PES University, Bengaluru

UE18CS312 - Data Analytics

List of concepts that should be known for any test/ exam on UE18CS312

80-85% TC: Direct application of concepts taught in class
15-20% GS: General skill (how much can students extrapolate and analyze based on what has been taught?)

No direct questions on coding (either writing or interpreting code) in the written test/ exam

Unit 1

1. [GS] Given a problem (for example, factors that influence performance of students in various subjects in the eighth standard), what data can be collected and how? (Data: marks in various subjects (from the class register), number of people in the house, educational background and occupation, etc. (from Census data), #hrs of tv, #play, etc., from an actual survey, etc.
2. [GS - descriptive] Given some data, posing problems on that – (in the above example, economic health of various regions can be inferred from the education, occupation and income data)
3. [TC] Identifying the format of data (numeric, ordinal, nominal OR continuous vs discrete OR ratio vs interval, etc.) and operations allowed such as on data cube, etc.
4. [TC] Given data, basic stocktaking – summary of the data, finding outliers using box-and-whisker plots, identifying any data that is incorrect or inconsistent, finding missing data and suggesting ways to fill this in with least error
5. [TC] Checking for independence of variables using chi-square; use of hypothesis testing to compare significance of impact of one factor (such as dosage of a drug, etc.) using t-test (from IDS and Chapter 6 of textbook)
6. [TC] Data transformation – when, why and how? Problem on given data in a range, mapping it into a new range
7. [TC] Data reduction – dimensionality reduction using wavelets and PCA (application level) and feature subset selection and generation
8. [TC] Data reduction – numerosity reduction using binning, histogram and sampling (understanding of different sampling techniques)
9. [TC] Data cube operations and detection and removal of duplicate records when integrating data
10. [TC] Different forms of visualization (given data, what is an appropriate visualization technique to get insights to the patterns in this data?)
11. [GS] Given a visualization, what inferences can be made from the graph, plot, etc.?
12. [TC] Effect of interval width on histogram interpretation – modality of Gaussian function through density approximation of the histogram and descriptors such as skew and kurtosis

Unit 2

1. [TC] Given some data (10 points or so), figuring out the most appropriate correlation measure to be used and calculating correlation coefficient
2. [TC] Interpreting the correlation coefficient
3. [GS] Correlation vs causation and nature of relationship given some scenario

4. [TC] Simple linear regression – calculation of parameters given a small data set
5. [TC] Assumptions for linear regression and evaluation of results
6. [TC] Assumptions for multiple regression and computing redundancy of features using correlation
7. [TC] Comparing multiple models and selecting one of them
8. [TC] Bias/ Variance trade off and feature selection using ridge/ lasso
9. [TC] **Compute** Mahalanobis distance, Minkowski's distance (Manhattan or Euclidean) between data points, compute R, R2; **interpret** R2, adjusted R2, Cp-Mallows, Cook's distance, DFFit and DFBeta, Leverage, t statistic (significance), F statistic, Durbin-Watson statistic, AIC/ BIC
10. [TC] Odds, odds ratio and logistic regression – simple calculation (such as parameters or probability, given the parameters or a prediction given the data)
11. [TC] Confusion matrix – making entries given some data and calculating Accuracy, Recall, Precision, sensitivity, specificity, Youden's index, F1 score, etc.
12. [TC] Selection of a model based on AUC of RoC; RoC vs Confusion Matrix for evaluation
13. [GS] Interpretation of R2, SSE, etc., and residual plots
14. [GS] Given a small problem, being able to set up the logistic regression solution path (i.e., identify the need for transformation of data and explain the computation of parameters, etc., or given the coefficients compute probability/ given the predictor and probability, compute odds and infer B1 (as change in ln odds) and B0 (as avg(y)-B1)avg(x)) or given B0 and B1 come up with a prediction)

Unit 3

1. [TC] Given a sample signal being able to recognize if it is additive or multiplicative and what its components are (sketch schematically by hand)
2. [GS] Identifying components such as seasonality, cyclic, etc.
3. [TC] Given a signal, performing simple calculation for predicting future demand or forecasting with exponential smoothing (alpha =1 type simple case that can be worked by hand)
4. [TC] Assumptions for stationarity and methods to convert a nonstationary signal to a stationary one
5. [TC] Given an averaging filter (simple, weighted), computing the moving average to find a forecast
6. [TC] Given alpha, compute single exponential smoothing forecast and computing initialization for level and trend for Holt and Holt-Winter's methods, compute seasonality index.
7. [TC] Be able to answer (theory) of Double and triple exponential smoothing and Croston's forecast (importance, when this can be used, how it works, advantages/ limitations)
8. [TC] Given a small data set, using regression to compute a forecast or interpreting the results of regression for forecasting
9. [TC] Choice of parameters for AR, MA. ARMA model based on ACF and PACF
10. [TC] Given a small data set, making predictions with small order models and computing the error (MAE, MSE, etc.)
11. [TC] Given an expression (like ARIMA(1,1,1) being able to write the corresponding forecasting equation)

12. [GS] Being able to suggest a solution approach for a small/ hypothetical modeling/ forecasting problem
13. [TC] Interpreting results of a DF or ADF test for stationarity and methods to convert non stationary signals to stationary
14. [TC] Brief note on Ljung box and Theil's coefficient
15. [TC] Spectral analysis – application or basic idea of computing the Fourier transform of a signal
16. [GS descriptive] What can be used for feature extraction in a time series signal? Or what sort of features can be extracted using wavelets in a time series signal? (If there is an MCQ question (on quiz, etc.), it would probably be output of a 2 channel filter bank with a simple Haar filter (of size 2))
17. Box Jenkins methodology and use of SARIMA and ARIMAX for forecasting

*****