CS 5703 Machine Learning Practice
Midterm 2
November 10, 2026

General instructions:

- Please wait to open this exam booklet until you are told to do so.
- This examination booklet has 6 pages.
- Write your name and the date, and sign your name below.
- The exam is closed book, notes, and electronic devices. The exception is that you may have one page of personal notes (double sided).
- The exam is worth a total of 150 points (and 15% of your final grade).
- You have 1.25 hours to complete the exam. Be a smart test taker: if you get stuck on one problem, go on to the next.

On my honor, I affirm that I have neither given nor received inappropriate aid in the completion of this exam.

Signature: _____ **Name:** _____

Date: _____

Question	Points	Score
Non-Linear Preprocessing	32	
Over-fitting and Regularization	38	
Cross-Validation and Hyper-parameter Selection	42	
Hypothesis Testing and Model Comparison	38	
Total:	150	

Part I. Non-Linear Preprocessing

1. (8 points) Assume a two-feature input space and a polynomial transformation of degree $d = 3$. How many distinct polynomial features are there?

   Solution: x_0^3, x_1^3, x_0^2, x_1^2, $x_0^2x_1$, $x_0x_1^2$, x_0x_1, x_0, x_1, 1. Total = 10.

2. (8 points) Assume a three-feature input space, and a polynomial transformation of degree $d = 2$. How many distinct polynomial features are there?

   Solution: x_0^2, x_1^2, x_2^2, x_0x_1, x_0x_2 x_1x_2, x_0, x_1, x_2, 1. Total = 10.

3. (8 points) Assume a two-feature input space, a $d = 2$ polynomial transformation, and the following coefficients for a classifier. What is the binary output of the classifier (positive or negative) for input $[3, 0]$?

Coefficient	**Term**
2	x_0^2
1	x_1^2
2	x_0x_1
3	x_0
-1	x_1
-20	1

   A. positive B. negative

4. (8 points) What is the output of the above classifier for input $[-2, 5]$?

   A. positive **B. negative**

Part II. Over-fitting and Regularization

5. (6 points) What does *over-fitting* refer to?

Solution: Over-fitting refers to when a learned model performs well on a training data set, but not well on independent data.

6. (10 points) A single model is clearly over-fitting its training set. List four approaches you can take to possibly compensate for this.

   1. Increase training set size
   2. Decrease model complexity
   3. Decrease the input feature dimensionality
   4. Increase/add regularization
   5. (For gradient based methods) add early stopping

7. (6 points) Assuming that alpha is high enough, which of Lasso, Ridge, and Elastic Net will set coefficients to exactly zero? Mark all that are true.

   √ **Lasso** ◯ Ridge √ **Elastic Net**

8. (8 points) Briefly explain why training set performance can decrease as the number of training samples increases.

Solution: With a small training set, it is possible for a model to effectively memorize each of the samples. As new samples are added to the set, the model is forced to accommodate the variance within the sample.

9. (8 points) Briefly explain why validation set performance can increase as the magnitude of the regularization parameter (e.g., alpha) increases.

Solution: Especially when the training data set size is small, increasing the regularization parameter directs the learning process toward a *smoother* solution that tends to capture the general trend in the training data, rather than the specific details of the training data.

Part III. Cross-Validation and Hyper-parameter Selection

10. (6 points) For Holistic $N = 30$ Fold Cross-Validation, what is the maximum number of folds used for testing purposes for a given rotation?

Solution: In the class, we have talked only about using one fold.

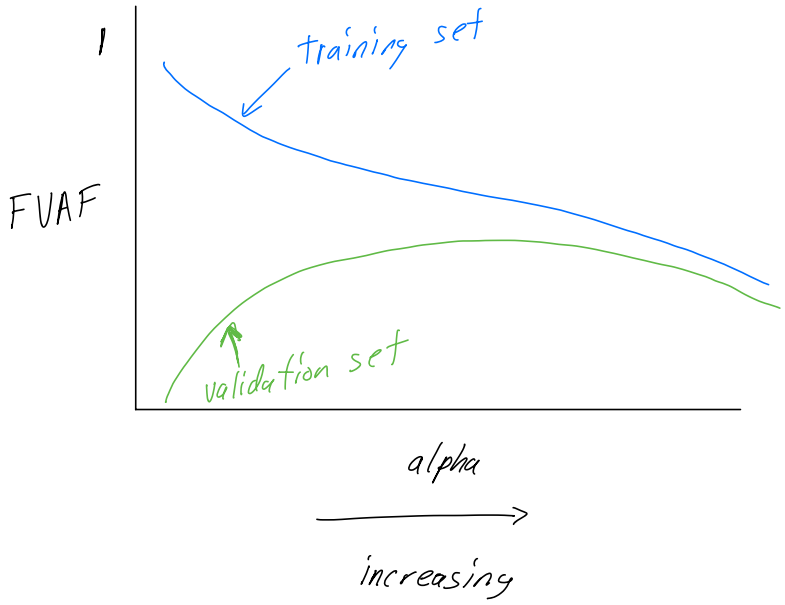
11. (8 points) In N-Fold Cross-Validation, which data set do we use for hyper-parameter selection? Explain why we do not use either of the other two data sets.

Solution: We use the validation data set performance to select hyper-parameters.
Training data set: we cannot detect over-fitting.
Testing data set: the testing data set is no longer a proper simulation of data that the model might see in the future.

12. (8 points) In creating the N folds of data for cross-validation, what must we ensure?

Solution: The folds should generally be balanced (similar number of samples), and the data from one fold must be independent of the data from the other folds.

13. (8 points) Sketch typical average training and validation performance as a function of a regularization hyper-parameter. Label your axes.



14. (6 points) Describe where testing performance would be placed on the above figure.

Solution: We would typically choose the hyper-parameter value corresponding to the best validation set performance. Ideally, the test set performance at this point will be comparable to the validation set performance.

15. (6 points) Suppose a regression model is trained to predict the value of a commodity in US dollars. What are the units of the FVAF measure of model performance?

Solution: FVAF is a unitless measure.

Part IV. Hypothesis Testing and Model Comparison

16. (8 points) What are the requirements for using a paired sample test in the context of comparing two model types?

Solution: We use the same test data sets to evaluate the performance of models A and B, which allows us to measure direct contrasts between the models.

17. (6 points) Which one of the following is most true for a probability function?
A. $0 \leq p(y|...)$ B. $p(y|...) \leq 1$ **C. Both A and B** D. Neither A nor B

18. (8 points) Briefly explain why we formulate our hypothesis testing procedure in terms of a null hypothesis that we want to reject, as opposed to an alternative hypothesis that we want to accept.

Solution: The number of null hypotheses is small (one), whereas the number of alternative hypotheses is hard to enumerate.

19. (8 points) In model comparison, explain briefly why the test statistic for the difference in mean model performance is assumed to follow a t-distribution under the null hypothesis, rather than a standard normal (z) distribution.

Solution: While we define the mean of the distribution to be zero, the standard deviation is not specifically defined. Instead, we estimate this from the data itself. Using the t-distribution compensates for the uncertainty of this estimate.

20. (8 points) Contrast hold-out cross-validation from holistic cross-validation.

Solution: With hold-out cross-validation, a single test data set is held out and used for the final evaluation of the N models, whereas holistic cross-validation uses N independent testing data sets. This gives us some confidence that the holistic performance metrics are statistically independent of one-another, whereas the hold-out test statistics are definitely not independent.