Introduction to Machine Learning – Autumn 2024

Homework Assignment – 2

Due Date: 10th November 2024(midnight)

Question 1: (12 Marks)

Open the link below and you will find a notebook, implement the **TODO** sections and according to the results you get, answer the following questions.

https://colab.research.google.com/drive/1P6gOOYk68NqP5B1ZMgZcYiCIBWA6_RzN?usp=sharing

1. What patterns do you observe in the K-fold cross-validation results?
2. How does the polynomial degree affect the bias-variance trade-off?
3. What is the optimal polynomial degree for this dataset and why?
4. Explain the relationship between model complexity and overfitting based on your results.

Question 2: (10 Marks)

Consider a piece of data collected over the course of 14 days where the features are Outlook, Temperature, Humidity, Wind and the outcome variable is whether any sport was played on that day. Using this data, construct a decision tree manually which takes in above 4 parameters to predict whether a sport will be played based on the conditions for a given day.

Day	Outlook	Temperature	Humidity	Wind	Play
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Question 3: (8 Marks)

Maximal possible entropy is achieved when all states are equally probable (prove it yourself for a system with 2 states with probabilities p and 1−p). What's the maximal possible entropy of a system with N states?

Entropy (D) is given by:

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}. \qquad (8.7)$$

Question 4: (10 Marks)

You are given a set of points and asked to apply the K-means clustering algorithm with K=2. Use the specified initial centroids to assign points to clusters and update centroids iteratively until convergence.

Initial Centroids C1= (1,1) and C2= (5,7)

Points= {(2,2), (4,4), (5,5), (7,8)}