Theoretical Assignment 3

Intro to Machine Learning, TAU

December 17, 2024

Question 1

The Hyperbolic Tangent kernel is defined as $k(x, y) = tanh(x^T y + \gamma)$. Where $\gamma \in \mathbb{R}$ is some constant scalar.
Show that this is not a valid kernel according to Mercer's theorem.
Hint: Develop the conditions for a set of 2 samples: $S = \{x_1, x_2\}$.

Question 2

The hinge loss $\max(0, 1 - y(w^T x + b))$ is commonly used in SVMs. However, consider an alternative quadratic loss:

$$L_i(w, b) = \left[\max(0, 1 - y_i(w^T x_i + b))\right]^2 .$$

(a) In soft-SVM, what do the slack variables represent? What is the geometrical meaning of the different value ranges that the slack variables can have? How are the slack variables related to the hinge loss?

(b) Write the primal optimization problem for a soft-margin SVM using this new quadratic loss. Hint: Consider about your answers to section a to understand how it would be different from the primal problem with the hinge loss.

(c) Write the Lagrangian, KKT conditions, and the dual problem for a soft-margin SVM using this new quadratic loss.

(d) Derive the gradient of the loss function $L_i(w, b)$ with respect to w and b.

(e) Discuss the implications of using this loss function instead of the hinge loss. How might this impact the sparsity of support vectors and the generalization performance?

Question 3

(a) What is the sample complexity of hard margin SVM in the linear case (without a kernel) and why? Explain your answer, don't just quote from class.

(b) Briefly explain how will using an RBF kernel affect the sample complexity. For an SVM using the RBF kernel, discuss how the sample complexity is influenced by the kernel hyperparameter σ, which controls the spread of the RBF kernel.

Question 4

Given a dataset with two classes, C_1 and C_2, and the following class distributions:

Feature Value	Class C_1 Count	Class C_2 Count
A	8	2
B	3	7

(a) Compute the entropy $H(S)$ of the entire dataset and for the sunsetes S_A S_B (the subsets corresponding to feature values A and B).
(b) Calculate the information gain $G(S, A)$ when splitting the dataset on the feature value.
(c) Interpret your results: What does the information gain tell you about the usefulness of splitting on this feature?
(d) Compute the Gini impurity for the entire dataset S, and for the subsets S_A and S_B **before** splitting.
(e) Calculate the weighted average Gini impurity **after** splitting the dataset by the feature values A and B.
(f) Compare the results from entropy-based information gain and Gini impurity. Are the rankings of the splits the same? Explain why or why not.

Question 5

Consider a binary decision tree used for binary classification, such as the ones we've seen in class, where each internal node splits the data based on a single feature threshold, and each leaf represents a class label. Suppose the dataset has n features, and each feature can take on real values.
(a) Prove that a binary decision tree with depth d can divide the feature space into at most 2^d regions.
(b) Explain how this result relates to the expressiveness of decision trees and their capacity to overfit.
(c) Using this result, determine the minimum depth d required for a decision tree to perfectly classify a dataset of m unique points (no two points share the same feature values).
(d) Based on the previous sections, what is the VC dimension of a tree with depth d?