SVM & Naive bayes

Assignment Questions:

Support Vector Machines (SVMs)

1. What is a Support Vector Machine (SVM)?

- An SVM is a supervised machine learning algorithm used for classification and regression.[1]
- It aims to find the optimal hyperplane that best separates data points of different classes.[2]
- In high-dimensional spaces, SVMs are effective.[3]

2. What is the difference between Hard Margin and Soft Margin SVM?

- **Hard Margin SVM:**
    - Assumes data is perfectly linearly separable.
    - Aims to find a hyperplane with the maximum margin, with no misclassifications.
    - Highly sensitive to outliers.[4]
- **Soft Margin SVM:**
    - Allows for some misclassifications to handle non-linearly separable data or outliers.
    - Introduces a "slack" variable to control the degree of misclassification.[5]
    - More robust in real-world scenarios.

3. What is the mathematical intuition behind SVM?

- The core idea is to maximize the margin between the hyperplane and the closest data points (support vectors).
- This involves solving a constrained optimization problem.
- The hyperplane is defined by a weight vector (w) and a bias term (b).[6]
- The margin is related to the norm of the weight vector (||w||).
- The optimization problem aims to minimize ||w||, while ensuring that all data points are correctly classified (or within the soft margin).

4. What is the role of Lagrange Multipliers in SVM?

- Lagrange multipliers are used to solve the constrained optimization problem in SVM.[7]
- They transform the constrained problem into an unconstrained one, making it easier to solve.
- They help identify the support vectors, as their corresponding Lagrange multipliers are non-zero.

5. What are Support Vectors in SVM?

- Support vectors are the data points that lie closest to the hyperplane.[8]
- They are crucial in defining the hyperplane and the margin.[9]
- Only support vectors influence the position and orientation of the hyperplane.[10]
- If you remove non support vectors, the hyperplane remains the same.

6. What is a Support Vector Classifier (SVC)?

- SVC is the application of SVM for classification tasks.[11]
- It finds the optimal hyperplane to separate data points into different classes.

7. What is a Support Vector Regressor (SVR)?

- SVR is the application of SVM for regression tasks.[12]

- It aims to find a function that approximates the target values within a specified margin of tolerance.

8. What is the Kernel Trick in SVM?

- The kernel trick allows SVMs to handle non-linearly separable data by implicitly mapping the data into a higher-dimensional space.[13]
- It avoids the explicit computation of the mapping, which can be computationally expensive.
- Common kernel functions include linear, polynomial, and radial basis function (RBF) kernels.[14]

9. Compare Linear Kernel, Polynomial Kernel, and RBF Kernel:

- **Linear Kernel:**
    - Suitable for linearly separable data.
    - Simple and efficient.
    - Equivalent to a standard linear classifier.
- **Polynomial Kernel:**
    - Maps data to a higher-dimensional space using polynomial combinations of the original features.[15]
    - Can capture non-linear relationships.
    - Degree of the polynomial is a parameter.
- **RBF Kernel (Radial Basis Function):**
    - Maps data to an infinite-dimensional space.[16]
    - Can capture complex non-linear relationships.[17]
    - Uses a Gaussian function.[18]
    - Gamma is a parameter.[19]
    - Very powerful, and widely used.

10. What is the effect of the C parameter in SVM?

- The C parameter controls the trade-off between maximizing the margin and minimizing the training error (misclassifications).[20]
- A small C value leads to a wider margin but allows more misclassifications (soft margin).[21]
- A large C value leads to a narrower margin but tries to minimize misclassifications (hard margin).[22]
- Higher C values can lead to overfitting.

11. What is the role of the Gamma parameter in RBF Kernel SVM?

- The gamma parameter in the RBF kernel controls the influence of a single training example.[23]
- A small gamma value means a larger radius of influence, making the decision boundary smoother.[24]
- A large gamma value means a smaller radius of influence, making the decision boundary more complex and potentially leading to overfitting.

Naïve Bayes Classifiers

12. What is the Naïve Bayes classifier, and why is it called "Naïve"?

- Naïve Bayes is a probabilistic classifier based on Bayes' theorem.[25]
- It assumes that features are conditionally independent given the class,[26] which is often not true in real-world scenarios.[27]
- It's called "naïve" because of this strong independence assumption.

13. What is Bayes' Theorem?

- Bayes' theorem calculates the conditional probability of an event based on prior knowledge of related events.[28]

- Mathematically: P(A|B) = [P(B|A) * P(A)] / P(B)
    - P(A|B): Posterior probability of A given B.[29]
    - P(B|A): Likelihood of B given A.
    - P(A): Prior probability of A.
    - P(B): Prior probability of B.

14. Explain the differences between Gaussian Naïve Bayes, Multinomial Naïve Bayes, and Bernoulli Naïve Bayes:

- **Gaussian Naïve Bayes:**
    - Assumes features are continuous and follow a Gaussian (normal) distribution.
    - Used for continuous data.
- **Multinomial Naïve Bayes:**
    - Assumes features represent counts or frequencies (e.g., word counts in text).[30]
    - Used for discrete data, especially text classification.[31]
- **Bernoulli Naïve Bayes:**
    - Assumes features are binary (e.g., presence or absence of a word).
    - Used for binary data, such as document classification with binary feature vectors.

15. When should you use Gaussian Naïve Bayes over other variants?

- When your features are continuous and you believe they follow a normal distribution.
- Examples:
    - Predicting the length of a flower petal.
    - Analyzing sensor data.

16. What are the key assumptions made by Naïve Bayes?

- **Conditional Independence:** Features are conditionally independent given the class.
- **Prior Distributions:** Prior probabilities of classes are known or can be estimated.[32]
- **Feature Distributions:** The distribution of features within each class is known or can be estimated.

17. What are the advantages and disadvantages of Naïve Bayes?

- **Advantages:**
  - Simple and easy to implement.
  - Fast training and prediction.
  - Effective with high-dimensional data.[33]
  - Performs well with categorical inputs.
  - Works well with relatively small training data sets.
- **Disadvantages:**
  - Strong independence assumption may not hold.
  - Can suffer from the "zero frequency" problem (Laplace smoothing helps).
  - Less accurate than more complex models in some cases.

18. Why is Naïve Bayes a good choice for text classification?

- It performs well with high-dimensional data (word counts).
- It's relatively fast, which is important for large text datasets.
- Multinomial Naïve Bayes is well-suited for word frequency data.[34]

19. Compare SVM and Naïve Bayes for classification tasks:

- **SVM:**
  - Effective in high-dimensional spaces.[35]

    - Can handle non-linear relationships with the kernel trick.[36]
    - Robust to outliers.
    - More computationally expensive.
- **Naïve Bayes:**
    - Simple and fast.
    - Works well with high-dimensional data.
    - Assumes feature independence.
    - Less accurate than SVM in some cases.
    - Works well with relatively small data sets.

20. How does Laplace Smoothing help in Naïve Bayes?

- Laplace smoothing (or additive smoothing) addresses the "zero frequency" problem.[37]
- If a feature value is not present in the training data for a particular class, its probability becomes zero, which can cause issues.
- Laplace smoothing adds a small constant (usually 1) to all counts, preventing zero probabilities.[38]
- This improves the model's robustness and generalization ability.