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] **Compute** the Pearson's correlation coefficient and **compute** the Phi Coefficient and **interpret** 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 (accuracy of prediction, various properties of the residuals, etc.)
6. [TC] Assumptions for multiple regression and computing redundancy of features using correlation
7. [TC] Comparing multiple models and selecting one of them based on various parameters
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, ARIMA model based on ACF and PACF and analysis of the signal (does it need differencing? how much?)
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] A brief note on the Ljung box and Theil's coefficient – what are these used for? How are the tests applied? How are the results interpreted?
15. [TC] Spectral analysis –application or basic idea of computing the Fourier transform of a signal
16. [GS] 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 anMCQ 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

Unit 4

1. [GS] Given a scenario, to be able to identify the most appropriate recommendation technique to apply and give the rationale
2. [GS] Given a scenario to be able to identify what sort of a recommendation system this is (constrained-based, case-based, etc.) and suggest the workflow
3. [TC] Compute user-user or item-item similarity using cosine, Jaccard or SMC or distances such as Euclidean, Manhattan or correlation coefficient
4. [TC] To use collaborative filtering to predict the rating for a movie (as a weighted average of others' ratings that are available)
5. [TC] Finding the best question to ask (or feature to test) at a node based on classification error, Gini index, entropy or information gain. Similarly, to be able to decide whether it is worth splitting a node based on a feature (i.e., comparing entropy or Gini for child nodes with parent node)
6. [TC] Deciding the best split for a categorical or continuous variable
7. [TC] Compute hierarchical clustering (single linkage, complete linkage, average linkage, centroid method) and draw the dendrogram, kmeans (for a given set of initial points) or DBSCAN clustering for given data
8. [TC] How can we identify noise points or outliers using DBSCAN or hierarchical clustering? (varying parameters to zero in on points that are repeatedly marked as 'noise points' or applying thresholds to distance measures in the dendrogram to identify outliers)
9. [GS] Use of clustering to find similar documents or statements (social media postings, for example)
10. [TC, GS] Pros and cons of various clustering and classification approaches and boosting and bagging – when can we use a particular method? Given a scenario, what method may be most suitable?
11. [GS] Use of cross-validation and accuracy measures to determine the parameters for knn, ANN, SVM; objective measures (such as SSE) in case of determining the most suitable clustering configuration for DBSCAN, kmeans
12. [TC] Use of bagging and boosting to improve accuracy (differences between or pros and cons of using bagging, boosting)
13. [TC] Given some text data, being able to identify the preprocessing steps, compute features such as TF, IDF and TF-IDF and apply simple classification techniques (such as Naïve Bayes) to solve a problem

14. [GS] Given transaction data, to be able to find frequent k-itemsets and arrive at reasonable association rules for a given minsup or minconf
15. [TC] Computing support or confidence for a given data set or association rules
16. [TC] Being able to come up with a contingency table for a word problem or given a contingency table to be able to compute probabilities such as P(tea | coffee) etc., and determine what rule is most 'interesting' (i.e., to measure interestingness or lift)
17. [TC] Evaluating a recommender systems (various measures and what they mean)
18. [GS] Designing a recommender system (or suggest an ML pipeline for a given scenario) that relies on data-driven approaches or leverages automation for parameter tuning, etc.

Unit 5

1. [TC] LSA – steps involved in the process: preparation of data (preprocessing techniques + computing TF-IDF, preparing a term-document matrix, SVD computation and finding correlations, finding top keywords for a document or finding the documents that are most similar to each other based on LSA and computation of cosine similarity or Pearson's correlation coefficient).
2. [GS] When and where can it be used: Given a problem, being able to delineate preprocessing steps for that problem and how LSA can be used to compute similar documents or find topics given the document
3. [TC] How can sparse data be handled? Different data representations – pros and cons for dense vs sparse representations. Being able to delineate the steps for sparse PCA.
4. [TC] Write a note on hidden and confounding variables – be able to recognize dependent, independent, hidden and confounding variables in an experiment setting.
5. [TC] What is the bias that a confounding variable can create in a regression problem?
6. [TC] How can experiments be designed to overcome confounding/ hidden variables?
7. [GS] AB Testing – given a problem how will you set up a test?
8. [TC] Application of modeling with Poisson process – computing expected demand (number of customers, spare parts, etc.) in a given duration
9. [TC] Compound Poisson process – being able to identify variables and computing mean and variance for a compound Poisson process
10. [TC] Markov chains – be able to draw the state transition diagram from a probability matrix or to derive the transition probability matrix given a state transition diagram or word problem
11. [TC] Compute the state of a vector after n transitions (given the transition probability matrix), compute the stationary distribution for a matrix, check whether a matrix is regular or not (only 3-4 powers max so it can be completed in reasonable time)
12. [TC] Given some data, writing the transition probability matrix in the Canonical form (identifying **I**, **0**, **Q** and **R**), computing the fundamental matrix (i.e., $(\mathbf{I}-\mathbf{Q})^{-1}$) and compute probability of eventual absorption (**FR**) and expected time to absorption (**Fc**).
13. [TC] Being able to write the system of linear equations for computing the expected time to reach a state j from a state i
14. [TC] Compute customer lifetime value given the steady state retention probability, distribution of customers at the current time and margins generated by customers
15. [GS] Given a scenario, being able to identify key performance indices for a business or asking questions that can bring business value or suggesting how an A/B test can be set up to measure value or identifying any mistake in the workflow to measure/ interpret business value

*****