Machine learning assignment 3- Alireza Mohammadshafie

Short questions:

1. What is the kernel function, and what are its advantages? Is a kernel a type of similarity measure between training samples? (2 points)

I think when we face complication in using svm and can't solve the problem with just svm we have to use kernel which is a function that compute similarity of our samples in another dimensions. In this method there is no need to compute the transformation. We have to just calculate the similarity of our point with kernel which itself transform our samples in another dimensions and help us to solve classification problems.

2. Is it necessary to standardize features with different scales for decision trees and k-Nearest Neighbors (KNNs)? Please explain the reason. (1 point)

When we use KNN, standardizing features is necessary because KNN calculates distances and features with bigger scales affect the result more. But for decision trees, feature scale doesn't matter because trees split data by conditions on each feature, so we don't need to standardize for decision trees.

3. What are Maximum Likelihood Estimation (MLE) and Maximum A Posteriori (MAP)? What is the relationship between them? (1 point)

MLE finds the parameter values that make the observed data most likely, only based on data. MAP also includes prior beliefs we have about the parameters. MAP becomes MLE if we assume no prior information.

4. Please describe the advantages of stochastic gradient descent over gradient descent. (1 point)

Stochastic gradient descent is faster and works better for large datasets because it updates the model based on small batches or single examples instead of using all data at once. This makes it more efficient when we have complex data like pictures and large dataset.

5. Why is L2-regularized robust regression viewed as Laplace likelihood + Gaussian prior? (1 point)

Because the regression loss uses a Laplace distribution for the error, and adding L2 regularization is like assuming our model parameters are from a Gaussian distribution (the prior). So, together, it combines Laplace likelihood with a Gaussian prior.