LATIHAN FINAL

MATA KULIAH	: Kecerdasan Mesin (IN242)	TANGGAL RILIS	: 7 Juni 2022
SEMESTER	: Genap 2021/2022	WAKTU	: Tengah Malam
DOSEN	: HB & HT	POIN	: 100

Problem 1. What is the fundamental idea behind Support Vector Machines?

Problem 2. What is a support vector?

Problem 3. Why is it important to scale the inputs when using SVMs?

Problem 4. Can an SVM classifier output a confidence score when it classifies an instance? What about a probability?

Problem 5. Say you've trained an SVM classifier with an RBF kernel, but it seems to underfit the training set. Should you increase or decrease γ (`gamma`)? What about `C` ?

Problem 6. What is the approximate depth of a Decision Tree trained (without restrictions) on a training set with one million instances?

Problem 7. Is a node's Gini impurity generally lower or greater than its parent's? Is it generally lower/greater, or always lower/greater?

Problem 8. If a Decision Tree is overfitting the training set, is it a good idea to try decreasing `max_depth`?

Problem 9. If a Decision Tree is underfitting the training set, is it a good idea to try scaling the input features?

Problem 10. If it takes one hour to train a Decision Tree on a training set containing 1 million instances, roughly how much time will it take to train another Decision Tree on a training set containing 10 million instances?

Problem 11. If your training set contains 100,000 instances, will setting `presort=True` speed up training?

Problem 12. If you have trained five different models on the exact same training data, and they all achieve 95% precision, is there any chance that you can combine these models to get better results? If so, how? If not, why?

Problem 13. What is the difference between hard and soft voting classifiers?

Problem 14. Is it possible to speed up training of a bagging ensemble by distributing it across multiple servers? What about pasting ensembles, boosting ensembles, Random Forests, or stacking ensembles?

Problem 15. What is the benefit of out-of-bag evaluation?

Problem 16. What makes Extra-Trees more random than regular Random Forests? How can this extra randomness help? Are Extra-Trees slower or faster than regular Random Forests?

Problem 17. If your AdaBoost ensemble underfits the training data, which hyperparameters should you tweak and how?

Problem 18. If your Gradient Boosting ensemble overfits the training set, should you increase or decrease the learning rate?

Problem 19. What are the main motivations for reducing a dataset's dimensionality? What are the main drawbacks?

Problem 20. What is the curse of dimensionality?

Problem 21. Once a dataset's dimensionality has been reduced, is it possible to reverse the operation? If so, how? If not, why?

Problem 22. Can PCA be used to reduce the dimensionality of a highly nonlinear dataset?

Problem 23. Suppose you perform PCA on a 1,000-dimensional dataset, setting the explained variance ratio to 95%. How many dimensions will the resulting dataset have?

Problem 24. In what cases would you use vanilla PCA, Incremental PCA, Randomized PCA, or Kernel PCA?

Problem 25. How can you evaluate the performance of a dimensionality reduction algorithm on your dataset?

Problem 26. Does it make any sense to chain two different dimensionality reduction algorithms?

Problem 27. How would you define clustering? Can you name a few clustering algorithms?

Problem 28. What are some of the main applications of clustering algorithms?

Problem 29. Describe two techniques to select the right number of clusters when using K-Means.

Problem 30. What is label propagation? Why would you implement it, and how?

Problem 31. Can you name two clustering algorithms that can scale to large datasets? And two that look for regions of high density?

Problem 32. Can you think of a use case where active learning would be useful? How would you implement it?

Problem 33. What is the difference between anomaly detection and novelty detection?

———— "*Genius = 1% Inspiration + 99% Perspiration ...*" -Thomas A. Edison ————