Environmental Data Science Concept Checklist

December 9, 2021

DISCIPLINE
Topic
Subtopic
☐ Core Concept
☐ Specialized Concept

MATHEMATICS

Calculus

- ☐ Derivatives
- ☐ Integration
- ☐ Polar coordinates
- ☐ Complex numbers
- ☐ Gradient
- ☐ Limits
- ☐ Sequences
- ☐ Series
- ☐ Multiple integrals
- ☐ Vector calculus
- ☐ Partial derivatives
- ☐ Differential equations
- ☐ Jacobian matrix
- ☐ Hessian matrix

Linear Algebra

- ☐ Systems of linear equations
- ☐ Vectors
- ☐ Matrix multiplication
- ☐ Projection
- ☐ Inner products
- ☐ Outer products
- ☐ Trace, rank, and transpose
- ☐ Linear independence
- ☐ Linear transformations
- ☐ Determinant
- ☐ Matrix inversion
- ☐ Change of basis
- ☐ Singular values
- ☐ Eigenvalues and Eigenvectors
- ☐ Orthogonality

MACHINE LEARNING & STATISTICS

Probability & Statistics

Probability Theory

- ☐ Set theory
- ☐ Sample spaces
- ☐ Axioms of Probability
- ☐ Combinatorics
- ☐ Conditional probability
- ☐ Correlation
- ☐ Covariance
- ☐ Expected value
- ☐ Mean, median, standard deviation, and variance
- ☐ Independence
- ☐ Order statistics

Random Variables

- ☐ Discrete and continuous distributions
- ☐ Probability mass/density function (PDF)
- ☐ Cumulative distribution function / hazard function
- ☐ Joint density
- ☐ Moment generating function
- ☐ Characteristic function

Discrete Probability Distributions

- ☐ Bernoulli
- ☐ Binomial
- ☐ Geometric
- ☐ Poisson
- ☐ Negative binomial
- ☐ Hypergeometric

Continuous Probability Distributions

- ☐ Normal/Gaussian
- ☐ Uniform
- ☐ Exponential
- ☐ Chi-squared
- ☐ Student's t
- ☐ Weibull
- ☐ Beta
- ☐ Gamma
- ☐ F

Joint Probability Distributions

- ☐ Multinomial
- ☐ Multivariate normal
- ☐ Dirichlet
- ☐ Wishart

Density Estimation

- ☐ Mixture Model
- ☐ Gaussian Mixture Model
- ☐ Kernel Density Estimation
- ☐ Parzen Window

Goodness-of-fit Tests

- ☐ All parameters known
- ☐ All parameters unknown
- ☐ Analysis of Variance
- ☐ Multiple comparisons (Tukey's Method)
- ☐ F-test

Hypothesis Testing

- ☐ Type I and Type II Errors
- ☐ Likelihood ratio test
- ☐ Generalized likelihood ratio
- ☐ Two sample t test
- ☐ Power of the Test

Estimation

- ☐ Degrees of Freedom
- ☐ Interval estimation
- ☐ Maximum Likelihood
- ☐ Method of Moments
- ☐ Minimum variance estimators
- ☐ Sufficient Statistics

Bayesian Statistics

- ☐ Bayes' Theorem
- ☐ Conjugate prior
- ☐ Evidence approximation
- ☐ Non informative priors

Nonparametric Statistics

- ☐ Friedman Test
- ☐ Kruskal-Wallis Test
- ☐ Sign Test
- ☐ Testing of Randomness
- ☐ Wilcoxon test

Information theory

- ☐ Mutual information
- ☐ Entropy
- ☐ Kullback-Leibler divergence

Variable / Feature Selection

Resampling Methods

- ☐ Bootstrap
- ☐ K-fold cross validation
- ☐ Leave one out cross validation
- ☐ Markov Chain Monte Carlo (MCMC)
- ☐ Gibbs sampling
- ☐ Jackknife

Dimensionality Reduction

- ☐ Curse of dimensionality
- ☐ Principal components regression
- ☐ Partial least squares

Subset Selection

- ☐ Best subset selection
- ☐ Stepwise selection

Regularization / Shrinkage

- ☐ Lasso
- ☐ Ridge Regression

Supervised Learning

Linear Regression

- ☐ Least Squares
- ☐ Confidence intervals
- ☐ Correlation
- ☐ P-value
- ☐ R squared statistic
- ☐ Residual
- ☐ t-statistic

Nonlinear Regression

- ☐ Polynomial Regression
- ☐ Nonparametric regression
- ☐ Generalized additive models
- ☐ Generalized linear model
- ☐ Regression Splines
- ☐ Smoothing Splines
- ☐ Local regression
- ☐ Fixed effects model
- ☐ Random effects model
- ☐ Mixed effects model
- ☐ Basis Functions
- ☐ Step Functions

Performance Evaluation

- ☐ Sensitivity
- ☐ Specificity
- ☐ Test and Training Error
- ☐ Bias/Variance Tradeoff
- ☐ Confusion Matrix
- ☐ Receiver Operating Characteristic (ROC) curve

Decision Theory

- ☐ Likelihood Ratio Test
- ☐ Minimax criterion
- ☐ Committees
- ☐ Decision fusion

Density Estimation

- ☐ Mixture Model
- ☐ Gaussian Mixture Model
- ☐ Kernel Density Estimation
- ☐ Minimax criterion
- ☐ Parzen Window

Graphical Models

- ☐ Markov Models
- ☐ Hidden Markov Models
- ☐ Bayesian Belief Network
- ☐ Markov Random Fields

Other Classification Methods

- ☐ K Nearest Neighbors
- ☐ Linear Discriminant Analysis
- ☐ Fisher's linear discriminant
- ☐ Bayes Classifier
- ☐ Naïve Bayes Classifier
- ☐ Quadratic Discriminant Analysis
- ☐ Partial Least Squares
- ☐ Discriminant Analysis
- ☐ Fuzzy Classification
- ☐ Probit model

Other Regression Methods

- ☐ Logistic Regression (Logit model)
- ☐ Multinomial Logistic Regression
- ☐ Multiple Logistic Regression
- ☐ Multinomial Logistic Regression
- ☐ Multiple Logistic Regression
- ☐ Relevance Vector Machines
- ☐ Multiple Linear Regression

Neural Networks

- ☐ Perceptron
- ☐ Error Backpropagation
- ☐ Feed-forward network functions
- ☐ Recurrent Neural Networks

Support Vector Machines

- ☐ Kernel Functions
- ☐ Maximal Margin Classifier
- ☐ Support Vector Classifier
- ☐ Separating hyperplane
- ☐ One versus all classification
- ☐ One versus one classification
- ☐ Polynomial kernel
- ☐ Radial kernel
- ☐ SVMs with more than 2 classes

Ensemble Methods

- ☐ Bagging
- ☐ Boosting
- ☐ AdaBoost
- ☐ Stacking
- ☐ Bayesian Model Averaging

Classification and Regression Trees (CART)

- ☐ Decision Trees
- ☐ Gini Index
- ☐ Out of Bag Error Estimation
- ☐ Tree Pruning
- ☐ Random Forests

Unsupervised Learning

Component Analysis

- ☐ Dimensionality Reduction
- ☐ Factor Analysis
- ☐ Principal component analysis
- ☐ Proportion of Variance Explained
- ☐ Independent component analysis
- ☐ Kernel Principal Component Analysis
- ☐ Low-dimensional representations and Multidimensional scaling
- ☐ Nonlinear component analysis
- ☐ Self-organizing maps

Clustering

- ☐ K-means Clustering
- ☐ Hierarchical clustering
- ☐ Mean Shift
- ☐ Agglomerative hierarchical clustering
- ☐ Dendrograms
- ☐ Dissimilarity measures
- ☐ Expectation Maximization
- ☐ Inversion
- ☐ Linkages (complete, single, average, centroid)
- ☐ On-line clustering
- ☐ Stepwise-optimal hierarchical clustering

Model Selection & Evaluation

Performance criteria

- ☐ Adjusted R squared
- ☐ Akaike Information Criterion
- ☐ Bayesian Information Criterion
- ☐ Mallow's Cp
- ☐ Variance Influence Factor

Common Data Challenges

- ☐ Collinearity
- ☐ Multicollinearity
- ☐ Outliers
- ☐ High Leverage Points
- ☐ Heteroscedasticity

Selection techniques

- ☐ Forward Selection
- ☐ Backward Selection
- ☐ Mixed Selection

Time Series Modeling

Characteristics of Time Series

- ☐ Autocorrelation
- ☐ Cross-correlation
- ☐ Stationarity
- ☐ Partial Autocorrelation

Spectral Analysis and Filtering

- ☐ Fourier Analysis / Fourier Transform
- ☐ Spectral Density
- ☐ Smoothing
- ☐ Periodogram
- ☐ Nonparametric Spectral Estimation
- ☐ Wavelets

Time series models

- ☐ Autoregressive Models (AR)
- ☐ Moving Average Models (MA)
- ☐ Autoregressive Moving Average Models (ARMA)
- ☐ Autoregressive Integrated Moving Average (ARIMA)
- ☐ Seasonal ARIMA

- ☐ Generalized Autoregressive Conditional Heteroskedasticity (GARCH) Models
- ☐ Multivariate Autoregressive moving average with exogenous inputs (ARMAX) Models
- ☐ Lagged regression models
- ☐ State-space Models
- ☐ Dynamic linear models with switching

Other Concepts

Other Statistical Learning Approaches

- ☐ Reinforcement Learning
- ☐ Online Learning
- ☐ Kernel Methods
- ☐ Anomaly Detection
- ☐ Multiple Instance Learning
- ☐ Bag of words model
- ☐ Network analysis
- ☐ Recommender systems

Principles

- ☐ No free lunch theorem
- ☐ Occam's Razor
- ☐ No Silver Bullet

Programming

Basic Concepts & Syntax

- ☐ Data types
- ☐ Arrays
- ☐ File Input/Output
- ☐ Functions
- ☐ Logic and conditionals
- ☐ Loops
- ☐ Math and assignment operators
- ☐ Random number generation
- ☐ Regular Expressions

Languages

- ☐ Python
- ☐ R
- ☐ MATLAB
- ☐ Shell scripting (e.g., Bash)
- ☐ Julia
- ☐ Mathematica
- ☐ C/C++
- ☐ FORTRAN
- ☐ IDL

Numerical Analysis

- ☐ Difference equations
- ☐ Interpolation
- ☐ Extrapolation
- ☐ Methods for solving linear and nonlinear systems of equations
- ☐ Monte Carlo methods
- ☐ Numerical integration
- ☐ Fourier analysis and spectral methods

Optimization

- ☐ Gradient Descent
- ☐ Linear Programming
- ☐ Lagrange Multipliers
- ☐ Boltzmann Learning
- ☐ Boltzmann networks
- ☐ Evolutionary methods
- ☐ Genetic algorithms
- ☐ Graphical models
- ☐ Simulated annealing
- ☐ Stochastic methods

Version Control

- ☐ Git
- ☐ Branch
- ☐ Clone
- ☐ Commit
- ☐ Merge
- ☐ Push
- ☐ Pull

Web Programming

- ☐ Application Programming Interface (API)
- ☐ Markdown language
- ☐ CSS
- ☐ HTML
- ☐ JavaScript
- ☐ JSON
- ☐ Scalable Vector Graphics (SVG)
- ☐ XML
- ☐ LaTeX
- ☐ Model View Controller (MVC) architecture

Web Scraping

- ☐ DOM parsing
- ☐ HTML parsing
- ☐ Computer vision web-page analyzers
- ☐ Semantic annotation recognition

Databases

Relational Databases

- ☐ SQL
- ☐ Schema
- ☐ Queries
- ☐ Insert, Update, Select, Delete
- ☐ Joins
- ☐ Indexes
- ☐ Integrity constraints
- ☐ Authorizations
- ☐ Transactions
- ☐ Triggers
- ☐ Views

Big Data

- ☐ Distributed File Systems (i.e. Hadoop)
- ☐ Map Reduce
- ☐ NoSQL
- ☐ Extract, Transform, Load (ETL)

Paradigms

Object-oriented programming

- ☐ Class
- ☐ Inheritance
- ☐ Methods
- ☐ Properties

Other approaches

- ☐ Functional programming
- ☐ Imperative programming

Natural Language Processing

- ☐ Optical character recognition (OCR)
- ☐ Grammatical Inference
- ☐ Parsing
- ☐ Part-of-speech tagging
- ☐ Sentiment analysis
- ☐ Topic segmentation

Visualization

Theory

- ☐ Color theory
- ☐ Gestalt Principles
- ☐ Small multiples
- ☐ Data density
- ☐ Data-Ink Maximization
- ☐ Human visual perception

Techniques and Styles

- ☐ Correlation analysis
- ☐ Deviation analysis
- ☐ Distribution analysis
- ☐ Multivariate analysis
- ☐ Time series analysis
- ☐ Stacked time series
- ☐ Geo-spatial analysis
- ☐ Mapping
- ☐ Part-to-a-whole
- ☐ Rankings

Datasets

Remote Sensing

- ☐ Landsat-7 & Landsat-8
- ☐ MODIS Terra & Aqua
- ☐ Sentinel-2
- ☐ ICESat & ICESat-2

Re-Analysis

- ☐ ERA-5
- ☐ GLODAP

Other

- ☐ CMIP5/6 output

Formats

Vector Data

- ☐ ESRI Shapefiles

Raster Data

- ☐ GeoTIFF
- ☐ NetCDF

Tabular Data

- ☐ CSV
- ☐ TSV
- ☐ XLSX

Other

- ☐ JSON
- ☐ XML

Software & Tools

- ☐ MS Excel
- ☐ ArcGIS (or QGIS)
- ☐ Climate Data Operators (CDO)
- ☐ NetCDF Opererators (NCO)

Workflow Management

- ☐ Snakemake
- ☐ Continuous Integration (CI)

—inspired by Kyle Bradbury's Data Science Concept Checklist)