

# 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.	<input type="checkbox"/>	<input type="checkbox"/>
Accuracy is a good metric for multi-class classification in the presence of heavily imbalanced classes.	<input type="checkbox"/>	<input type="checkbox"/>
Tuning two hyper-parameters with four options each using grid-search with 10-fold cross-validation requires exactly 80 model fits not counting refitting the best model.	<input type="checkbox"/>	<input type="checkbox"/>
Ridge regression does not work on data more features than samples	<input type="checkbox"/>	<input type="checkbox"/>
Hexbin plots are a way to resolve overplotting issues.	<input type="checkbox"/>	<input type="checkbox"/>
It is good practice to standardize sparse dataset so that each feature has zero mean.	<input type="checkbox"/>	<input type="checkbox"/>
Trees with larger maximum depth overfit more.	<input type="checkbox"/>	<input type="checkbox"/>
The one-vs-one classification heuristic for multi-class classification trains every binary classifier on the whole original training data.	<input type="checkbox"/>	<input type="checkbox"/>
Decision Trees are sensitive to the scaling of the data.	<input type="checkbox"/>	<input type="checkbox"/>
For a perfectly calibrated classifier, 80% of the data for which $p(y=1) = 0.8$ belong to class 1.	<input type="checkbox"/>	<input type="checkbox"/>

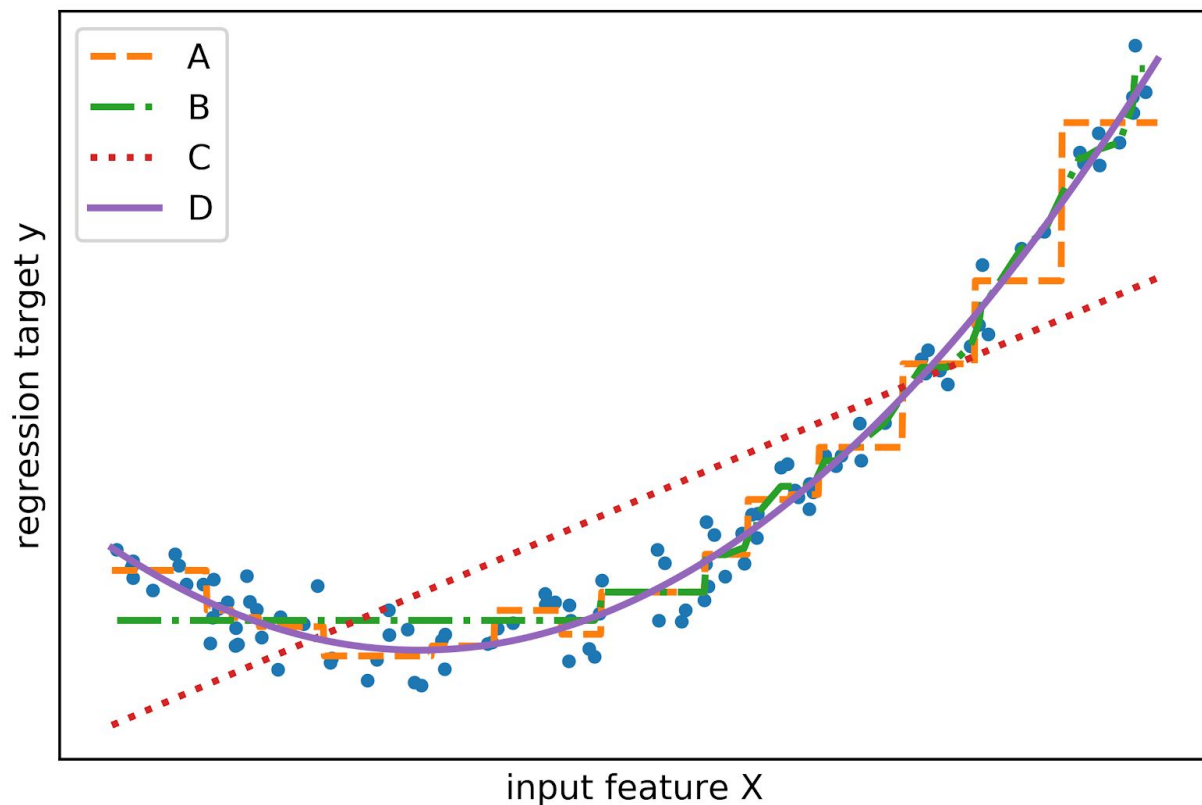
## 2 Multiple choice (30pt)

Select all choices that apply.

2.1 Given a fitted logistic regression model, assume we change the offset / intercept  $b$  by adding 100 to it. Which of the following metrics would be impacted on the test set?

- ☐ Average Precision
- ☐ F1 Score
- ☐ Macro Average Recall
- ☐ Brier Score
- ☐ ROC AUC

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

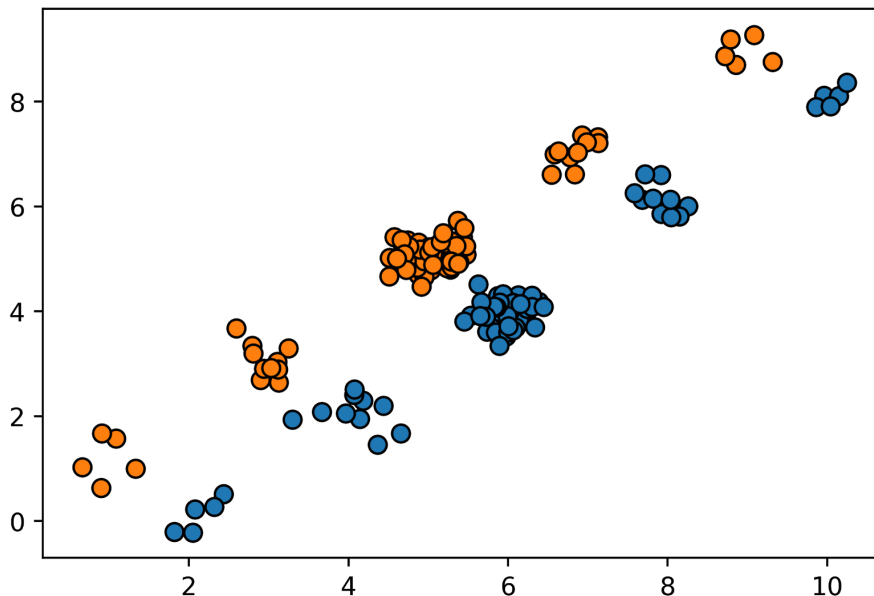
2.3 Which of the following variables should be treated as categorical?

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

2.4 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.

2.5 Which of the following models will be able to achieve 100% training accuracy on the following dataset?



- ☐ DecisionTreeClassifier(max\_leaf\_nodes=4)
- ☐ DecisionTreeClassifier(max\_depth=4)
- ☐ DecisionTreeClassifier(min\_samples\_split=100)
- ☐ ExtraTreesClassifier(n\_estimators=1000, max\_depth=1)

2.6 Which of the following statements is true about SMOTE?

- ☐ SMOTE can add new, synthetic samples, to your dataset.
- ☐ SMOTE duplicates existing samples.
- ☐ The main tuning parameter of SMOTE is the number of neighbors to consider when adding a new point.
- ☐ SMOTE will always improve accuracy on imbalanced datasets.

### 3 Debugging (10pt each)

For each code snippet, find and explain all errors given the task. There can be more than one. Assume all necessary imports are already made.

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.

```
1 | X_trainval, X_test, y_trainval, y_test = train_test_split(X, y)
2 | X_train, X_valid, y_train, y_valid = train_test_split(
3 |     X_trainval, y_trainval)
4 |
5 | best_score = 0
6 |
7 | for C in [0.001, 0.01, 0.1, 1, 10, 100]:
8 |     svm = LinearSVC(C=C)
9 |     svm.fit(X_train, y_train)
10 |    score = svm.score(X_test, y_test)
11 |    if score > best_score:
12 |        best_score = score
13 |        best_C = C
14 |
15 | svm = LinearSVC(C=best_C).fit(X_valid, y_valid)
16 | score = svm.score(X_test, y_test)
```

3.2 Task: Apply logistic regression to a dataset consisting only of categorical variables given as integers, and having missing values and visualize the 10 most important coefficients. Assume that feature\_names is an array of length n\_features containing strings describing the features and X\_train, y\_train are given.

```
1 | pipe = make_pipeline(SimpleImputer(strategy="mean"), OneHotEncoder(),
2 |                     LogisticRegression())
3 | pipe.fit(X_train, y_train)
4 | coef = pipe.named_steps['logisticregression'].coef_
5 | important = np.argsort(coef)[-10:]
6 | plt.barh(range(10), coef[important])
7 | plt.yticks(range(10), feature_names[important])
```

#### 4 Coding (10pt)

Assume all necessary imports are already made.

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  $y$ . Assume there are no missing values or categorical features.

## 5 Concepts (5pt each)

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

5.1 Why is macro-average recall a more useful metric for gridsearch on a binary classification problem than recall of the positive class?

5.2 Why are pipelines essential when working with scikit-learn?

5.3 Why is accuracy a bad metric for binary classification with imbalanced datasets?

5.4 Explain target encoding of categorical variables.