CSCI 416/516 Midterm Study Guide

Name:

Student ID:

---

1 K-Nearest Neighbors

- **Problem 1: KNN Concepts.**
  Describe the K-NN algorithm in your own words.
- **Problem 2: Euclidean Distance.**
  Given the following data points in a 2-dimensional space: $A = (1, 2)$, $B = (2, 4)$, $C = (2, 1)$, and $D = (3, 3)$, compute the Euclidean distance between point A and the other three points. Which point is the closest to point A?
- **Problem 3: Choosing K.**
  Discuss the implications of choosing a very small value for K versus a very large value.
- **Problem 4: Data Normalization.**
  Why is it important to normalize data when using the K-NN algorithm? Provide an example to support your answer.
- **Problem 5: Time Complexity.**
  K-NN has different computational costs for training and prediction. Explain the time complexity for both phases
- **Problem 6: Handling Categorical Data.**
  Discuss how you would handle categorical data in a dataset when applying the K-NN algorithm. What distance metric would you use?
- **Problem 7: Outliers and Noise.**
  Explain how outliers or noise in the data can affect the performance of the K-NN algorithm. What preprocessing steps can help mitigate these effects?
- **Problem 8: Curse of Dimensionality.**
  What is the curse of dimensionality in the context of K-NN? How does high dimensionality impact the effectiveness of the K-NN algorithm?
- **Problem 9: Imbalanced Datasets.**
  Discuss the challenges posed by imbalanced datasets in the context of K-NN.

- **Problem 10: K-NN's Assumptions.**
  K-NN, like all algorithms, makes underlying assumptions about the data. What are these assumptions, and how might they impact the model's performance in real-world scenarios?

2 Linear Regression

- **Problem 11: Linear Regression Concept.**
  Define linear regression. What are the primary components of a linear regression model?
- **Problem 12: Linear Regression Assumptions.**
  What are the primary assumptions underlying linear regression? List and briefly explain.
- **Problem 13: Measurement of Fitting.**
  How is the goodness of fit of a linear regression model measured?
- **Problem 14: Overfitting.**
  Explain the concept of overfitting in the context of linear regression. How can it be prevented?
- **Problem 15: Gradient Descent.**
  Explain the concept of gradient descent. How is it used in the optimization of linear regression?
- **Problem 16: Variations of GD.**
  Describe the difference between batch gradient descent, mini-batch gradient descent, and stochastic gradient descent.
- **Problem 17: Learning Rate.**
  How does learning rate affect the convergence of the gradient descent algorithm in linear regression optimization?
- **Problem 18: Regularization.**
  How is the L_2 regularization defined and why do we need it?
- **Problem 19: Polynomial Linear Regression.**
  What is polynomial linear regression? How does it differ from simple linear regression?
- **Problem 20: Polynomial Linear Regression.**
  Why are polynomial regression models particularly prone to overfitting? How can you detect and mitigate this?

3 Logistic Regression

- **Problem 21: Logistic Regression Concept.**
  Define logistic regression. How is it different from linear regression?

- **Problem 22: Sigmoid Function.**
  Explain the sigmoid function and its significance in logistic regression.

- **Problem 24: Cost Function.**
  How does the cost function for logistic regression differ from the one for linear regression?

- **Problem 24: Categorical Prediction.**
  How do you handle categorical predictors in logistic regression?

- **Problem 25: Gradient Descent.**
  Explain the steps in the Gradient Descent algorithm as it applies to logistic regression.

- **Problem 26: Evaluation.**
  How can we evaluate the performance of a logistic regression model?

- **Problem 27: Linear Model.**
  Why is logistic regression referred to as a "linear classifier" even though it models a nonlinear relationship between predictors and the probability outcome?

- **Problem 28: Training Set Size.**
  A colleague argues that logistic regression requires more samples to train effectively compared to linear regression. Do you agree? Explain your reasoning.

- **Problem 29: Residual.**
  Residual is the motivation for using logistic regression rather than linear regression. Why is that?

- **Problem 30: Decision Boundary.**
  What is the decision boundary in logistic regression? Provide a graphical illustration.

4 Support Vector Machine

- **Problem 31: Theory of SVMs.**
  Explain the principle of maximizing the margin in SVMs. How does this contribute to the model's generalization ability?

- **Problem 32: Kernel Tricks.**
  What is the kernel trick in SVMs? Provide examples of different types of kernels used in SVMs.

- **Problem 33: Support Vectors.**
  Define support vectors and explain their significance in the context of SVMs.

- **Problem 34: Parameter Tuning.**
  Discuss the role of parameters like C in SVMs. How do they affect the model's performance?

- **Problem 35: Hyperplane Decision Boundary.**
  What is a hyperplane in SVMs, and how does it help in classification tasks?

- **Problem 36: Dual Problem and Kernel Trick.**
  Explain how the dual problem in SVMs facilitates the use of the kernel trick. Why is this significant?

- **Problem 37: Lagrangian Multipliers.**
  Suppose a sample in the training dataset has a Lagrangian multiplier of 0.5. What does this say about this sample?

- **Problem 38: Primal and Dual.**
  How is the Primal in SVMs defined? And how is it related to the Dual?

- **Problem 39: Linear Separability.**
  Describe the techniques or strategies you can use to address the issue of linear inseparability when working with SVMs.

- **Problem 40: Kernel Tricks.**
  What are the advantages and disadvantages of the kernel tricks?