Name: _____

Model Evaluation

4. (10 points) Lisa trains models for classification problems. She is provided with different image data sets (e.g., trains, people, cars, cats, dogs) by Snapbook. Each data set has both positive and negative examples. In fact, Snapbook provides Lisa only a fraction of each data set, the remainder is left for internal Snapbook testing. Lisa trains a separate model on each data set. She measures model training accuracy, and she estimates test accuracy using cross-validation. For each model, Snapbook measures the accuracy of the model on the data that was held out (not provided to Lisa). These experiments yield the following results:

	training accuracy	cross-validation accuracy	held-out tests accuracy
data set 1	52%	54%	51%
data set 2	97%	71%	70%
data set 3	93%	92%	55%
data set 4	91%	91%	89%
data set 5	50%	53%	70%

For which data set(s):

(a) Lisa's model is overfitting (check all that apply):
◯ data set 1 √ **data set 2** ◯ data set 3 ◯ data set 4 ◯ data set 5

Solution: In data set 2, the training accuracy is extremely high, while the cross-validation and test accuracy are significantly lower, pointing to overfitting. None of the other data sets show this significant of a difference.

(b) It is likely that more training data drawn from the same distribution would improve the quality of the held-out accuracy (check all that apply):
◯ data set 1 √ **data set 2** ◯ data set 3 ◯ data set 4 ◯ data set 5

Solution: As above, Lisa's model for data set 2 is overfitting, and so having more training data can mitigate this issue.
Data set 3 also has a held-out accuracy that is much lower than the training accuracy, but the distinction is that here, the cross-validation accuracy is high. This implies that more training data would not help in generalization, and that instead there is a more fundamental problem with Lisa's training data not being drawn from the same distribution as the test data.

(c) Lisa's hypothesis class might not be expressive enough (check all that apply):
√ **data set 1** ◯ data set 2 ◯ data set 3 ◯ data set 4 √ **data set 5**

Solution: Low training and cross-validation accuracies can point to a hypothesis class not being expressive enough. This occurs for data sets 1 and 5.

(d) Held-out data set is not likely from the same distribution as Lisa's (check all that apply):
◯ data set 1 ◯ data set 2 √ **data set 3** ◯ data set 4 √ **data set 5**

Solution: This is usually the case when the cross-validation accuracy is very different from the held-out test accuracy. This occurs for data sets 3 and 5.