1 Multiple Choice Questions (MCQs)

1. **Feature Scaling in SVM**
   Scenario: Riya is building an SVM model to classify shopping behavior using age and monthly expenditure. She loads the data and trains the model without scaling the features:

```
from sklearn.svm import SVC
model = SVC(kernel='rbf', gamma='scale')
model.fit(X_train, y_train)
```

   Question: What is the most likely issue Riya will face?

   A) Increased training time
   B) Ineffective visualization
   C) Biased model due to unscaled features
   D) Better accuracy due to raw data

   Answer: **C**
   Explanation: SVMs rely on distance calculations to define margins. Without scaling, features like expenditure (in thousands) dominate over age (in tens), skewing the decision boundary. This bias leads to suboptimal classification. Scaling (e.g., standardization) ensures equal feature contributions, improving the model's accuracy and margin quality.

2. **Gamma's Effect on RBF Kernel**
   Scenario: In an SVM classification task for iris flowers, Dev uses the following kernel:

```
model = SVC(kernel='rbf', gamma=1.0)
```

   Question: What behavior does this choice of gamma encourage?

   A) Smoother decision boundary
   B) Linear decision boundary
   C) Highly localized influence, risk of overfitting
   D) Model ignoring support vectors

Answer: **C**
Explanation: In the RBF kernel, $K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$, a high gamma (e.g., 1.0) causes the similarity to decay rapidly with distance, creating narrow influence zones. This results in a complex, wiggly decision boundary that fits noise, increasing overfitting risk. Lower gamma values produce smoother boundaries.

3. **Adjusting Gamma for Overfitting**
Scenario: Neha observes her SVM classifier creates an overly complex boundary and performs poorly on new data.
Question: What should she adjust?

   A) Increase gamma
   B) Increase C
   C) Decrease gamma
   D) Use linear kernel

Answer: **C**
Explanation: A high gamma in the RBF kernel leads to tight influence zones, causing the model to overfit by capturing noise. Decreasing gamma widens these zones, smoothing the decision boundary and improving generalization to unseen data. Increasing gamma or C (margin penalty) worsens overfitting, and a linear kernel may not suit non-linear data.

4. **Kernel Choice for Non-Linear Data**
Scenario: Two datasets are given: one linearly separable and the other with curved class boundaries. The same SVM model is applied.
Question: What should be changed for better performance on non-linear data?

   A) Lower gamma
   B) Change kernel from linear to RBF
   C) Use fewer features
   D) Use Decision Trees

Answer: **B**
Explanation: A linear kernel assumes classes are separable by a straight line, which fails for curved boundaries. The RBF kernel maps data to a higher-dimensional space where a linear boundary can separate non-linear patterns, improving performance. Lowering gamma or reducing features doesn't address non-linearity, and Decision Trees are a different algorithm.

5. **Gamma Scale Heuristic**
Scenario: You use `gamma='scale'` in your SVM configuration.

Question: How is gamma computed internally?

Hint: Should the gamma value be a constant?

A) $\gamma = 1$
B) $\gamma = 0$
C) $\gamma = \frac{1}{n_{\text{features}} \times \text{Var}(X)}$
D) $\gamma = \text{mean}(X)$

Answer: **C**

Explanation: In scikit-learn, setting `gamma='scale'` means that

$$\gamma = \frac{1}{n_{\text{features}} \times \text{Var}(X)},$$

where n_{features} is the number of features and $\text{Var}(X)$ is the average variance across features. This heuristic adapts gamma based on the data's scale, balancing model complexity. Fixed values like 1 or 0, or using the mean, are incorrect.

6. **Noise and Support Vectors**
Scenario: Aarav adds random noise to his dataset. He notices a spike in the number of support vectors used by the SVM.
Question: Why did this happen?

A) Increased bias
B) Decreased gamma

C) Overfitting due to noise

D) Improved accuracy

Answer: **C**
Explanation: Noise introduces irregularities, causing more data points to lie near or within the margin, increasing the number of support vectors. The SVM tries to fit these noisy points, leading to a complex boundary and overfitting. Bias isn't directly affected, decreasing gamma smooths the boundary, and accuracy typically decreases with noise.

7. **Impact of Unscaled Features**
Scenario: Priya trains an SVM without feature scaling. "Age" varies between 18–60, and "Cost" between 1000–50000.
Question: What metric is most likely impacted?

A) Accuracy

B) Precision

C) Recall

D) Margin quality

Answer: **D**
Explanation: SVMs maximize the margin based on distances. Unscaled features (e.g., large "Cost" vs. small "Age") distort these distances, prioritizing the larger-scale feature and reducing margin quality. This suboptimal margin affects classification performance indirectly, but margin quality is the primary metric impacted. Accuracy, precision, and recall are downstream effects.

8. **Non-Linear Decision Boundary**
Scenario: A researcher plots their SVM decision boundary and finds it nearly linear.
Question: How can they make it more non-linear?

A) Increase gamma

B) Reduce C

C) Decrease gamma

D) Increase number of features

Answer: **A**
Explanation: In the RBF kernel, low gamma creates broad influence zones, approximating a linear boundary. Increasing gamma tightens these zones, allowing the boundary to bend and capture non-linear patterns. Reducing C softens the margin, not non-linearity; decreasing gamma makes the boundary more linear; and adding features doesn't guarantee non-linearity.

2 Numeric Type Questions (NTQs)

1. **Squared Distance Calculation**
Scenario: You need to compute the squared distance between points $A = (2, 4)$ and $B = (4, 6)$ for an SVM kernel.
Question: What is the squared distance?
Answer: **8**
Explanation: The squared distance is calculated as $\|A - B\|^2 = (2 - 4)^2 + (4 - 6)^2 = (-2)^2 + (-2)^2 = 4 + 4 = 8$. This is a key component in the RBF kernel, $K(x_i, x_j) = \exp(-\gamma\|x_i - x_j\|^2)$, determining the similarity between points.

2. **RBF Similarity**
Scenario: You are computing the RBF kernel similarity for two points with a squared distance of 8 and $\gamma = 0.1$.
Question: What is the RBF similarity?
Hint: The RBF kernel is $K = \exp(-\gamma\|x_i - x_j\|^2)$.

Answer: **0.4493**
Explanation: The RBF kernel is $K = \exp(-\gamma\|x_i - x_j\|^2)$. With $\gamma = 0.1$ and squared distance 8, compute $K = \exp(-0.1 \cdot 8) = \exp(-0.8)$. Using $e^{-0.8} \approx 0.4493$, the similarity is 0.4493, indicating moderate similarity due to the distance and gamma.

3. **Gamma Scale Computation**
Scenario: You have a dataset with two features, with variances 10 and 5. You compute gamma using the scale heuristic.
Question: What is the gamma value?
Answer: **0.0667**

Explanation: The scale heuristic is $\gamma = \frac{1}{n_{\text{features}} \cdot \text{Var}(X)}$, where $\text{Var}(X)$ is the average feature variance. For variances 10 and 5, average variance $= \frac{10+5}{2} = 7.5$. With $n_{\text{features}} = 2$, $\gamma = \frac{1}{2 \cdot 7.5} = \frac{1}{15} \approx 0.0667$. This adapts gamma to the data's spread.

3 Multiple Select Questions (MSQs)

1. **SVM Preprocessing Steps**
   Scenario: You are preparing a dataset for SVM classification, including numerical and categorical features.
   Question: Which steps are necessary for SVM preprocessing?

   A) Feature Scaling

   B) One-hot encoding target

   C) Removing missing values

   D) Label encoding categorical features

   Answers: **A, C, D**
   Explanation:
   - A) SVMs require scaled features to ensure equal distance contributions, true.
   - B) The target is typically label-encoded (e.g., 0/1), not one-hot encoded, false.
   - C) Missing values must be handled (e.g., imputed or removed) for SVMs, true.
   - D) Categorical features need numerical encoding (e.g., label or one-hot), true for input features.

2. **High Gamma Issues**
   Scenario: You train an SVM with `gamma=1.0` on noisy data.
   Question: What issues arise?

   A) Overfitting

   B) Smoother margins

   C) Sensitive to outliers

   D) Wide decision boundary

Answers: **A, C**
Explanation:
- A) High gamma creates tight influence zones, fitting noise and overfitting, true.
- B) High gamma produces wiggly, not smooth, margins, false.
- C) Outliers heavily influence the boundary due to localized sensitivity, true.
- D) High gamma narrows the boundary, not widens it, false.

3. **Support Vector Properties**
Scenario: You analyze the support vectors in an SVM classifier.
Question: Which statements are true about support vectors?

A) Lie on margin

B) Affect hyperplane

C) Cannot overlap classes

D) Define decision function

Answers: **A, B, D**
Explanation:
- A) Support vectors lie on the margin or violate it, true.
- B) They determine the hyperplane's position, true.
- C) In soft-margin SVMs, support vectors can be within the margin or misclassified, false.
- D) The decision function depends on support vectors via the kernel, true.

4. **Low Gamma Effects**
Scenario: You set a very low gamma in an RBF SVM.
Question: What happens?

A) Underfitting

B) Broad margins

C) Poor local separation

D) Wiggly boundary

Answers: **A, B, C**
Explanation:
- A) Low gamma creates a near-linear boundary, missing complex patterns, leading to underfitting, true.
- B) Broad influence zones result in wider margins, true.
- C) The model fails to capture local variations, true.
- D) Low gamma smooths the boundary, not makes it wiggly, false.