Shahid Beheshti University
Machine Learning
M.Sc - Fall 2024

Assignment 3

1 Theoretical Questions

1. **Exercise 1:** Is it possible for an SVM classifier to provide a confidence score or probability when making predictions for a specific instance? Please explain.

2. **Exercise 2:** What should you do if you have trained an SVM classifier using an RBF kernel and find that it is underfitting the training set? Specifically, should you increase or decrease the values of γ (gamma) or C, or both?

3. **Exercise 3:** What does it mean for a model to be ϵ-insensitive, specifically in the context of ignoring small errors while focusing more on significant losses? How can a simple regression model be made ϵ-insensitive? Please provide some examples of such models

4. **Exercise 4:** How ROC, AUC and F1-score are being used in the evaluation of classification performance? How are they computed? Explain different areas in ROC curve and how the model is performing (reaching a high value fast or slowly, how high does the curve goes, etc.)

5. **Exercise 5:**

   How does the threshold value used in classification (the decision boundary (e.g., a probability like 0.5) used to assign a predicted class based on the model's output scores or probabilities) affect the model's performance? This value specifies a cut-off for an observation to be classified as either 0 or 1. Can you explain the trade-off between false positive and false negative rates, and how the choice of threshold value impacts precision and recall?

6. **Exercise 6:**

   - Plot the RBF kernel $K(x_1, x_2) = \exp\left(-\gamma \|x_1 - x_2\|^2\right)$ for the following cases:

  - Fix $x_1 = 0$ and plot $K(x_1, x_2)$ as a function of $x_2 \in [-5, 5]$. Use $\gamma = 0.5$, $\gamma = 1$, and $\gamma = 5$.
  - Fix $x_1 = (0, 0)$ and plot $K(x_1, x_2)$ in 2D as a heatmap or contour plot, where $x_2 = (x, y)$ and $x, y \in [-3, 3]$. Use $\gamma = 1$.
- Explain how γ affects the "shape" or spread of the kernel. What does this imply about how the kernel measures similarity?

7. **Exercise 7:** How do precision, recall, and F1-score compare in terms of evaluating model performance? Please write their formulas and explain what each metric means, along with what they indicate about the model's performance.

8. **Exercise 8:** Min-Max Scaling, Z-score normalization and Robust scaling normalization are 3 different data normalization techniques commonly used in data analysis. For each technique, please provide a brief explanation, any applicable mathematical formulas, and describe the situations in which each technique is most appropriate.

9. **Exercise 9(Extra Point):** Explain nested cross-validation and 5x2 cross-validation in detail and when we should use them.

2 Practical Exercise

In this part, you are going to work with the **Vehicle Insurance Claim Fraud Detection** dataset. You will implement multiple classification models using the Scikit-Learn package to predict if a claim application is fraudulent or not, based on about 32 features. You are expected:

- Perform exploratory data analysis on the dataset.
- Try to tackle the problem using the following models:
  - Logistic Regression
  - SVM
  - KNN
  - Naive Bayes
  - Tree based
  - Other classifiers: Gradient Boosting, Voting Models, other models (Extra Point)
- Use stratified cross-validation to report your models' performance.
- Check whether this dataset is imbalanced or not, if yes, try some techniques to overcome this issue. (including over-sampling, under-sampling, weight-based approaches, etc.) (consider data leakage in this part and make sure no test data is leaked in the train set)

- Try to boost the performance of the SVM models that you have used in the above section by utilizing various methods (including hyperparameter tuning, different preprocessing methods, feature engineering, etc.). Don't limit yourself only to the aforementioned methods, based on the quality of your work, extra scores may be granted.