Practice Midterm - Applied Machine Learning COMS W4995

Date:

Name:

UNI:

For all choice boxes, please fill in the box you want to choose like this: ■
Otherwise your answer can not be graded.

1 True/False (+ 2pt each)

	True	False
If highly correlated but relevant features are present in a dataset Lasso regression will select one of them at random.	☐	☐
Accuracy is a good metric for multi-class classification in the presence of heavily imbalanced classes.	☐	☐
Tuning two hyper-parameters with four options each using grid-search with 5-fold cross-validation requires exactly 40 model fits.	☐	☐
Ridge regression does not work on data with collinear features	☐	☐
Hexbin plots are a way to resolve overplotting issues.	☐	☐
It is good practice to standardize sparse dataset so that each feature has zero mean.	☐	☐
A node in a decision tree always contains exactly half the samples of its parent.	☐	☐
Kernel support vector machines don't scale well to large datasets.	☐	☐
Decision Trees are very sensitive to the scaling of the data.	☐	☐
For a perfectly calibrated classifier, 80% of the data for which p(y=1) =.8 belong to class 1.	☐	☐

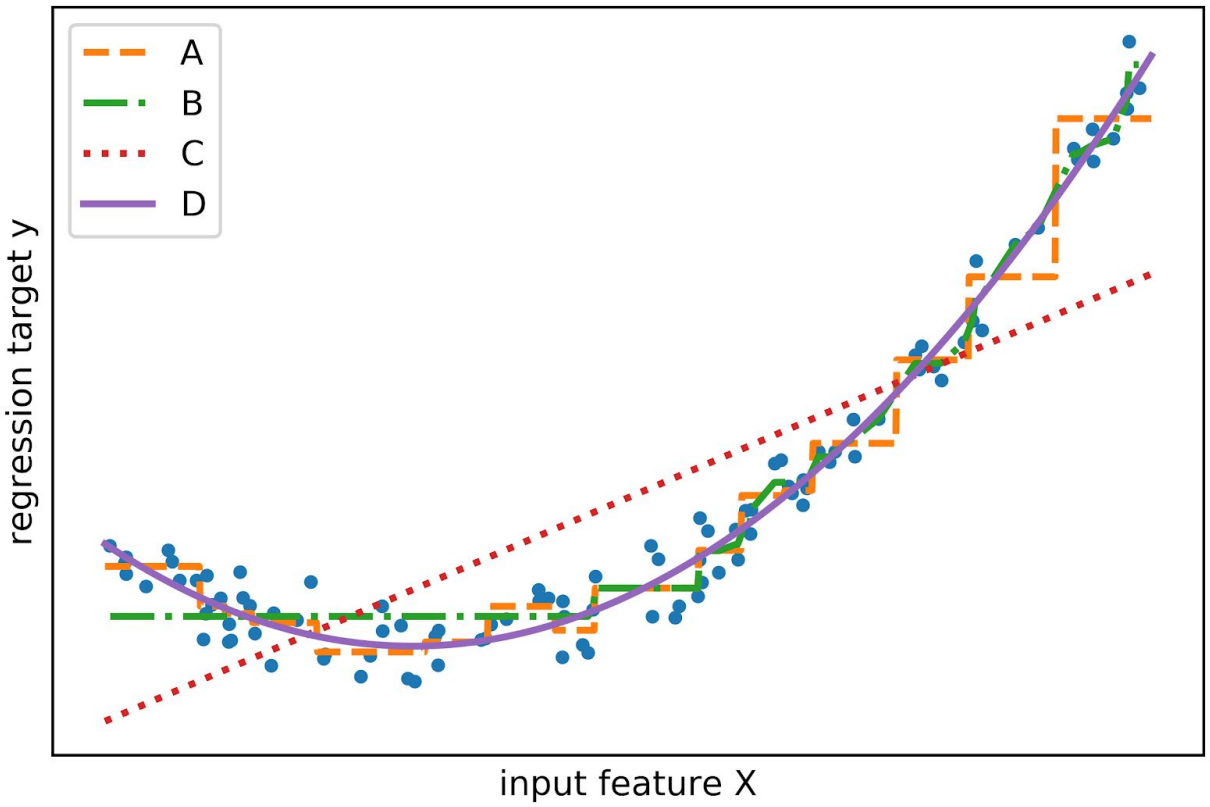
2 Multiple choice (20pt)

Select all choices that apply.

2.1 Which of the following are non-parametric models?

- ☐ Random Forest
- ☐ Linear Regression
- ☐ Logistic Regression
- ☐ Nearest Neighbors
- ☐ Nearest Shrunken Centroid

2.2 Given a 1d regression problem as follows (blue dots are training data), which of the following assignments of models to predictions is consistent with the graph:



- ☐ A is a tree
- ☐ A is isotonic regression
- ☐ B is a linear model
- ☐ B is isotonic regression
- ☐ C is a tree
- ☐ C is a linear model
- ☐ D is polynomial regression
- ☐ D is a random forest

2.3 Which of the following variables should be treated as categorical?

- ☐ Income
- ☐ Nationality
- ☐ Gender
- ☐ Age
- ☐ ZIP code

2.4 2.1 What are possible reasons that cross-validation could yield a very different accuracy than evaluating on an independent, unused test set?

- ❏ Data is not independently distributed, as in time series.
- ❏ Data is not linearly separable.
- ❏ Class balances are different between the cross-validation data and test data.
- ❏ Overfitting of hyper-parameters to the cross-validation.

3 Debugging (10pt each)

For each code snippet, find and explain all errors given the task. There can be more than one.

3.1 Task: Perform grid-search (without using the GridSearchCV class) using a split into training, validation and test data, with a final evaluation on the test set.

```
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)
score = svm.score(X_test, y_test)
```

3.2 Task: Use the PowerTransformer (implementing the box-cox transformation) transformer to preprocess data and learn a Ridge model, and visualize the coefficients.

```
pipe = make_pipeline(StandardScaler(), PowerTransformer(), Ridge())
scores = cross_val_score(pipe, X_train, y_train, cv=10)
plt.barh(range(X_train.shape[0]), pipe.coef_)
```

4 Coding (10 each)

4.1 Provide code to build a `LogisticRegression` model and evaluate its performance on a separate test set, given a classification dataset as numpy arrays X and y.

4.2 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 data as numpy arrays X and y.

5 Concepts (5pt each)

Answer each question with a short (2-5 sentences) explanation.

5.1 How are the “jet” and “viridis” colormaps different and why does it matter?

5.2 What is underfitting?

5.3 Why are nearest neighbor methods sensitive to the scaling of the data?

5.4 Explain target encoding of categorical variables.