This is a closed book exam.

Write clearly and justify your answers.

Incorrect statements in essay-type questions will incur negative points.

Time limit: 90 min.

(1) [5 points] *(essay-type question)* We have the following learning algorithms:

- **L1**: L1-regularized linear regression.
- **L2**: L2-regularized linear regression.
- **kNN**: k-nearest neighbors with Euclidean distance.
- **RF**: Random Forests.
- **FFNN**: A feed-forward neural network trained using backpropagation.

If you think some key detail about a learner is missing, you may assume which parameter was used or how it was implemented. However, the assumption must be sensible - something that could reasonably used as a default in some implementation (do not assume behavior that makes the learner clearly behave poorly or is not even feasible to implement). And, whatever you assume, applies to all the below questions.

Order the above learners from best to worst with respect to the following dimensions:

a. The time complexity of learning (consider both the number of observations and the number of input variables).

b. Time complexity of making a prediction (consider both the number of observations and the number of input variables).

c. Interpretability/ease of understanding the model and its predictions.

d. Difficulty of implementation if you had to implement it from scratch without third-party libraries.

e. Memory (space) complexity of storing the learned model.

Justify your answers.

(2) [5 points] *(essay-type question)* List 3 different dimensionality reduction/clustering techniques that you know. For each of them briefly summarize in which case you would prefer over the others (or why you would never prefer to use it). For each of them determine the time complexity of applying it (how it depends on the dimensionality of the data and the number of observations).

(3) [5 points] *(essay-type question)* Joe F. Random, a machine learning practitioner, is working on a classification problem. The constraints are as follows: He has data with 100 observations, 50 predictors (independent variables), and a binary class (dependent variable). He has to deliver a model that will be used in production without any possibility of modification. He can only choose between learner A or B, both of which have some continuous hyperparameters, and he can only use them in a black-box manner (train with some hyperparameters and use the trained model for predicting, repeating this an arbitrary number of

times for different hyperparameters). Additionally, if he can't with some certainty conclude that learner B is better than learner A, learner A must be delivered.

Joe decided on the following training and evaluation process: He selected (uniformly at random) 20 different sets of hyperparameters for A and B. The estimated the performance of these 40 combinations of learner and hyperparameters using leave-one-out cross validation. Learner B with some set of hyperparameters had the best loss so he retrained learner B with those parameters on all 100 observations and delivered that model to production.

a. Does Joe's model evaluation process have any potential sources of bias (positive/negative?) or variance? That is, identify all sources of error when estimating true risk (how the delivered model is expected to perform in production) with Joe's performance estimate.

b. Are there any other potential flaws in Joe's learning process (how the models were trained and hyperparameters chosen)? Suggest a better learning and evaluation process that is still within the constraints of the problem.

(4) [5 points] Some machine learning-related mathematics and algorithms:

a. Define Empirical Risk Minimization and show how Maximum Likelihood is a special case.

b. Write the likelihood of a Poisson GLM with the logarithm link function. Hint: $P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$.

c. Define the k-fold cross-validation estimator of a prediction model's generalization error and discuss how k affects the properties of the estimator.

d. Write the algorithm for boosting.

e. Define the Shapley value of a player in coalitional game theory.

(5) [bonus 3 points] Some Bayesian statistics:

a. Let's assume some data y and a parametric model $p(y|\theta)$, where θ are the model's parameters. Define Bayesian inference and Maximum Likelihood estimation.

b. What are the advantages and disadvantages of using Bayesian inference as opposed to using Maximum Likelihood inference or another form of Empirical Risk Minimization.

c. Derive the Bayes estimator (in the continuous case) for quartic loss ($\ell(y, \hat{y}) = (y - \hat{y})^4$, where y is the true and $\hat{y}$ the predicted value).

The End.