BTM KAIST **Name:** _____
Applied Machine Learning
Midterm Sample
Time Limit: 75 Minutes _____

1. This exam contains 7 pages (including this cover page) and 19 questions.
2. Total of points is 94.
3. Answer all questions. The marks for each question are indicated at the beginning of each question.
4. This **IS an CLOSED BOOK** exam.
5. Calculators are not allowed.

1 True/False (20 points)

1. (2 points) If highly correlated but relevant features are present in a dataset, Lasso regression will select one of them at random. **T**
2. (2 points) Tuning two hyper-parameters with four options each using grid-search with 5-fold cross-validation requires exactly 40 model fits. **F**
3. (2 points) It is good practice to standardize sparse datasets so that each feature has zero mean. **F**
4. (2 points) Ridge Regression adds an L_1-norm penalty to the cost function and often sets several of the weights to zero. **F**
5. (2 points) 5-NN has more overfitting (lower bias) than 1-NN. **F**
6. (2 points) It is important that exactly same scaling transformation is applied to the training set and the test set for the supervised model. **T**
7. (2 points) The distinction between the training set, validation set, and test set is fundamentally important to apply machine learning methods in practice. Any choices made based on the test set accuracy "leak" information from the test set into the model. **T**

2 Multiple Choice (20 points, 4pts each)

Select **all** choices that apply.

8. (4 points) After training a ridge regression model, you find the training and test accuracies are 0.97 and 0.55, respectively. Which of the following would be the best choice for the next ridge regression model you train?

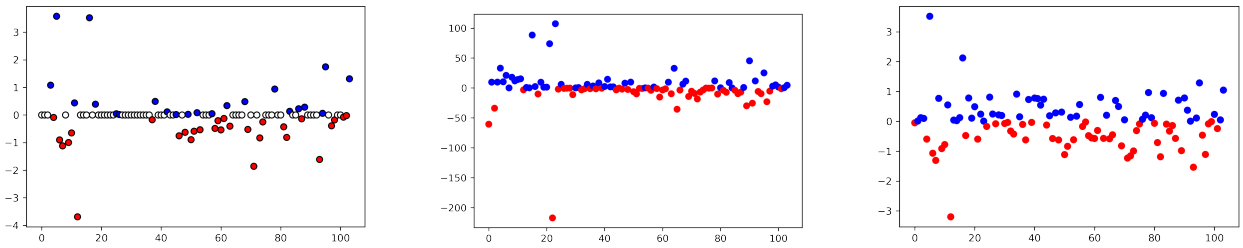
   A. You are overfitting, the next model trained should have a lower value for alpha

   B. You are overfitting, the next model trained should have a higher value for alpha

   C. You are underfitting, the next model trained should have a lower value for alpha

   D. You are underfitting, the next model trained should have a higher value for alpha

9. (4 points) Match the plots below to the correct regression types.



   A. Ridge, Lasso, OLS

   B. OLS, Ridge, Lasso

   C. Lasso, OLS, Ridge

   D. Ridge, OLS, Lasso

   E. OLS, Ridge, Lasso

10. (4 points) Which of the following variables should be treated as categorical?

   ☐ Income

   √ **Nationality**

   √ **Gender**

   ☐ Age

   √ **ZIP code**

11. (4 points) Suppose you are interested in finding a parsimonious model (the model that accomplishes the desired level of prediction with as few predictor variables as possible) to predict housing prices. Which of the following would be the best choice?

   A. Ridge Regression

B. Ordinary Least Squares Regression

C. Logistic Regression

D. Lasso Regression

12. (4 points) Which of the following is not part of data preprocessing?

A. Scaling

B. Data transformation

C. One-Hot-Encoding

D. Feature Selection

E. Cross-validation

3 Debugging

For each code snippet, find and explain **all** errors given the task. There can be more than one errors. You can write your answer on the empty spaces on this page.

13. (10 points) Task: Perform grid search (without using the `GridSearchCV` class) using a split into training, validation, and test data, with a final valuation on the test data.

```
X_trainval, X_test, y_trainval, y_test=train_test_split(X,y)
X_train, X_valid, y_train, y_valid=
train_test_split(X_trainval, y_trainval)

best_score=0

for C in [0.001, 0.01, 0.1, 1, 10, 100]:
    svm=LinearSVC(C=C)
    svm.fit(X_train, y_train)
    score=svm.score(X_test, y_test)
    if score > best_score:
        best_score=score
        best_C=C

svm=LinearSVC(C=best_C).fit(X_valid, y_valid)
```

Solution:

1. `score=svm.score(X_valid, y_valid)`
2. `svm=LinearSVC(C=best_C).fit(X_trainval, y_trainval)`

14. (10 points) Task: Create a pipeline of a scaler, imputer, and linear regression. Test the model with cross validation and get the score on the test dataset.

```
pipe=make_pipeline (StandardScaler(), SimpleImputer, LinearRegression())
mymodel=cross_val_score(pipe, X_train, y_test, cv=10)
mymodel.score(X_test, y_test)
```

Solution:

1. Forgot to put the parenthesis after SimpleImputer
2. Cross validation does not return a model.

4 Coding

Assume all necessary imports are already made.

15. (10 points) Provide the code to build `LinearRegression` model and evaluate its performance on a separate test set, given a dataset as numpy arrays `X` and `y`.

Solution:

```
X_train, X_test, y_train, y_test=train_test_split(X,y,stratify=y)
lr=LinearRegression()
lr.fit(X_train,y_train)
R2=lr.score(X_test,y_test)
```

16. (10 points) Provide code to implement grid-searching the parameters `C` and `gamma` of an SVC in a pipeline with a `StandardScaler`, and evaluating the best parameter setting on a separate test set, given a dataset as numpy arrays `X` and `y`.

Solution:

```
pipe=make_pipeline(StandardScaler(),SVC())
params={'svc__gamma':np.logspace(-3,2,6),'svc__C':np.logspace(-3,2,6)}
#you could also use n_features to change the range of gamma.
#Either is acceptable
#exact ranges dont matter, but should be logscale
X_train, X_test, y_train, y_test=train_test_split(X,y,stratify=y)
grid=GridSearchCV(pipe, param_grid=params,cv=5)
#cv not necessary, specifiying a different scoring would also be fine
grid.fit(X_train,y_train)
score=grid.score(X_test,y_test)
```

5 Concepts (20 Points)

17. (5 points) What is overfitting?

Solution: In statistics, overfitting is "the production of an analysis that corresponds too closely or exactly to a particular set of data, and may therefore fail to fit additional data or predict future observations reliably". An overfitted model is a statistical model that contains more parameters than can be justified by the data. The essence of overfitting is to have unknowingly extracted some of the residual variation (i.e. the noise) as if that variation represented underlying model structure.

18. (5 points) Why are nearest neighbor methods sensitive to the scaling of the data?

Solution: Nearest neighbors method usually rely on euclidean distances. Euclidean distances depend on the scale of the data. If one feature is on a much larger scale than he others, the distance in this feature will outweigh distances in all other features.

19. (10 points) A real estate firm would like to build a system that predicts the sale prices of a house. They create a spreadsheet containing information about 1460 house sales in the Santa Clara area. In addition to the price, there are 79 features describing the house, such as number of bedrooms, total indoor area, lot area, has a garage, location, etc. Explain how you would implement a machine learning model that would solve this prediction task. You don't need to show Python code, but please give a description of the system and explain all steps you would carry out when developing it.

Solution: We first convert the spreadsheet into a matrix using pandas library. This is clearly a regression problem, so pick a useful regression model, such as OLS, Ridge, or Lasso. Methodologically, we split the data into two parts, test, and train data sets. We select the best mode (hyperparameter tuning) via cross validation. Finally, we evaluate on the test set. The evaluation metric will be a regression metric, such as mean-squared error (R^2). To get the full score, you need to say that you will apply a regressor, and discuss how to evaluate it.