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.
- **Kernelized**: Kernelized linear regression with the RBF kernel.
- **SVR**: Support Vector Regression with the RBF kernel.
- **RF**: Random Forests.

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)* What is the curse of dimensionality, and what effect does it have on the performance of the machine learning algorithms? Consider the k-nearest neighbors algorithm and describe the problem of increasing the number of dimensions of input data for this algorithm, say, when the data includes 100 dimensions. To illustrate the problem, you may use an example of uniformly distributed data in data space that is constrained to a hyper-cube. Comment on how an increase of dimensions affects the non-myopic feature scoring technique of Relief.

(3) [**5 points**] *(essay-type question)* Joe F. Random is working on a prediction model for the Ministry of Agriculture, Forestry, and Food. He aims to create a model that can predict which plants farmers grow on a particular field from features encompassing its terrain, soil, and weather characteristics. Joe was given a perfectly square region to train and evaluate his model. He decides to split the region into $n \times n$ equally sized square subregions. For evaluation, he uses 70-30 train-test splits on subregions (random 70% of subregions are used for training and remaining 30% for testing), repeated 3 times.

a. Argue on the suitability of Joe's evaluation.

b. How would the results change if he used smaller or larger subregions (if he increased or decreased n for region split, respectively)? How many subregions would be the best with the given 70-30 train-test split?

c. Propose a better evaluation procedure.

When answering, use appropriate terminology regarding estimators and their properties.

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

a. Write the pseudo-code for k-fold cross-validation that also performs variance estimation.

b. Define a strictly proper scoring rule.

c. Write the Stochastic Neighbor Embedding algorithm (in pseudo-code). Be precise in the objective that we are trying to optimize.

d. Explain why should we remove strongly correlated features prior to using linear regression.

e. Explain why we typically set one of the categories to be the reference category (coefficients set to 0) in multinomial logistic regression.

(5) [**bonus 2 points**] Some Bayesian statistics:

a. Show that the maximum likelihood of the coefficients for L2 regularized linear regression is the same as the posterior mode of Bayesian linear regression with a normal prior on the coefficients.

b. List approaches for Bayesian inference—approaches for the actual computation of the posterior—and explain their advantages and disadvantages.

The End.