Coding Blocks | Online

Machine Learning Online

Assignment – 3: Naive Bayes, Decision Trees, and Ensemble Learning

---

Part-I: Naive Bayes Classifier

Objective:

To understand the probabilistic foundation of Naive Bayes classifiers and apply them to text and numerical datasets.

Assignment Tasks

Task 1: Theory Questions

Answer in 2–4 sentences:

1. What is the core assumption of Naive Bayes?
2. Differentiate between GaussianNB, MultinomialNB, and BernoulliNB.
3. Why is Naive Bayes considered suitable for high-dimensional data?

Task 2: Spam Detection using MultinomialNB

- Load a text dataset (e.g., SMS Spam Collection or any public text dataset).
- Preprocess using `CountVectorizer` or `TfidfVectorizer`.
- Train a **MultinomialNB** classifier.
- Evaluate:
    - Accuracy
    - Precision
    - Recall
    - Confusion Matrix

Task 3: GaussianNB with Iris or Wine Dataset

- Train a **GaussianNB** classifier on a numeric dataset.
- Split data into train/test sets.
- Evaluate model performance.
- Compare with Logistic Regression or Decision Tree briefly.

Part-II: Decision Trees

Objective:

To implement Decision Tree classifiers and understand their structure, splits, and overfitting characteristics.

Assignment Tasks

Task 4: Conceptual Questions

Answer briefly:

1. What is entropy and information gain?
2. Explain the difference between Gini Index and Entropy.
3. How can a decision tree overfit? How can this be avoided?

Task 5: Decision Tree on Titanic Dataset

- Load Titanic dataset (or use a similar dataset).
- Preprocess (handle missing values, encode categorical variables).
- Train a **DecisionTreeClassifier**.
- Visualize the decision tree using `plot_tree`.
- Evaluate the model using accuracy and confusion matrix.

Task 6: Model Tuning

- Use parameters like:
  - `max_depth`
  - `min_samples_split`
- Show how performance changes.
- Plot training vs testing accuracy to visualize overfitting.

Part-III: Ensemble Learning – Bagging, Boosting, Random Forest

Objective:

To understand and implement ensemble techniques like Random Forest, AdaBoost, and Gradient Boosting for improving classification performance.

Assignment Tasks

Task 7: Conceptual Questions

Answer:

1. What is the difference between Bagging and Boosting?
2. How does Random Forest reduce variance?
3. What is the weakness of boosting-based methods?

Task 8: Random Forest vs Decision Tree

- Train a **RandomForestClassifier** on the same dataset used for Task 5.
- Compare accuracy, precision, and recall with the standalone decision tree.
- Plot feature importances.

Task 9: AdaBoost or Gradient Boosting

- Train an **AdaBoostClassifier** or **GradientBoostingClassifier**.
- Use a suitable dataset.
- Compare it with Random Forest and Decision Tree in terms of:
    - Accuracy
    - F1-score
    - Training time (optional)

Submission Guidelines

- Submit:
  - Jupyter Notebook
  - All CSV datasets used or links to source
  - README with execution steps
- Use proper comments and section headers
- Upload it on Github.

Coding Blocks | Online