

Homework 3*Handed Out: September 27**Due: October 11, 8:00 p.m.*

- You are encouraged to format your solutions using \LaTeX . You'll find some pointers to resources for learning \LaTeX among the Canvas primers. Handwritten solutions are permitted, but remember that you bear the risk that we may not be able to read your work and grade it properly — do not count on providing post hoc explanations for illegible work. You will submit your solution manuscript for written HW3 as a single PDF file.
- The homework is **due at 8:00 PM** on the due date. We will be using Gradescope for collecting the homework assignments. Please submit your solution manuscript as a PDF file via Gradescope. Post on Ed Discussion and contact the TAs if you are having technical difficulties in submitting the assignment.

1 Multiple Choice & Written Questions

Note: You do not need to show work for multiple choice questions. If formatting your answer in \LaTeX , use our LaTeX template [hw3_template.tex](#) (This is a read-only link. You'll need to make a copy before you can edit. Make sure you make only private copies.).

1. [k Nearest Neighbors] Consider properties of k -NN models:
 - a. (2 pts) Suppose that we are using k -NN with just two training points, which have different (binary) labels. Assuming we are using $k = 1$ and Euclidean distance, what is the decision boundary? Include a drawing with a brief explanation.
 - b. (2 pts) For binary classification, given infinite data points, can k -NN with $k = 1$ express any decision boundary? If yes, describe the (infinite) dataset you would use to realize a given classification decision boundary. If no, give an example of a decision boundary that cannot be achieved.
 - c. (2 pts) Suppose we take $k \rightarrow \infty$; what is the resulting model family?
 - d. (2 pts) What effect does increasing the number of nearest neighbors k have on the bias-variance tradeoff? Explain your answer. [Hint: Use parts (b) and (c) in your explanation.]
 - e. (2 pts) In logistic regression, we learned that we can tune the threshold of the linear classifier to trade off the true negative rate and the true positive rate. Explain how we can do so for k -NNs for binary classification. [Hint: By default, k -NN uses majority vote to aggregate labels of the k nearest neighbors; consider another option.]
2. [Decision Trees] In class, we discussed early stopping of generating splits and post-pruning. Here, we consider how they interact.

- a. (4 pts) Naïvely, one might expect that if we are using post-pruning on a validation set, then there is never any benefit to using early stopping conditions (e.g., maximum depth to split). Explain why this is not the case.
- b. (4 pts) Suppose we are training a decision tree with both early-stopping conditions and post-pruning. For each of the following, indicate whether it increases bias or variance: (i) increase the maximum depth of the decision tree, (ii) increase the minimum number of samples needed to split, (iii) disable post-pruning, and (iv) assuming we are using a feature map, add more features to the feature map.
3. [Decision Trees] You own a movie theater and are trying to understand your market: what types of people frequently go to the movies? You start with the following dataset with data about 6 people with different age groups, income levels, and professions, and whether or not they frequently go to movie theaters. In particular, you are going to build a decision tree to predict whether or not someone is a frequent movie-goer.

No	High Income?	Engineer?	Movie Goer?
1	T	T	Yes
2	F	F	Yes
3	T	T	Yes
4	F	T	No

Recall the following definitions of entropy and information gain, respectively, which are useful for this problem:

$$H(Z) = - \sum_y P(Y = y) \log_2 P(Y = y)$$

$$\text{IG}(Z, j, t) = H(Z) - H(Z[x_j = t])P(x_j = t) - H(Z[x_j \neq t])P(x_j \neq t).$$

- a. (4 pts) Based on the principle of information gain, which attribute is to be used for the first split? Be sure to show your computations. You can round the entropy and information gain values to two decimal places.
- b. (4 pts) Draw the complete (unpruned) decision tree, showing the class predictions at the leaves. Assuming you are using LaTeX, you may (i) very neatly hand draw the tree, photograph it, and include it as a figure, (ii) draw it using a graphics program or PowerPoint, or (iii) express the tree in a series of if statements, preferably using LaTeX's verbatim environment.
- c. (2 pts) From the Decision Tree constructed in the previous question, predict whether a person who has high income but is not an engineer is a movie goer.
4. [Decision Trees] (6 pts) In class, we covered how to learn binary decision trees. In this question, you will modify the decision tree learning algorithm to obtain ternary decision trees (i.e, each internal node has three children). Assume all features are real-valued, and the labels are categorical (i.e., classification).

- a. (2 pts) Each internal node should split a single feature into three intervals $x_j \geq t_1$, $t_1 > x_j \geq t_2$, and $t_2 > x_j$. Write down all parameters for this split.
- b. (2 pts) Write down the information gain for a single split. You do not have to expand the entropy (e.g., for traditional decision trees, the information gain is $IG(j, t) = H(Z) - H(Z[x_j \geq t])P(x_j \geq t) - H(Z[x_j < t])P(x_j < t)$).
- c. (2 pts) Does this strategy increase, decrease, or keep unchanged the model capacity compared to traditional decision trees? What if the depth of the decision tree is bounded by 3?