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 10-fold cross-validation requires exactly 80 model fits not counting refitting the best model.	□	■
Ridge regression does not work on data more features than samples	□	■
Hexbin plots are a way to resolve overplotting issues.	■	□
It is good practice to standardize sparse dataset so that each feature has zero mean.	□	■
Trees with larger maximum depth overfit more.	■	□
The one-vs-one classification heuristic for multi-class classification trains every binary classifier on the whole original training data.	□	■
Decision Trees are 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.	■	□

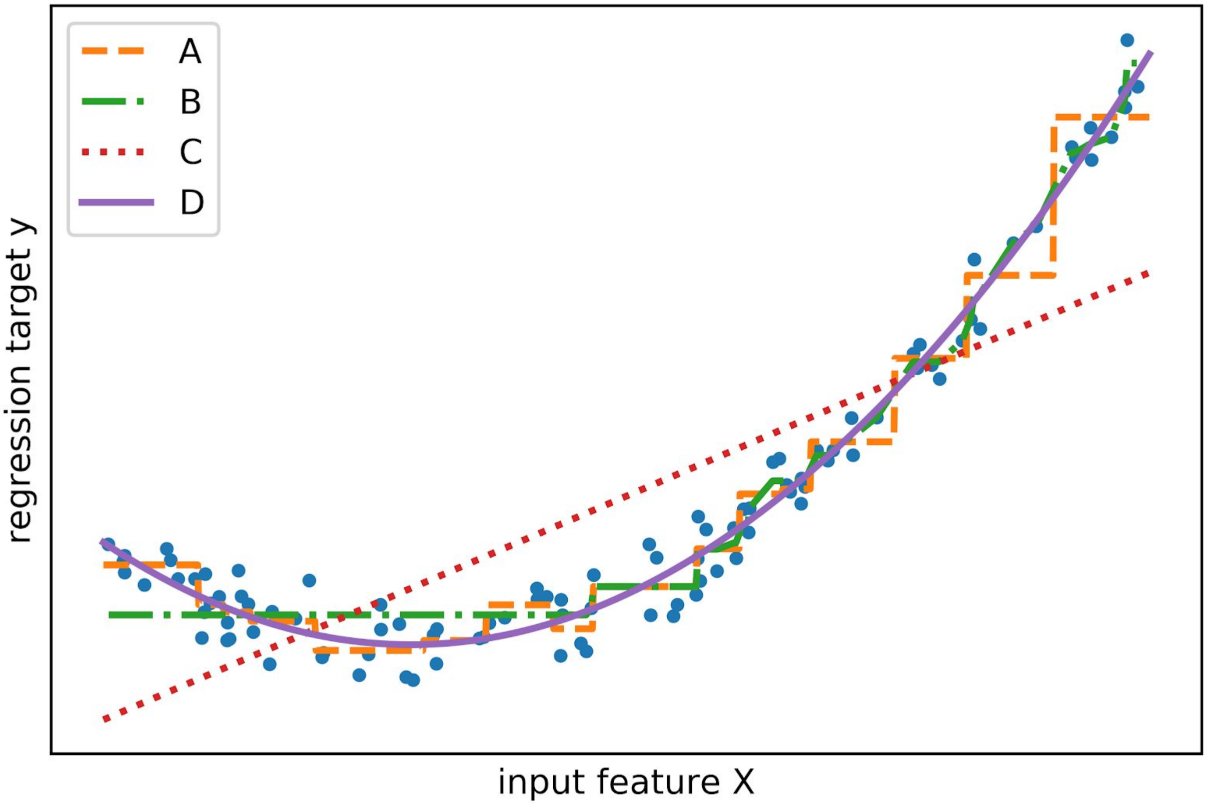
2 Multiple choice (30pt)

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

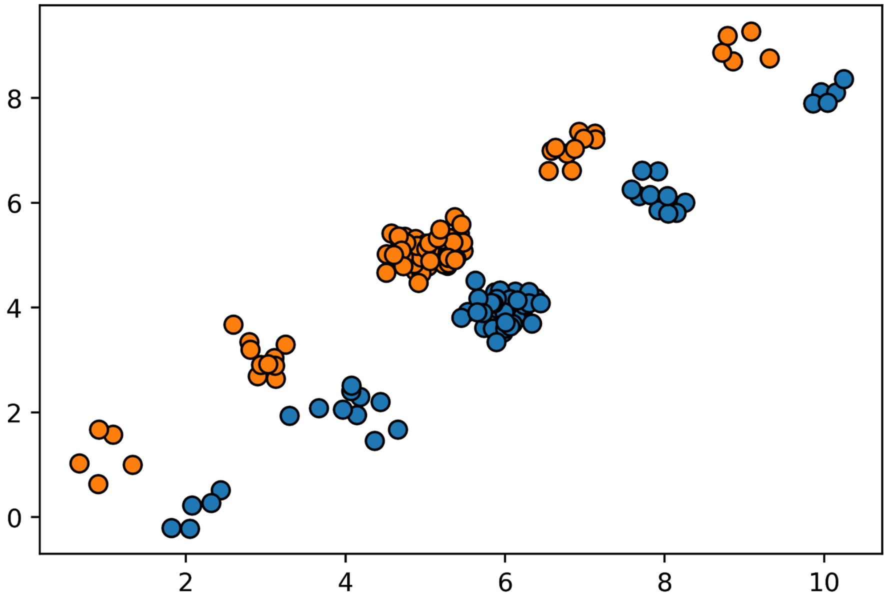
2.3 Which of the following variables should be treated as categorical?

- □ 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.

```
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) <- should use validation set
   if score > best_score:
            best_score = score
            best_C = C

svm = LinearSVC(C=best_C).fit(X_valid, y_valid)<-should use X_trainval
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.

```
pipe = make_pipeline(SimpleImputer(strategy=”mean”), OneHotEncoder(),
                                   LogisticRegression())
pipe.fit(X_train, y_train)
coef = pipe.named_steps[‘logisticregression’].coef_
important = np.argsort(coef)[-10:]
plt.barh(range(10), coef[important])
plt.yticks(range(10), feature_names[important])
```

Mean imputation is meaningless for categorical data
Argsort should use abs for important features.
Feature names don’t correspond to one-hot encoded features.

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.

```
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 don’t 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, specifying a different scoring would also be fine
grid.fit(X_train, y_train)
score = grid.score(X_test, y_test)
```

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?

Recall could be high if only predicting one class, which is not useful. Macro-average recall requires recall to be high for both classes simultaneously

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

Otherwise cross-validation can not be performed correctly over workflows that include preprocessing. Pipelines allow encapsulating the whole workflow within a single model, allowing all the learning to take place within the cross-validation loop and avoiding leaking information from the validation part.

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

Accuracy is hard to interpret and discards many important aspects of the result in imbalanced classification. Given a 99:1 imbalanced dataset, a model with 99% accuracy could either be constant or very useful, but accuracy does not allow to distinguish the two.

5.4 Explain target encoding of categorical variables.
Target encoding replaces a categorical column by the mean response in that category in the training data. For regression, this is the mean prediction, for binary classification, it is the prevalence of one of the classes, and for multi-class classification one feature is created for the prevalence of each class.
Often the prevalence is estimated using a cross-validation scheme or smoothing to avoid overfitting.