Michael Dang – 16257750

COMP-SCI 5565

Midterm review

10/18/2023

1.

* Prediction: this allows us to make future predictions or forecasts based on the given data points.
* Interpolation: this allows us to understand the underlying behavior of data points, and how each data point relates to each other.

2.

* Steps:
* Collect the data, then do data cleaning/feature engineering/preprocessing.
* Modeling fitting. Fit the model into the data, for linear regression: y = aX + b + e, for multiple linear regression: y = aX + bY + cZ + … + e.
* Calculate the residual and then minimize the error. Validation.
* Add regularization, because linear regression is considered an ‘ill-conditioned’ since the condition number for linear regression is too large, hence it converges very slowly. So, we add an L1 or L2 regularization term (from Numerical Analysis Optimization).
* Tools:
* Rstudio
* Python
* Metrics:
* RSS
* R\_squared / R\_squared adjusted.
* MAE
* MSE

3.

a. Overfitting

b.

* Dection: Validation set, visualization.
* Avoidance: Feature selection, regularization, CV, hyperparameter tunning.

4.

a. This is the estimated or predicted value.

b. It helps us to classify the predicted value vs the actual value.

c. Yes

5.

a. During training, the model learns from the training data, hence we can adjust and reduce the errors (MSE, MAE, R\_squared). The goal is to fit the model to the training data at the optimum value, i.e., not poorly and not perfect in between. Then in the testing, we will use that model to perform on the unseen data to see how well it deals with new data.

b. Because it will make the model ‘overfitting’.

c. Evaluating a model using testing data is crucial for understanding how well it generalizes.

d. The R\_squared = 1 – (SSE/TSS), based on the formula we can see that the SSE calculates the distance between the true value and the predicted value, and the TSS calculates the distance between the true value and mean. i.e., it is the ratio between the explained variance and total variance. (0 < R\_squared < 1) the model with the R\_squared closer to 1 indicates a good fit and small residual and vice versa.

6.

* To see if there are correlations between features and if the features are highly correlated to each other so we can apply dimensionality reduction (PCA, truncated SVD, etc) to capture the most meaningful data.

7.

a. Regression. Since we predict the continuous target variable based on other features. There are 500 samples (n = 500) and there are 5 features, budgets, profits, number of employees, industry, and CEO salary (p = 5). The strongest predictor is likely the company's profit.

b. Classification. Since we predict the discrete target variable (0 or 1) based on their feature. There are 20 samples (n = 20), and there are 9 features, including price charged, marking budget, competition price, time to market, total development hours, space allocated, total personnel need, long-term support, and customer rating of satisfaction. The strongest predictor is likely to be the customer rating of satisfaction.

c. Clustering. Since we want to factor the users into the group based on their characteristics, without a target. There are 1000 samples (n = 1000) and there are 3 features, including subscribers, hours per week, and generating/responding to messages (p = 3). The strongest differentiating factor is likely whether the user is a paid subscriber or not.

8.

a. Qualitative data

b. Since there are X1 and X2, hence p = 2.

c. Increasing n would like to improve the model performance. With more samples, the boundaries between groups can be better defined.

d. I would argue that if we increase p, the dataset might fall into the ‘curse of dimensionality’ hence it would resolve more noise and make visualization difficult.

e. I’m a big fan of unsupervised learning, even though it’s harder. So I would go for unsupervised learning.

9.

a. The average residual error. This value measures the distance between the actual value and the predicted value.

b. Sample sizes

c. No this only includes the reducible error because it is missing the term + Var(e) which assumes iid with N (0, sigma^2).

d. No, RSE depends on the degree of freedom of the parameter, i.e., it is the standard deviation of the residuals meanwhile the equation above is MSE (the average square errors).

10.

Confusion matrix or misclassification matrix. Metrics cost analysis, i.e., to see the False Positive and False Negative.

11.

The False Negative is 5 since by looking at the graph there are 5 triangles that’s been misclassified.

12.

It describes the correlation between each feature in the data. Correlation coefficient or correlation matrix. Yes, R\_squared depends on RSE and RSS, as shown in the formula:

R\_squared = 1 – (RSS/TSS) and RSS = (n – 2) \*RSE^2

13.

a. Probabilities outside (0,1) range, logistic converts to (0,1)

b. Thresholding functions for converting the continuous outputs of a model into discrete class predictions. A threshold, usually 0.5, is used to classify values above the threshold as 1 class and below as other.

c. Intercept, coefficient, threshold value.

14.

When we there 2 < classes < 5 and when the classes are linearly separable.

15.

LDA is also known as multivariate Gaussian (STAT5572 – Multivariate Analysis) and uses Bayes’ theorem to compute discriminant functions.

1. For a k class LDA, the general rule is that k – 1 discriminant functions are required to discriminate between the classes.
2. The discriminant functions in LDA act similarly to step functions in that they divide the discriminant space into regions associated with each class and create step-like separation between classes in the reduced dimensionality space.

16.

The prior distribution, i.e., the pre-existing probability of an event or class before observing the data.

1. Pi (.

17.

Likelihood ratios relate the prior and density to the evidence in order to compute the posterior probability estimate from Bayes’ theorem.

18.

The main drawback of Naïve Bayes for multi-class systems is that it assumes the features are conditionally independent of each other.

1. The algorithm generates a k separate model, one for each class, for a k class system.
2. Assumes features are independent in each class.

19.

K-fold cross-validation is used to estimate the performance of the model. i.e., K-fold helps to find the optimal model parameters and provides a realistic estimate of model performance on new data.

1. K-fold averages the performance across folds to get an overall estimate of model generalization errors.
2. LOOCV, each test fold contains just a sample, hence the performance depends heavily on that individual sample’s properties.
3. Computational costs, high variance with small dataset.
4. Validation methods assume the distribution of the resampled subsets is representative of the original sample distribution.