CS M146 Quiz Week 2 Solutions: Decision Tree, Nearest Neighbors and ML Basics

January 30, 2023

True/False (Overfitting, 2 points)

Training errors will increase, and the test error will decrease(True statement).
Overfitting occurs when you achieve a good fit of your model on the training data, while it does not generalize well on new, unseen data(test dataset). Hence, if a model is **overfitting**, then adding additional training examples can improve the model performance on unseen data (decreasing the test error) while increasing the training error. On the other hand, if a model is **underfitted**, then adding new training examples does not help.

True/False (Cross validation, 2 points)

In general, we choose the hyperparameter corresponding to the best-averaged performance. In this example, MaxDepth=2 achieves the best-averaged performance across the folds: $\frac{0.8+0.8}{2} = 0.8$

(Decision tree, 7 points)

- **Entropy**: $H(Play) = -(\frac{3}{5}log_2(\frac{3}{5}) + \frac{2}{5}log_2(\frac{2}{5})) = 0.97$
- **Conditional Entropy**:

$$\begin{aligned} H(Play|Outlook) &= \sum_{x\in\{S,O,R\}} P(Outlook = x)H(Play|Outlook = x) \\ &= \frac{2}{5}H(Play|Outlook = S) + \frac{2}{5}H(Play|Outlook = O) + \frac{1}{5}H(Play|Outlook = R) \\ &= \frac{2}{5}0 + \frac{2}{5}1 + \frac{1}{5}0 \\ &= 0.40 \end{aligned}$$

- **ID3 Algorithm**: Gain[Play,Outlook] is greater than Gain[Play,Temperature].Therefore, we choose Outlook.

(KNN, 5 points)

- **Concept**: For the large values of k, the classifier is more likely to overfit than underfit(False statement).
- **KNN practice**: The three nearest neighbors of $(-1, 2)$ are $(-1, 1)$,$(0, 1)$, and $(0, 2)$ points. Two of them have labels $-$ and one of them has label $+$. Therefore, $(-1, 2)$ will be labeled by $-$.