ALGO MODEL5 COMMANDS for PRODIGY

!! RULES: do not include method, algorithm, or model in any labels, including common acronyms (kNN, SVM, etc.), included NN although it could be neural network or nearest neighbor, include all 'new' algorithm names

This is kind of a combo of known ML algorithms + anything that comes before method, algo, etc. included 'regression' and 'classification', not model, algo, method or technique

Tried this, but not sure how to load the seeds in a file: prodigy terms.train-vectors

Tried this, but got an error, only uses tokens, not vectors (multiple words): prodigy terms.teach algo_seeds en_vectors_web_lg --seeds "logistic_regression, random_forest, artificial_neural_networks, decision_tree, random_forest"

Old code, gave errors when I tried to use it with ner.teach or ner.match: prodigy sense2vec.teach algo_pattern5 /Users/sashaqanderson/Dropbox/USGS/NER_Work/s2v_old --seeds "logistic_regression, random_forest, artificial_neural_networks, decision_tree"

ERROR

{'text': 'logistic_regression|NOUN', 'word': 'logistic regression', 'sense': 'NOUN', 'meta': {'score': 1.0, 'sense': 'NOUN'}, 'answer': 'accept', '_input_hash': 1642957147, '_task_hash': 798897654}

Not necessary, no seeds: prodigy db-out algo_pattern5 > /Users/sashaqanderson/Dropbox/USGS/NER_Work/algo_model5/algo_patterns5.jsonl

#old version of prodigy can't use patterns with ner.manual. First create dataset using ner.manual. Teach it a bit, then use ner.match to use the patterns… (trained 10 manually first, then tried match)

Prodigy ner.manual algo_data5 blank:en /Users/sashaqanderson/Dropbox/USGS/NER_Work/algo_model4/arxiv_train.jsonl --label ALGO

NOTES FOR ALGO LABELS: (train 201)

Labeled	Did not Label	Rejected
Binary classification	Glasso	To
Lasso	Multi armed bandit problems	Model
NN classifier		Machine learning
Nearest neighbor		Corresponding clustering
NN		Unsupervised learning
NN classification		Algebraic geometry
Soft-margin linear binary support vector machine		VC dimension
Bruhat-Tits tree		Random words and sentences
Stochastic classification		
Hidden Markov (not models)		
Baldi-Chauvin (not algorithm)		
Bayesian		
Naïve bayes		
Mixture		
Simulated annealing		
Markov chain monte carlo		
Nested Chinese restaurant (not process)		
Bayesian nonparametric (no model not method)		
Frequentist (methods)		
Online (algorithm)		

Message-passing (algo)		
Binary regression trees		
Random forest		
Parametric (model)		
Support vector machines		
Classification constrained dimensionality reduction (algo)		
CCDR (algo)		
$ k$ nearest neighbors		
Latent Dirichlet Allocation		
Interaction component (model for communities)		
Dirichlet Process (not priors)		
Kernel Hilbert spaces		
K-nearest neighbors		
KNN, k-NN, kNN		
MCMC		
Bayesian KNN		
BKNN		
Information Preserving Independent Component Analysis		
Local linear embedding (not algorithm)		
Manifold-learning (not algorithms)		
Manifold-based embedding (not algorithms)		
High-dimensional generalized additive (not models)		
LLE		
K-nearest neighbors		
KNN, k-NN, kNN		
Clustering (algo)		
Principal component analysis		
PCA		
Decomposable Gaussian graphical (models)		
Computer vision		
Image processing		
Sparse linear (model)		
SLM		
Online boosting (algo)		
AdaBoost		
Markov Random Fields		
Nadaray-Watson regression (not technique)		
Min-cut clustering		
Information Bottleneck (not method)		
Exponential random graph (not models)		
ERG (models)		
Epsilon-machines		
REMAPF (not algorithm)		
Tree-based regressor		
RPtree partitioning		
High-dimensional causal (modeling)		
Linear non-Gaussian causal (model)		
Boosting (algo)		
High-dimensional Linear (model)		
High-dimensional linear (not model)		
Regularized linear (not model)		
Tree ensembles		

Kernel (methods)		
Kernel ridge regression		
KNIFE		
Latent variable (model)		
Network (models)		
Kernel-smoothing l1-penalized logistic regression		
Cellular automata (model)		
Fuzzy logic		
Binary classification		
Kmeans clustering		
K-means		
neares-prototype classifier		
prototype vector machine		
PVM		
Kriging		
Sparse PCA		
Sparse dictionary learning		
Prim's (not algorithm)		
Lloyd's (not algorithm)		
$ k$ means		
Spectral clustering		
Dirichlet Process Mixtures of Generalized Linear (Models)		
CART		
Laplacian support vector machine		
LapSVM		
Expectation propagation		
Linear regression		
LASSO		
Forward step-wise regression		
Additive noise (not models)		
Slow Feature Analysis		
SFA		
Covariance (model)		
Full-rank unconstrained (model)		
Expectationmaximization		
Optimal aggregation (not algorithm)		
Bayesian Canonical Correlation Analysis		
Cascading Indian buffet (not process)		
Deep/directed/nonlinear gaussian Belief Networks		
Forest structured undirected graphical (models)		
Forest		
Gaussian Graphical (model)		
Independent component analysis		
FastICA		
RobustICA		
Online centroid anomaly detection		
Latent supervised Dirichlet allocation		
Regularized least squared ((RLS) regression and classification)		
Least-squares support vector machine		
Ridge regression		
Greedy RLS		
Distance-based discriminant (algo)		

Multidimensional scaling (algo)		
Divisive Information-Theoretic Feature Clustering (model)		
Linear Bayesian network (models)		
Sparse Linear Identifiable Multivariate (modeling)		
LiNGAM		
Sparse Non-linear Identifiable Multivariate (model)		
Linear ranking support vector machine		
RankSVM		
Restricted Boltzmann Machine		
Deep Belief Networks		
$ k$ fold cross-validation		
Akaike information criterion		
Bayesian Information Criterion		
StARS		
Stability-based (method)		
Density (not modeling)		
Gaussian Process Latent variable		
State-space (model)		
Reduced rank multivariate (not model)		
Rank constrained vector generalized linear (not model)		
nonparametric classification		
nonparametric regression		
semiparametric(model)		
additive (model)		
quantile regression		
kernel based nearest neighbor (approach)		
Kernel based nearest neighbor (not approach)		
Kernel Induced Random Survival Forest		
Random Survival Forest		
Deep Boltzmann machine		
Deep network		
Two-layer sparse group Bolztmann machine		
Normal means (model)		
Iterative Detection estimation		
Parametric kernel-based (not method)		
Nonparametric state-space (not model)		
Nonparametric classification		
Regularized kernel (methods)		
Sparse linear regression		
Convergent optimixation algo)		
Lasso		
Binary classifier		
Supervised binary classification		
Online proximal (not algorithms)		
Multiple kernel learning		
Penalized regression (not method)		
Single Line Search (not (SLS) algorithm)		
Signed Single Line Search (not (SSLS) algorithm)		
Leave-one-out cross validation		
Hold out cross validation		
Leave$upsilon out cross validation		
Heteroscedastic (model)		

Sparse Poisson-like (model) Conditional random fields Natural language processing Computer vision Temporally varying coefficient varying structure ((VCVS) graphical model) TESLA (loss) Temporally smoothed L1 regularized regression VCVS (model) proximal gradient (method) Simultaneous Orthogonal Matching Pursuit ((S-OMP) procedure) Multi-task regression Bayesian Information Criterion Multi-task learning Adapaptive Lasso Greedy forward regression Orthogonal Matching Pursuit Multi-output regression Automatic target recognition ATR (algo) Vector autoregressive (VAR model) VAR (followed by model) Linear acyclic (not models) Threshold based correlation screening (methods) Density (modeling) Autoencoders Non-convolutional network Gaussian process classifier Convergent optimization (algo) Gaussian process (GP models) GP (model) Generalized Additive (model) Additive (model) Sparse regression Discrete infinite logistic normal distribution Mixed membership (model) Hierarchical Dirichlet (process) DILN topic (model) Correlated topic (model) Online inference (algo) Topic (modeling) Sufficient component analysis Epanechnikov kernel Multiple kernel learning Elastic-net MKL (methods) Additive non-Gaussian noise (models) Least-squares independence regression Additive noise (model) Auto-associative (models) Projection pursuit (algo) Least-trimmed squares regression		

Linear non-Gaussuan structural equation (model)		
LiNGAM		
Spatially constrained agglomerative clustering		
Sparse coding 9algo)		
EM (algo)		
Probabilistic PCA		
Gaussian noise (model)		
Univariate regression		
Generalized linear (model)		
Single-index hazard rate (model)		
Penalized regression		
Metropolis-Hastings (algo)		
Infinite Hidden markov (models)		
Multivariate autoregressive (processes)		
Group lasso		
Dirichlet Diffusion Tree		
MCMC		
Bayesian EM search (algo)		
ANOVA kernels		
Nonnegative matrix factorization		
Gaussian graphical (model learning)		
Adaptive gradient-based (method)		
Dirichlet (process)		
Homogenous Poisson (process)		
Random dot product graph (model)		
Latent position (model)		
Laplacian spectral clustering		
Gaussian Scale mixtures		
GSM (model)		
Bessel K (model)		
Real-value signal (model)		
Concave regularization (methods)		
Concave high-dimensional sparse estimation (procedures)		
Forest		
Boosted decision trees		
Adaboost		
Friedman’s gradient boosting		
Boosting (algo)		
Gradient boosting		
Decision forests		
Probabilistic (model)		
Directed graphical (model)		
Occlusion based (model)		
Directed (model)		
Dirichlet Variable Length Markov (model)		
Time Convolutional Restricted Bolztmann Machine		
Factor analysis (model)		
Nonparametric regressors		
k-NN regression		
k nearest neighbors		
k-NN classifier		
k approximate nearest neighbor classifier		

ideal regression graphical Lasso Hedge (algo) L_1 regularized maximum likelihood (method) Expectation Propagation Information-maximization clustering Probabilistic classifier Provably convergent EP (algo) Hierarchical divisive clustering Graph-based (algo) Rulefit (algo) Ensemble of prediction (models) Sparse clustering (algo) Sparse k-means (ago) Trimmed k-means (algo) Robust sparse k-means (algo)		

prodigy db-out algo_data5 >
/Users/sashaqanderson/Dropbox/USGS/NER_Work/algo_model5/algo_data_model5.jsonl

used pretraining from model4
python -m spacy pretrain /Users/sashaqanderson/Dropbox/USGS/NER_Work/algo_model5/arxiv_train.jsonl en_vectors_web_lg /Users/sashaqanderson/Dropbox/USGS/NER_Work/algo_model5/pretrain_algo_model5/ --use-vectors

Prodigy ner.batch-train algo_data5 en_vectors_web_lg --init-tok2vec /Users/sashaqanderson/Dropbox/USGS/NER_Work/algo_model4/pretrain_algo_model4/model999.bin --output algo_model5 --eval-split 0.2 --label ALGO

56.9% (best yet for first round!!) Afraid to make-gold :-\ Made a copy ofalgo_model5 just in case. I screw it up…

Just testing this out to see what it saves. It saves the annotations of the current model. I'm not sure how to get my annotations.
prodigy db-out algo_data5 >
/Users/sashaqanderson/Dropbox/USGS/NER_Work/algo_model5/annotations201.jsonl

Prodigy ner.make-gold algo5 ./algo_model5 /Users/sashaqanderson/Dropbox/USGS/NER_Work/algo_model4/arxiv_train.jsonl --label ALGO

Prodigy ner.batch-train algo5 en_vectors_web_lg --init-tok2vec /Users/sashaqanderson/Dropbox/USGS/NER_Work/algo_model4/pretrain_algo_model5/model999.bin --output algo_model5b --eval-split 0.2 --label ALGO

41.9% why does it go down? Keep making gold…

Prodigy ner.print-stream algo_model5b /Users/sashaqanderson/Dropbox/USGS/NER_Work/algo_model4/arxiv_test.jsonl --label ALGO

Prodigy ner.print-stream algo_model5b /Users/sashaqanderson/Dropbox/USGS/NER_Work/ner_text_train40.jsonl --label ALGO

Prodigy ner.batch-train algo5 en_vectors_web_lg --init-tok2vec /Users/sashaqanderson/Dropbox/USGS/NER_Work/algo_model4/pretrain_algo_model5/model999.bin --output algo_model5 --eval-split 0.2 --label ALGO

Prodigy ner.teach algo_data5 ./algo_model5 /Users/sashaqanderson/Dropbox/USGS/NER_Work/algo_model4/arxiv_train.jsonl --label ALGO --patterns /Users/sashaqanderson/Dropbox/USGS/NER_Work/algo_model5/algo_patterns5.jsonl

---

Started again with Model 5, with manual labeled data but new make_gold.

Model 5b

Prodigy ner.make-gold algo5b ./algo_model5b /Users/sashaqanderson/Dropbox/USGS/NER_Work/algo_model4/arxiv_train.jsonl --label ALGO

Make gold additional accepted terms:
Mahalanobis metric learning (algo)
Compression (method)
Posterior inference (algo)
Adaptive supervised classification
PACBayesian (approach)
Classification (model)
Sparse additive (models)
Data-generating (models)
Prediction (model)
Stochastic (algo)
Lanczos (algo)
Backward (algo)
Hierarchical clustering
Document clustering
Latent variable (models)
Language (model)
Fully connected (model)
Bayesian network
Constraint-based (algo)
Score-based (algo)
Conditional independence (algo)
Bayesian trees
Manifold regularization (approach)
Second order exponential (models)
Multitask learning
Multisensor networks
Greedy RLS
Distance based clustering
Linear latent variable
Thermodynamic soft graph clustering
Singular value penalized (models)
Kernel based nearest neighbor (approach)

Probabilistic topic (model)
Cost-insensitive classification
SLS (algo)
Cross-domain object matching
CDOM (method)
Continuous-variable graphical (models)
Structural equation (models)
Kernel-free (framework)
Convex (model)
Threshold-based correlation screening (methods)
Correlation screening (approach)
Gaussian process classification
Variational inference (algo)
Distribution free SDR (method)
Generative (model)
Partial information (model)
gLasso
Pitman Yor Diffusion Tree
Gaussian Scale mixtures
VMM
Hierarchical Bayesian approach
Reproducing kernel Hilbert space
Asymptotic pseudo-trajectory (approach)
Post processing (methods)
Multi-view predictive partitioning
(model of) stochastic one-armed bandit
Two-block least squares (TBPLS regression model)
MVPP (slgo)
TB-PLS model
Ising (models)
Chow-Lui (algo)
High dimensional undirected graphical (models)
Gaussian Copula graphical (models)
Higher order PCA
Sparse HOPCA
Alternating augmented Lagrangian (method)
Fused Lasso
Muti-dimensional classification
Variational message-passgin (algo)
Hierarchical prior (models)
Linear non-Gaussian acyclic (model)
Partial least squares
Sparse PLS (method)
Regularized PLS
Non-negative PLS
Generalized PLS
Discrimative probabilistic (model)
Semiparametric (model)
Logistic regression
Sequential monte carlo (sampler)
Conditional independence test based (algo)
Convex subspace recover (algo)
Low-dimensional latent mixture (model)

Probabilistic multinomial probit classification
Multiclass GP classficication
multinomial probit GP classification
nested EP (approach)

Trained 1,430

Prodigy ner.batch-train algo5 en_vectors_web_lg --init-tok2vec /Users/sashaqanderson/Dropbox/USGS/NER_Work/algo_model4/pretrain_algo_model5/model999.bin --output algo_model5b --eval-split 0.2 --label ALGO

41.9%

Prodigy ner.batch-train algo5 en_vectors_web_lg --init-tok2vec /Users/sashaqanderson/Dropbox/USGS/NER_Work/algo_model4/pretrain_algo_model5/model999.bin --output algo_model5b --eval-split 0.2 --n-iter 10 --label ALGO

5 hours for nothing ☹