Question 13

5 pts

Flexibility/Complexity of the model:

Edit View Insert Format Tools Table

12pt Paragraph

The flexibility and complexity of a decision tree model grow as the number of terminal nodes increases. This arises because a tree with more nodes may construct more sophisticated decision boundaries that can accommodate a wide range of data patterns and circumstances.

As a result, the model's capacity to closely match the training data is improved, allowing it to capture a greater range of data features. While **this complexity allows** for a more personalized fit to the training data, it may also lead to overfitting, in which the model becomes overly specialized in training data properties at the price of generalizability.

p

101 words

Question 14

5 pts

Training Error:

Edit View Insert Format Tools Table

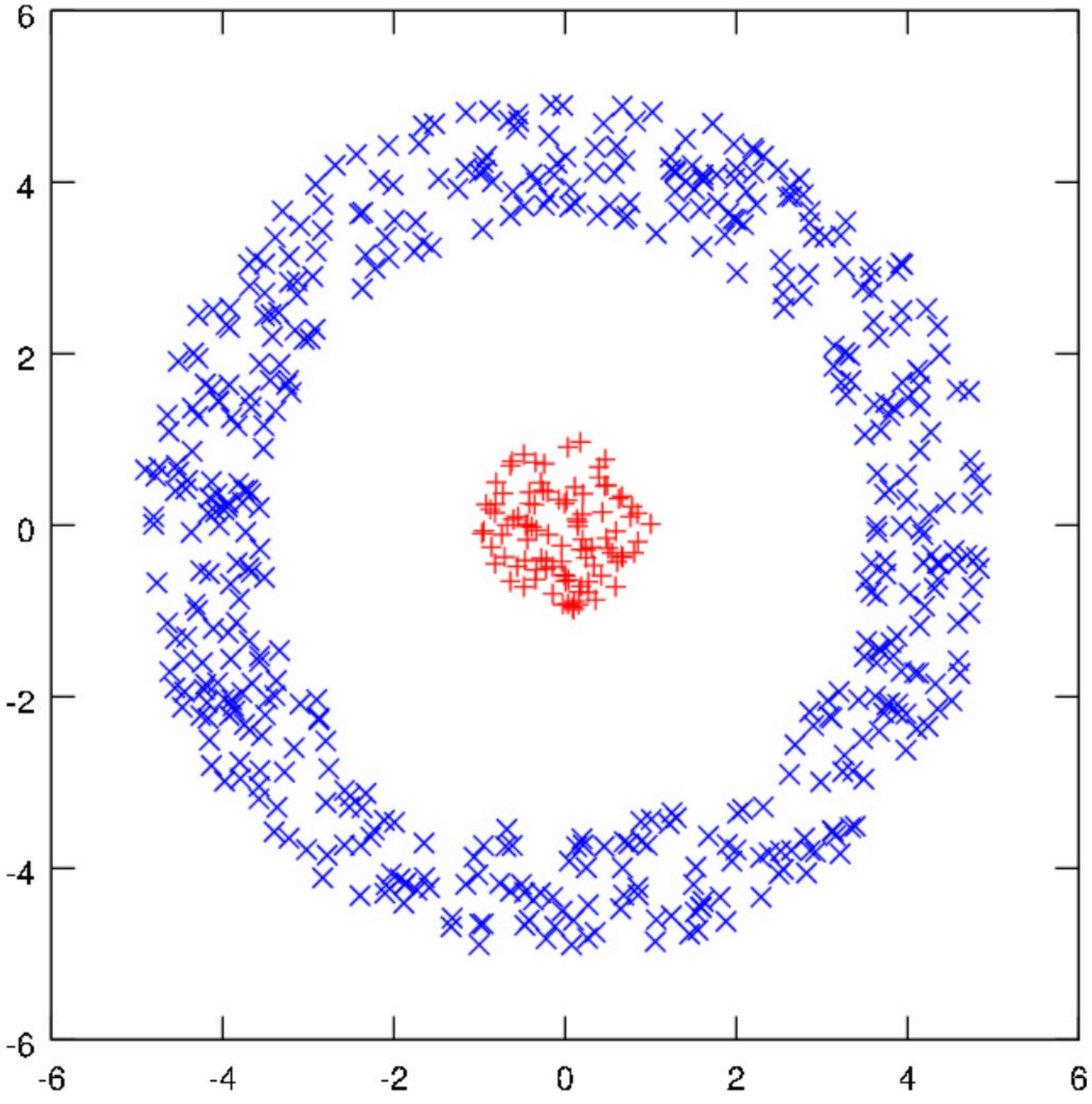
12pt Paragraph

A decision tree's **training error** normally lowers as the number of terminal nodes grows. This reduction in error is due to the tree's improved capacity to match the training data more precisely.

With more nodes, the tree becomes better at capturing particular patterns and correlations in the training data, resulting in a tighter alignment between predicted and actual values. While this improves performance on training data, it does not always transfer to higher performance on unseen test data.

Suppose you have a binary classification problem with only two predictors.

Plotting the data in the two predictor dimensions, and coloring by target variable category, yields the following:



Your goal is to come up with a way to classify future observations as one of the two categories, based on the predictor values.

For each of the following four methods, state whether you think the method would be a good

Question 15

5 pts

Test Error:

Edit View Insert Format Tools Table

12pt Paragraph

Test error exhibits an initial drop followed by a possible increase in response to a growing number of terminal nodes in a decision tree. Initially, as the tree becomes more complicated, it catches the underlying trends in the dataset better, resulting in higher accuracy on unseen test datasets.

However, once the tree reaches a certain degree of complexity, it tends to overfit the training dataset. Overfitting appears as a model that is too fitted to the details of the training data, hindering its capacity to generalize to new datasets. once a result, the initially decreasing test error begins to grow again once the model's complexity reaches the ideal threshold for generalization.

p

111 words

Question 6 — 5 pts

K-Nearest Neighbors

Edit View Insert Format Tools Table

12pt Paragraph

This method is distinguished by its flexibility and non-parametric nature, which enables it to negotiate non-linear borders that may exist between two separate categories. This property makes it particularly well-suited to tackling challenges in which linear assumptions about data connections do not hold. It is sensitivity to outliers can be a major irregularity in the data that might have a disproportionate effect on the model's performance and decision-making process. Second, the computing cost of this strategy can be significant. This is especially important when working with huge datasets or sophisticated models, when computing demand can quickly grow, thereby affecting the approach's efficiency and scalability.

p

104 words

For the remaining questions please refer to the following.

Here we have some Python code to do some of the things we talked about this term.

Consider the following code:

```
ct = ColumnTransformer(
   [
      ("step1", OneHotEncoder(sparse_output = False), ["variable1"]),
      ("step2", StandardScaler(), ["variable2", "variable3"]),
      ("step3", PolynomialFeatures(), ["variable2"])
   ],
   remainder = "drop"
)

p = Pipeline(
  [("bigstep1", ct),
   ("bigstep2", LinearRegression())]
).set_output(transform="pandas")

degs = {'bigstep1__step3__degree': np.arange(1, 10)}

gscv = GridSearchCV(p, degs, cv = 5, scoring='r2')
gscv_fitted = gscv.fit(X, y)
```

Question 16

5 pts

Describe this process (i.e. what the code is doing as a whole). Feel free to create a list of bullet points if you wish.

Edit View Insert Format Tools Table

Question 7

5 pts

Linear Discriminant Analysis

Edit View Insert Format Tools Table

12pt Paragraph

If the data meets specific criteria, LDA can be a viable solution for classification difficulties. It is assumed that the data follows a Gaussian distribution with identical covariances for each dataset and that the variances of the datasets are similar. If these criteria are fulfilled, LDA can generate a linear decision boundary that effectively divides the classes. However, if the assumptions are broken, LDA may underperform.

p

66 words

Question 1 — 10 pts

Consider the following code:

```
ct = ColumnTransformer(
   [
      ("step1", OneHotEncoder(sparse_output = False), ["variable1"]),
      ("step2", StandardScaler(), ["variable2", "variable3"]),
      ("step3", PolynomialFeatures(), ["variable2"])
   ],
   remainder = "drop"
)

p = Pipeline(
  [("bigstep1", ct),
   ("bigstep2", LinearRegression())]
).set_output(transform="pandas")

degs = {'bigstep1__step3__degree': np.arange(1, 10)}

gscv = GridSearchCV(p, degs, cv = 5, scoring='r2')
gscv_fitted = gscv.fit(X, y)
```

How many separate times is a model fitting process being run in this code?

45

Question 17

15 pts

What are more informative names for "step1", "step2", and "step3" in this code? And why?

Edit View Insert Format Tools Table

12pt Paragraph

distribution with a mean of zero and a standard deviation of one is known as standard scaling (or Z-score normalization).

Step3 associated with "poly_features_variable2" :

Assuming that "step3" and "step4" are supposed to be different stages and that there is an error in the code, "step3" applies PolynomialFeatures to "variable2". The term "poly_features_variable2" defines the generation of polynomial features for "variable2" clearly.

In machine learning, polynomial features are used to improve the model's capacity to capture more complicated interactions by taking into account not just the variable but also its powers and interaction terms.

p ▸ strong

189 words

Question 18

10 pts

What are more informative names for "bigstep1" and "bigstep2" in this code? And why?

Edit View Insert Format Tools Table

12pt Paragraph

"bigstep1" to "feature_preprocessing":

This stage includes the ColumnTransformer, which does feature preprocessing. The term "feature_preprocessing" appropriately describes how this stage prepares and transforms input features before they are fed into the model. This term conveys a clear, high-level understanding of the stage's role in the pipeline, suggesting that various transformations (such as encoding, scaling, and polynomial feature generation) are performed on various columns.

Question 8

5 pts

Support Vector Classifiers

Edit View Insert Format Tools Table

12pt Paragraph

The goal of SVC is to find a hyperplane that optimizes the margin between distinct classes. It works well when the data is separable and has kernel functions that can handle both linear and non-linear decision boundaries. SVC, on the other hand, might be sensitive to the kernel and regularization settings used. This technique has the benefit of avoiding overfitting and outliers by employing a soft margin parameter that allows for some misclassification.

p

73 words

For the next 4 questions, please refer to the following.

Recall the following quantity that we seek to minimize when fitting a Ridge Regression model:

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \hat{\beta}_j^2$$

where *i* indexes the observations, *j* indexes the predictors, and λ is a tuning parameter that we pick the value of.

Question 2

5 pts

Describe the fitted model when $\lambda = 0$.

Edit View Insert Format Tools Table

12pt Paragraph B I U A T²

λ=0 With Ridge Regression, the ridge penalty factor essentially vanishes, making the fitted model equal to a regular linear regression model. This model does not employ regularization; its only goal is to minimize the sum of squared residuals.

In this, the model will not have the regularizing effect that Ridge Regression typically provides. It means that the model will behave exactly like a linear regression model, seeking to minimize the residual sum of squares without any penalty for the size of the coefficients.

p

83 words

Question 9 — 5 pts

Support Vector Machines

Edit View Insert Format Tools Table

12pt Paragraph

This technique, an advanced extension of the support vector classifier, distinguishes itself by its ability to handle complicated and non-linear borders between categories, making it an appealing solution for difficult classification issues. This complexity, however, comes with significant trade-offs. The approach can be computationally costly, especially when dealing with huge datasets or complicated models that need substantial computer resources and time. Furthermore, the intricacy that promotes its adaptability makes it difficult to comprehend and fine-tune.

p

75 words

Question 19

5 pts

What will the gscv_fitted variable contain?

Edit View Insert Format Tools Table

12pt Paragraph

The record of the parameter values that resulted in the best model performance is crucial to gscv_fitted. This information is critical for determining the model's ideal complexity level concerning the provided data, especially the number of polynomial transformations that best reflect the underlying patterns in the data.

Performance Metrics: This variable contains a detailed summary of the cross-validation findings, including R-squared values for each parameter combination examined.

Comprehensive Cross-Validation Data: gscv_fitted offers thorough cross-validation data for each parameter combination, such as mean test scores and standard deviations.

Refitted Model on Complete Data: In gscv_fitted, the grid search process culminates with the refitting of the best model on the whole dataset. This guarantees that the final model makes use of all available data, improving prediction power and generalizability to new data.

p

220 words

Question 20

10 pts

How do we use the value of gscv_fitted to proceed in this process?

Edit View Insert Format Tools Table

12pt Paragraph

performance in real-world circumstances.

Refinement and iteration: Refine the modeling method based on the insights gathered from the grid search results. This might involve modifying the grid search parameter range, integrating extra preprocessing stages, or experimenting with other modeling methodologies.

Documentation and reporting: Keep a record of the findings, including the parameters used, model performance, and any insights gleaned from the cross-validation results. This documentation is critical for ensuring transparency, repeatability, and conveying outcomes to

Question 3

5 pts

Describe the fitted model when $\lambda = \infty$.

Edit View Insert Format Tools Table

12pt Paragraph

When λ is set to ∞ in Ridge Regression, the model is driven to zero in all coefficients, making it non-predictive and limited to producing constant or almost constant values. This case demonstrates regularization's end, in which the model completely ignores the input characteristics.

The model loses its predictive power as it does not effectively use the input features to make predictions. Instead, it produces a constant or near-constant output regardless of the input.

p

74 words

Question 10 — 10 pts

For the classification problem in the picture above, consider when might you prefer to use

(a) ROC-AUC,

(b) F1 Score, or

(c) Accuracy

as your metric to measure model success.

For each metric, give a real-data example where you would prefer that metric, and explain why this is so.

For example, you might say "If the red points represent dogs and the blue points represent cats, I prefer ROC-AUC because..."

Edit View Insert Format Tools Table

12pt Paragraph B I U A T²

a) **ROC-AUC**: When the cost of false positives and false negatives vary dramatically, you must analyze the trade-off between true positive rate and false positive rate across multiple thresholds.

Consider a credit scoring system whose purpose it is to forecast if a person would default on a loan. If the red dots reflect those who will default and the blue points represent those who will not, ROC-AUC is an appropriate statistic. This is critical in the financial industry, where erroneously categorizing a non-defaulter as a defaulter (false positive) may result in a missed opportunity, but incorrectly classifying a defaulter as safe (false negative) may result in severe financial loss.

b) **F1 Score:** The dataset is unbalanced, and both precision (the proportion of accurate positive identifications) and recall (the fraction of correct positive identifications) must be considere

p 2 307 words </>

Question 4

5 pts

Suppose we wanted to find the "best" choice of λ.

Why would it be a bad idea to simply pick the value that minimizes the MSE on the training data, i.e., the one that leads to predictions that minimize

$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

Edit View Insert Format Tools Table

12pt Paragraph

Selecting the "best λ" It is not a good idea to choose a solution based just on whatever yields the lowest Mean Squared Error (MSE) on the training data. Because it may simply be overfitting being very adept at predicting the training dataset which will become less adept at predicting unforeseen datasets.

To find the "best λ" for a certain dataset, use techniques like cross-validation, which involves training and validating the model on different subsets of the data. This process provides a more balanced view of how different values of λ affect the model's performance, leading to a choice that better balances bias and variance and improves the model's ability to generalize.

p

112 words

Question 11

20 pts

Suppose you are presented with a dataset consisting of samples from 100 individuals, and for each sample, you have gene expression measurements for 20,000 genes. Of the individuals, 50 are carriers for a rare genetic disease, while 50 are not.

You are interested in studying which genes are likely drivers of the disease.

How would you approach this? Be sure to give at least one advantage and one disadvantage of the suggested analysis.

Edit View Insert Format Tools Table

12pt Paragraph B I U A T²

interpretability of the generated models.

6) Integrative Multi-Omics Analysis: Lastly, combining RNA-Seq data with information from other omics fields, such as proteomics or genomics, might provide a more comprehensive knowledge of the illness process. The investigation of the interactions between various biological layers and how they affect the illness phenotype is made possible by this integration. The primary disadvantage in this case is the intricacy of data integration and analysis, which calls for sophisticated computational techniques and interdisciplinary knowledge.

The suggested method employs a combination of statistical, computational, and experimental methodologies to elucidate the genetic complexities of the uncommon hereditary ailment. This multi-pronged approach not only strengthens the findings' validity but also offers a thorough grasp of the disease's molecular basis.

13

p

1

473 words

</>

Question 5 — 10 pts

Explain the difference between *regression* problems and *classification* problems.
Be sure to include possible differences in
(1) modeling techniques available to us and
(2) the error/performance metrics
that are most appropriate to each set of problems.

Edit View Insert Format Tools Table

12pt Paragraph

1) modeling techniques available to us

a) Regression problems are those in which a continuous output has to be predicted. The objective is to estimate the mapping function between continuous output variables as well as input variables. Decision trees, polynomial regression, and linear regression are examples of common regression approaches. Neural networks built for regression problems, Random Forests, and Gradient Boosting are examples of more sophisticated techniques.

b) Classification problems are the classification forecasts outcomes that fall into one of two categories: discrete values, labels, or classes. Classification techniques include Decision Trees, Support Vector Machines, Neural Networks designed specifically for classification, and Logistic Regression (yes, despite its name). Classification challenges also often use deep learning techniques such as Recurrent Neural Networks for sequence data and Convolutional Neura

p

260 words

For the next 4 questions please refer to the following.

We can control the complexity of a decision tree model by adjusting how deep the tree is grown (i.e. the number of terminal nodes). Explain how the following characteristics of decision trees change **as the number of terminal nodes increases**, and why.

Question 12

5 pts

Sensitivity to training data:

Edit View Insert Format Tools Table

12pt Paragraph

A decision tree's model gets more perceptive to the subtleties and particular patterns found in the training data as the number of terminal nodes rises. This **increased sensitivity results** from the fact that a tree with more nodes can capture more precise information by drawing sharper distinctions.

Consequently, the model begins to capture the unique characteristics and noises present in the training sample in addition to the overall trends. The model's fit to the training set is improved by this enhanced sensitivity, but it may also limit the model's capacity to generalize to new, untested data.

p

96 words