Name:	Roll No:	

Data Mining and Machine Learning

Quiz 2, II Semester, 2024–2025

20 February, 2025

Some questions below may have more than one correct answer. You get full credit if you select all correct options. You get partial credit if you select a non-empty, strict subset of correct options. You get zero credit if you select any incorrect option.

1. Which of the following help to control overfitting in decision trees.

   (a) Restrict the height of the tree. ✔
   (b) Restrict the width of the tree.
   (c) Insist that leaf nodes must have a minimum number of samples. ✔
   (d) Insist on a minimum percentage impurity gain when splitting a node. ✔

   Explanation:

   There is no mechanism to restrict the width of the tree, nor is there any justification to do so. Option (a) corresponds to pruning. Options (c) and (d) are available in the `scikit-learn` implementation of decision trees, and were discussed in class.

2. We use regression to fit a degree-three polynomial to inputs with three components (x_1, x_2, x_3). How many derived non-linear features do we have to add?

   (a) 9
   (b) 27
   (c) 36
   (d) 39

   Explanation:

   None of the options is correct. This question is not counted for marking. The correct answer is 16.

   There are 6 quadratic features: $\{x_1^2, x_2^2, x_3^2, x_1x_2, x_1x_3, x_2x_3\}$

   There are 10 cubic features: $\{x_1^3, x_2^3, x_3^3, x_1^2x_2, x_1^2x_3, x_2^2x_1, x_2^2x_3, x_3^2x_1, x_3^2x_2, x_1x_2x_3\}$

3. In linear regression, normalizing all attributes to the same range of values is helpful for:

   (a) Comparing the relative significance of attributes based on the final coefficients computed by regression. ✔
   (b) Ensuring that stochastic gradient descent converges.
   (c) Justifying the use of a uniform step size for gradient descent across all "directions". ✔
   (d) Avoiding overfitting.

   Explanation:

   Discussed in class, Lecture 9.

4. In logistic regression:

   (a) If we use squared error as the loss function, gradient descent will not converge.
   (b) We use cross entropy as the loss function to make gradient descent converge faster. ✔
   (c) The input to the sigmoid function has to be a linear function of the attribites.
   (d) We use the sigmoid function because the step function is not differentiable. ✔

   Explanation:

   (a) — gradient descent will converge, but the rate of convergence will be slow.

   (c) — sigmoid will convert any increasing sequence of inputs into a "smooth" step. It is not necessary that the inputs to the sigmoid be generated by a linear function.