedback**.
- **What is the Q-Function?**
  - $Q(s, a)$ represents the **expected maximum future reward** for taking action (a) in a given state (s).
- **Define Exploration vs. Exploitation.**
  - **Exploration** is trying new, unknown actions. **Exploitation** is choosing the known action that currently yields the highest reward. The agent must balance the two.
- **What is the Agent-Environment loop?**
  - The continuous cycle: Agent observes **State (s)** → Agent takes **Action (a)** → Environment returns **New State (s')** and a **Reward (r)** → Agent updates its policy.