MATERIALS

for

REVIEW

- Python
  (a) objects
  (b) namespaces and modules
  (c) simple types and containers
  (d) control flow

Additional References:

1. `https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1194/readings/python-review.pdf`
2. `https://www.learnpython.org/`

- Numeric Python (NumPy)
  (a) lists vs. Numpy arrays
  (b) universal functions
  (c) vectorized computations
  (d) broadcasting
  (e) sequences with *linspace*() and *arange*()
  (f) sorting and searching
  (g) vectors and matrices
  (h) distances (Euclidean, street, Minkowski)
  (i) inner products and cosine similarity
  (j) statistical functions
  (k) properties of mean and variance
  (l) covariance and correlation

Additional References:

1. `https://www.python-course.eu/numpy.php`
2. `https://www.kdnuggets.com/2017/08/comparing-distance-measurements-python-scipy.html`

3. https://www.kdnuggets.com/2019/06/optimization-python-money-risk.html
4. https://www.kdnuggets.com/2019/06/speeding-up-python-code-numpy.html

- Numpy indexing and slicing
  (a) reshaping
  (b) transposing arrays
  (c) element selection
  (d) row(s), column(s) selection
  (e) row(s), column(s) slicing
  (f) row(s), column(s) reversing

Additional References:

1. https://www.tutorialspoint.com/numpy/numpy_indexing_and_slicing.htm
2. https://machinelearningmastery.com/index-slice-reshape-numpy-arrays-machine-learning-python/
3. https://www.w3schools.com/python/numpy_array_slicing.asp

- Pandas overview
  (a) *Series* object
  (b) broadcasting
  (c) Pandas *DataFrame*
  (d) column creation, indexing, sort
  (e) *head*() and *tail*() functions
  (f) data selection (label, position)
  (g) lambda functions
  (h) filtering, counting, aggregation

Additional References:

1. `https://www.kdnuggets.com/2018/10/beginner-data-visualization-exploration-using-pandas-beginner.html`
2. `https://www.kdnuggets.com/2020/04/stop-hurting-pandas.html`
3. `https://www.kdnuggets.com/2019/04/pandas-dataframe-indexing.html`
4. `https://www.kdnuggets.com/2020/04/python-data-analysis-really-that-simple.html`

- Pandas graphics
  (a) histogram
  (b) scatter plot
  (c) density
  (d) multi-variate density
  (e) counting
  (f) bar and violin plots
  (g) pair plots

Additional References:

1. `https://datatofish.com/plot-dataframe-pandas/`
2. `https://www.kdnuggets.com/2018/06/7-simple-data-visualizations-should-know-r.html`
3. `https://www.kdnuggets.com/2019/04/data-visualization-python-matplotlib-seaborn.html`

- Model Concepts and Definitions
  (a) prediction vs. classification
  (b) numerical vs. categorical data
  (c) loss function
  (d) testing and training data
  (e) bias elimination and stratification
  (f) cross and k-fold validation
  (g) bias vs. variance trade-offs

Additional References:

1. `https://www.kdnuggets.com/2018/04/supervised-vs-unsupervised-learning.html`
2. `https://www.kdnuggets.com/2016/09/data-science-basics-3-insights-beginners.html`
3. `https://www.kdnuggets.com/2016/05/machine-learning-key-terms-explained.html`
4. `https://www.kdnuggets.com/2019/06/choosing-error-function.html`
5. `https://www.kdnuggets.com/2020/05/dataset-splitting-best-practices-python.html`

6. https://www.kdnuggets.com/2018/01/training-test-sets-cross-validation.html
7. https://www.kdnuggets.com/2017/08/dataiku-predictive-model-holdout-cross-validation.html

- Model Selection

  (a) true and false positive

  (b) true and false negatives

  (c) sensitivity (or recall)

  (d) specificity, precision

  (e) type I and II error

  (f) confusion matrix

  (g) F_1 score

  (h) receiver operating characteristic (ROC)

  (i) area uner curve (AUC)

Additional References:

1. `https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/`
2. `https://www.kdnuggets.com/2018/10/confusion-matrices-quantify-cost-being-wrong.html`
3. `https://www.kdnuggets.com/2020/05/model-evaluation-metrics-machine-learning.html`

- Iris dataset
  (a) iris labels and features
  (b) statistics
  (c) histograms
  (d) scatterplots and counts
  (e) counting
  (f) bar and violin plots
  (g) pair plots

Additional References:

1. `https://www.ritchieng.com/machine-learning-iris-dataset/`
2. `https://scikit-learn.org/stable/auto_examples/datasets/plot_iris_dataset.html`

- Maximum Likelihood estimation
  (a) how is MLE used?
  (b) MLE for Bernouilli distribution
  (c) MLE for Poisson distribution
  (d) MLE for Normal distribution

Additional References:

1. https://towardsdatascience.com/probability-concepts-explained-maximum-likelihood-estimation-c7b4342fdbb1
2. https://www.kdnuggets.com/2019/11/probability-learning-maximum-likelihood.html

- gradient descent
  (a) optimization by iterations
  (b) gradient
  (c) curvature
  (d) stopping criteria
  (e) learning rate
  (f) oscillations

Additional References:

1. https://ml-cheatsheet.readthedocs.io/en/latest/gradient_descent.html
2. https://www.kdnuggets.com/2020/05/5-concepts-gradient-descent-cost-function.html
3. https://www.kdnuggets.com/2017/04/simple-understand-gradient-descent-algorithm.html
4. https://www.kdnuggets.com/2018/11/mastering-learning-rate-speed-up-deep-learning.html

- data scaling
  (a) need for scaling
  (b) min-max scaling
  (c) standard scaling

Additional References:

1. `https://machinelearningmastery.com/scale-machine-learning-data-scratch-python/`
2. `https://www.kdnuggets.com/2020/04/data-transformation-standardization-normalization.html`

- describing the data
  (a) discrete vs. continuous data
  (b) probability distributions
  (c) mean and standard deviation
  (d) Bernouilli, uniform, binomial, Poisson, Normal
  (e) outliers
  (f) bounds

Additional References:

1. https://www.kdnuggets.com/2018/08/basic-statistics-python-probability.html
2. https://www.kdnuggets.com/2018/11/5-basic-statistics-concepts-data-scientists-need-know.html
3. https://www.kdnuggets.com/2017/02/datascience-introduction-correlation.html
4. https://www.kdnuggets.com/2019/07/annotated-heatmaps-correlation-matrix.html
5. https://www.kdnuggets.com/2020/02/probability-distributions-data-science.html

- logistic regression
  (a) linear separability
  (b) logistic vs. linear regression
  (c) odds and logit function
  (d) computing weights
  (e) analysis of categorical data

Additional References:

1. https://towardsdatascience.com/understanding-logistic-regression-step-by-step-704a78be7e0a
2. https://www.kdnuggets.com/2019/10/build-logistic-regression-model-python.html
3. https://www.kdnuggets.com/2018/02/logistic-regression-concise-technical-overview.html
4. https://www.kdnuggets.com/2019/01/logistic-regression-concise-technical-overview.html

- nearest neighbor classification
  (a) distances and neighbors
  (b) nearest neigbor intuition
  (c) need for scaling
  (d) how to choose k
  (e) analyzing categorical data

Additional References:

1. https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_knn_algorithm_finding_nearest_neighbors.htm
2. https://www.kdnuggets.com/2020/04/introduction-k-nearest-neighbour-algorithm-using-examples.html
3. https://www.kdnuggets.com/2017/09/rapidminer-k-nearest-neighbors-laziest-machine-learning-technique.html
4. https://www.kdnuggets.com/2018/03/introduction-k-nearest-neighbors.html

- predictions with linear models
  (a) linear functions vs. linear models
  (b) polynomial functions
  (c) fitting with polynomials
  (d) effect of noise
  (e) bias vs. variance
  (f) overfitting
  (g) trade-offs in linear models

- ordinary linear regression
  (a) linear prediction
  (b) residuals and loss function
  (c) geometric meaning of slope and intercept
  (d) correlation and covariance
  (e) error variance
  (f) computation of parameters

Additional References:

1. https://towardsdatascience.com/a-beginners-guide-to-linear-regression-in-python-with-scikit-learn-83a8f7ae2b4f
2. https://intellipaat.com/blog/what-is-linear-regression/
3. https://www.kdnuggets.com/2016/11/linear-regression-least-squares-matrix-multiplication-concise-technical-overview.html

- Naive Bayesian classification
  (a) conditional probability
  (b) Bayes formula
  (c) prior and posterior probabilities
  (d) feature independence assumption
  (e) naive bayesian classification
  (f) discrete cases
  (g) continuous case

Additional References:

1. https://www.kdnuggets.com/2019/10/bayes-theorem-applied-machine-learning.html
2. https://www.kdnuggets.com/2020/06/naive-bayes-algorithm-everything.html
3. https://dzone.com/articles/naive-bayes-tutorial-naive-bayes-classifier-in-pyt

- decision trees
  (a) root, interior and leaf nodes
  (b) entropy and information gain
  (c) gini impurity
  (d) decision trees for categorical data
  (e) advantages and disadvantages of trees

Additional References:

1. https://www.kdnuggets.com/2018/12/guide-decision-trees-machine-learning-data-science.html
2. https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html
3. https://www.kdnuggets.com/2019/08/understanding-decision-trees-classification-python.html
4. https://www.kdnuggets.com/2020/02/decision-tree-intuition.html
5. https://www.kdnuggets.com/2019/02/decision-trees-introduction.html

- ensemble methods
  (a) ensemble learning
  (b) bagging
  (c) advantages and disadvantages
  (d) hyperparameters (estimators, max features, depth)
  (e) Random Forest classification
  (f) AdaBoost classification

Additional References:

1. https://towardsdatascience.com/simple-guide-for-ensemble-learning-methods-d87cc68705a2
2. https://www.kdnuggets.com/2016/11/data-science-basics-intro-ensemble-learners.html
3. https://www.kdnuggets.com/2017/10/random-forests-explained.html

- support vector machines (SVM)
  (a) linear separability
  (b) margin and support vectors
  (c) soft vs. hard margins
  (d) kernel transformations
  (e) kernel "trick"
  (f) linear, polynomial and Gaussian SVM

Additional References:

1. https://pythonprogramming.net/support-vector-machine-intro-machine-learning-tutorial/
2. https://www.kdnuggets.com/2016/07/support-vector-machines-simple-explanation.html
3. https://www.kdnuggets.com/2017/08/support-vector-machines-learning-svms-examples.html
4. https://www.kdnuggets.com/2016/06/select-support-vector-machine-kernels.html

- clustering
  (a) inertia
  (b) centroid
  (c) knee method
  (d) hierarchical clustering
  (e) dendogram
  (f) density-based clustering

Additional References:

1. https://www.kdnuggets.com/2018/07/clustering-using-k-means-algorithm.html
2. https://www.kdnuggets.com/2018/06/5-clustering-algorithms-data-scientists-need-know.html
3. https://www.kdnuggets.com/2020/04/dbscan-clustering-algorithm-machine-learning.html
4. https://www.kdnuggets.com/2019/09/hierarchical-clustering.html
5. https://www.kdnuggets.com/2020/02/understanding-density-based-clustering.html

- recommended books with reviews:
  1. `https://www.kdnuggets.com/2020/04/10-best-machine-learning-textbooks-data-scientists.html`
  2. `https://www.kdnuggets.com/2018/05/10-more-free-must-read-books-for-machine-learning-and-data-science.html`
- resources for data science seld-study
  1. `https://www.kdnuggets.com/2020/02/data-science-curriculum-self-study.html`

- interview study guide
    1. https://www.kdnuggets.com/2020/01/data-science-interview-study-guide.html

- data science interview questions and answers
  1. https://365datascience.com/wp-content/uploads/2019/02/Interview_FAQ_365datascience.pdf
  2. https://svrtechnologies.com/120-data-science-interview-questions-pdf/
  3. https://intellipaat.com/blog/interview-question/data-science-interview-questions/
  4. https://www.kdnuggets.com/2016/02/21-data-science-interview-questions-answers.html
  5. https://www.kdnuggets.com/2017/02/17-data-science-interview-questions-answers.html
  6. https://www.kdnuggets.com/2017/02/17-data-science-interview-questions-answers-part-2.html
  7. https://www.kdnuggets.com/2017/02/17-data-science-interview-questions-answers-part-3.html

- introduction to sklearn
    1. `https://scikit-learn.org/stable/tutorial/basic/tutorial.html`