|  |  |
| --- | --- |
| סוגי אלגוריתם |  |
| Supervised |  |
|  | Classification |
|  | Regression |
| Unsupervised |  |
|  | Clustering |
| Reinforcement |  |

|  |
| --- |
| תצוגת הנתונים |
| Features Matrix |
| Target Array |

|  |
| --- |
| Estimator |
| Estimator |
| Transformer |
| Predictors |

|  |
| --- |
| תהליכים |
| בחירת מודל |
| בחירת hyper parameters של המודל |
| סידור הנתונים במבנה שיתאים למודל |
| יישום המודל על הנתונים – fit |
| בחירת מדד למדידת ביצועי המודל – כמו ה Mse |

|  |  |  |  |
| --- | --- | --- | --- |
| Feature engineering |  |  |  |
| Feature selection |  |  |  |
|  | VarianceThreshold – הסרת מאפיינים עם שונות נמוכה |  |  |
|  | בחינות סטטיסטיות:   * SelectKBest * SelectPercentile * SelectFpr/SelectFdr | Score functions עבור הבחינות :  רגרסיה:   * f\_regression * mutual\_info\_regression   סיווג:   * Chi\_2 * F\_classif * Mutual\_info\_classif |  |
| Feature extraction | יצירת מאפיין חדש מתוך המאפיינים הקיימים |  |  |
|  |  |  |  |
| Data Preprocessing |  |  |  |
|  | Changing the representation of the data |  |  |
|  |  | LabelEncoder |  |
|  |  | OneHotEncoder |  |
|  |  |  |  |
|  | Imputation of missing data |  |  |
|  |  | SimpleImputer | Mean  Median  Most\_frequent  constant |
|  |  |  |  |
|  | Discretization |  |  |
|  |  | KBinsDiscretizer | Strategy:   * Uniform * Quantile * Kmeans   Encode:   * onehot * Onehot-dense * ordinal |
|  |  |  |  |
|  | Feature scaling |  |  |
|  |  | MInMaxScaler |  |
|  |  | StandardScaler |  |
|  |  |  |  |
| Text Pre Processing  NLTK library |  |  |  |
|  | Removing stopwords  nltk.corpus.stopwords.words('english') |  |  |
|  | Removing special characters  re.sub(<regex>) |  |  |
|  | lower casing  doc.lower.strip()  מסיר רווחים מההתחלה והסוף |  |  |
|  | Text tokenization  בעזרת הפיצול אפשר לבדוק ולהסיר את ה stop words |  |  |
|  | * Contraction expansion * Correcting spellings * Stemming חותך את הסיומות של המילים * Lemmatization מוציא שורש |  |  |
| Text Feature extraction | Bag of words / Bag of N-Gram  CountVectorizer Class | רוחב המטריצה הוא כגודל כמות המילים  Get\_feature\_names() – מציג את הפלט כ data frame  N – gram יכול להיות bi-gram, tri-gram...  ngram\_range – מציין את הטווח הנדרש |  |
|  | TF-IDF Model – TfidVectorizer Class |  |  |
|  | Latent Semantic Analysis )LSA( | חיפוש משמעות חבויה |  |
|  | Latent Dirichlet Allocation (LDA) = pLSA |  |  |
|  | Word2Vec  t-SNE (t-distributed Stochastic Neighbor Embedding) | Tsne - כלי לתצוגה של נתונים במימד גבוה. אפשר להציג את ה word2vec באמצעותו |  |

|  |  |
| --- | --- |
| TestSet |  |
| train\_test\_split |  |
| StratifiedKFold |  |

|  |  |
| --- | --- |
| Pipeline |  |
| make\_pipeline |  |
| ColumnTransformer | יכול לאחד מספר pipelines, וכך לאפשר טרנספורמציה שונה לכל עמודה |
| CustomTransformer | Implements from TransformerMixin or BaseEstimator (added set & get params methods) |

|  |  |
| --- | --- |
| Validation |  |
| cross\_val\_score function |  |
|  | cv – ניתן לשלוט על ה split  StratifiedKFold - Stratified sampling  Score='neg\_mean\_square\_error' לרגרסיה לינארית  Score = 'accuracy' ל classification |
| Evaluating The Test Set |  |
|  | final\_prediction = Final\_model.predict(test)  acuuracy\_score(y\_test,final\_prediction) |
| cross\_val\_predict function |  |
|  | דומה ל cross\_val\_score רק במקום להחזיר את הניקוד, הוא מחזיר לכל אלמנט את הניבוי שהשיג כשהיה ב test set בזמן החלוקה של ה cross |
| Confusion matrix |  |
|  | Y\_pred = cross\_val\_predict() or predict()  Confusion\_matrix(y\_train,y\_pred) |
| Precision/recall/f1 score/accuracy | accuracy\_score() / precision\_score() /recall\_score() |
| precision\_recall\_curve() |  |
| roc\_curve() | TPR vs FPR |

|  |  |  |
| --- | --- | --- |
| Fine Tuning The Model |  |  |
| GridSearchCV |  |  |
|  | estimator  param\_grid  scoring  cv  refit | אפשר לשמור את המודל הנבחר |
| RandomizedSearchCV |  |  |

|  |
| --- |
| The Bias-Variance Tradeoff |
| Underfit the data <-> the model has high bias  Overfit the data <-> the model has high variance |
| במילים אחרות:  למודלים עם high bias , ביצועי המודל על ה validation set דומים לאלו שעל ה training set  למודלים עם שונות גבוהה, ביצועי המודל על ה validation set רחוקים מאד מאלו שעל ה training set |
| Validation curve |
| המטרה למצוא את האיזון בין ה Bias ל variance |

|  |  |
| --- | --- |
| מודלים מסוג linear Regression |  |
| LinearRegression | Batch GD |
| SGDRegressor | Stochastic GD/ / Mini Batch GD |
| Polynomial Features + LinearRegression |  |
| Gaussian Features + LinearRegression | אין מחלקה של Gaussian . יש דוגמא בספר |
| Ridge Regression | Ridge class או SGDRegressor עם l2 |
| Lasso Regression | Lasso class או SGDRegressor עם l1 |
| Elastic Net | ElasticNet class |
| Early Stopping | SGDRegressor עם early stopping = true |
| KNN regression | ממוצע הערכים של K השכנים הקרובים |
| DecisionTree  DecisionTreeRegressor() |  |

|  |  |
| --- | --- |
| מודלים מסוג Logistic Regression |  |
| Logistic Regression class |  |
| Softmax regression / multinomial logistic regression | LogisticRegression class עם multi\_class=multinomial  Solver=lbfgs |

|  |  |
| --- | --- |
| מודלים מסוג Classification |  |
| Logistic Regression class עבור binary classifier |  |
| SGDClassifier | SGDClassifier.decision\_function() מחזיר ניקוד כל מופע, אפשר לבצע ניבוי בהתבסס על הניקוד תוך שימוש ב threshold מסויים  ב"מ OvA |
| Multinomial (multiclass) classifiers  OneVsOneClassifier class | OvA / OvO  בגלל שה SDG תומך ב OvA כברירת מחדל, יש מחלקה ייעודית ל OvO |
| SVM ב"מ OvO |  |
| RandomForestClassifier | אין לו צורך ב OvA או ב OvO כי הוא מנבא ישירות |
| KNN  K-Nearest Neighbors  KNeighborsClassifier | לפי הרוב קובע |
| Naïve Bayes Classification |  |
| Gaussian Naïve Bayes |  |
| Multinomial Naïve Bayes  MultinomialNB class | בשימוש לעיתים קרובות עבור טקסט  אפשר להפעיל את זה אחרי שקוראים ל TfidfVectorizer() |
| Decision Tree  DecisionTreeClassifier() | Info Gain  Gain Ratio  Gini Index  Pre-Pruning/Post pruning  export\_graphviz() |
| SVM  SVC class  LinearSVC class  Polynomial Kernel  Gaussian Kerne | LinearSVC = SVC (kernel linear) = SGDClassifier (loss hinge)  דוגמא, Polynomial Features+ Standard Scaler + LinearSVC  SVC (kernel poly)  SVC (kernel rbf)  SVC (kernel sigmoid) – ברשתות עצבים |
| Perceptron | SGDClassifier (loss perceptron, learning rate,eta0) |
| MLP – Multilayer Perceptron | אפשר להוסיף regulariztion |

|  |  |
| --- | --- |
| מודלים מסוג Clustering |  |
| K-Means /  Mini Batch K-Means | Inertia  Elbow  מדד Silhouette לבחירת מספר אשכולות |
| Hierarchical clustering | Agglomerative :עם linkage measures שונים.  divisive |
|  |  |
|  |  |