ALGO MODEL4 COMMANDS for PRODIGY

!! RULES: do not include method, algorithm, or model in any labels, including common acronyms (kNN, SVM, etc.), do not include NN – 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

prodigy sense2vec.teach algo_method4 /Users/sashaqanderson/Dropbox/USGS/NER_Work/s2v_old --seeds "logistic_regression, random_forest, artificial_neural_networks, decision_tree, genetic_algorithm "

NOTES FOR ALGO SEEDS:

Labeled	Did not Label	Rejected
SVM Dynamic programming Gradient descent ANNs Fuzzy logic (finite) State machines RNNs Backprop Deep learning Reinforcement learning MCMC Natural language processing Image recognition FPGAs	Cellular automata (include next time)	Evolutionary algorithms Bayesian (methods) ANOVA Simulations Statistical models Clustering NNs Pattern matching Fourier transforms Fourier analysis

prodigy terms.to-patterns algo_method4 --label ALGO

{"label":"ALGO","pattern":[{"lower":"logistic_regression|NOUN"}]}

{"label":"ALGO","pattern":[{"lower":"random_forest|NOUN"}]}

{"label":"ALGO","pattern":[{"lower":"artificial_neural_networks|NOUN"}]}

{"label":"ALGO","pattern":[{"lower":"decision_tree|NOUN"}]}

{"label":"ALGO","pattern":[{"lower":"genetic_algorithm|NOUN"}]}

{"label":"ALGO","pattern":[{"lower":"genetic_algorithms|NOUN"}]}

{"label":"ALGO","pattern":[{"lower":"neural_networks|NOUN"}]}

{"label":"ALGO","pattern":[{"lower":"neural_network|NOUN"}]}

{"label":"ALGO","pattern":[{"lower":"linear_regression|NOUN"}]}

{"label":"ALGO","pattern":[{"lower":"neural_nets|NOUN"}]}

{"label":"ALGO","pattern":[{"lower":"SVMs|NOUN"}]}

{"label":"ALGO","pattern":[{"lower":"decision_trees|NOUN"}]}

{"label":"ALGO","pattern":[{"lower":"random_forests|NOUN"}]}

{"label":"ALGO","pattern":[{"lower":"SVM|ORG"}]}

{"label":"ALGO","pattern":[{"lower":"dynamic_programming|NOUN"}]}

{"label":"ALGO","pattern":[{"lower":"gradient_descent|NOUN"}]}

{"label":"ALGO","pattern":[{"lower":"ANNs|NOUN"}]}

{"label":"ALGO","pattern":[{"lower":"fuzzy_logic|NOUN"}]}

{"label":"ALGO","pattern":[{"lower":"state_machines|NOUN"}]}

{"label":"ALGO","pattern":[{"lower":"state_machine|NOUN"}]}

{"label":"ALGO","pattern":[{"lower":"RNNs|NOUN"}]}
{"label":"ALGO","pattern":[{"lower":"finite_state_machines|NOUN"}]}
{"label":"ALGO","pattern":[{"lower":"neural_net|NOUN"}]}
{"label":"ALGO","pattern":[{"lower":"backprop|NOUN"}]}
{"label":"ALGO","pattern":[{"lower":"deep_learning|NOUN"}]}
{"label":"ALGO","pattern":[{"lower":"reinforcement_learning|NOUN"}]}
{"label":"ALGO","pattern":[{"lower":"backpropagation|NOUN"}]}
{"label":"ALGO","pattern":[{"lower":"natural_language_processing|NOUN"}]}
{"label":"ALGO","pattern":[{"lower":"turing_machines|NOUN"}]}
{"label":"ALGO","pattern":[{"lower":"LSTM|ORG"}]}
{"label":"ALGO","pattern":[{"lower":"MCMC|ORG"}]}
{"label":"ALGO","pattern":[{"lower":"RNN|NOUN"}]}
{"label":"ALGO","pattern":[{"lower":"Turing_machine|NOUN"}]}
{"label":"ALGO","pattern":[{"lower":"image_recognition|NOUN"}]}
{"label":"ALGO","pattern":[{"lower":"FPGAs|NOUN"}]}

prodigy db-out algo_method4 > /Users/sashaqanderson/Dropbox/USGS/NER_Work/algo_model4/algo_patterns4.jsonl

Prodigy ner.manual algo_data4 blank:en /Users/sashaqanderson/Dropbox/USGS/NER_Work/algo_model4/arxiv_train.jsonl --label ALGO

NOTES FOR ALGO LABELS: (train 200)

Labeled	Did not Label	Rejected
Binary classification	Random tree	Framework
Lasso	Tree	Dendograms
Nearest neighbor	CCDR (algorithm)	Mean squared error
Hidden Markov (not models)	Kernel Hilbert spaces	Rival forceasts theorem
Baldi-Chauvin (not algorithm)	(include next time)	Modeling techniques
Naïve bayes	SLM	Vector
Mixture (not models)	Max likelihood (if it is not	Clustering
Compression (not method)	before algorithm)	JSD
Logistic sequence prediction (not model)	Markov equivalence	Data-generating
Logisitic classification (not model)	Reproducing Kernel	E.Coli
Simulated annealing	Hilbert spaces (include	m
Markov chain monte carlo	next time)	Optimal
Nested Chinese restaurant (not process)	Multi-task learning	A, recent, the (method)
Bayesian nonparametric (no model not method)	Online HDP	Statistical, causal, (models)
Binary regression trees	Online LDA	*used causal inference*
Random forests	Elastic-net MKL	*(algorithm)*
PACBayesian (not approach)	Lp-MKL	Several
Classification constrained dimensionality reduction	Probabilistic model	Scorebased
$ k$ nearest neighbors		Random words
Latent Dirichlet Allocation		Learning
Interaction component model for communities		Independence
Dirichlet Process (not priors)		Stochastic
K-nearest neighbors		Deterministic
KNN, k-NN, kNN		Gaussian
MCMC		Non-Gaussian

Bayesian KNN BKNN Information Preserving Independent Component Analysis Manifold-learning (not algorithms) Manifold-based embedding (not algorithms) High-dimensional generalized additive (not models) Sparse additive (not models) Local linear embedding (not algorithm) LLE Laplacian eigenmap Local tangent space alignment Hessian eigenmaps Diffusion maps Principal component analysis PCA Sparse linear (not model) Online boosting (not algorithm) Hierarchical Dirichlet (process) Infinite hidden markov (models) Metropolis-hastings (algorithm) AdaBoost Markov Random Fields Nadaray-Watson regression (not technique) Min-cut clustering Information Bottleneck (not method) Autoencoder neural network Genetic (not algorithm) Maximum likelihood (not algorithm) Decision forest Exponential random graph (not models) Epsilon-machines REMAPF (not algorithm) Kernel Partial Least Squares Lanczos (not algorithm) Tree-based regressor RPtree partitioning Boosting (not algorithm) High-dimensional linear (not model) Regularized linear (not model) Tree ensembles Feature Importance Ranking Measure Kernel ridge regression Support vector machine Bayesian (not model) Mote carlo Network (not models) Logistic regression Kernel-smoothing Cellular automata network k-means clustering k-means		Approximation Heuristic Popular Technical Corresponding stability local linear graphical probabilistic inference author names

neares-prototype classifier prototype vector machine kriging (not models) sparse PCA sparse dictionary learning Prim's (not algorithm) Lloyd's (not algorithm) $ k$ means Spectral clustering Diriclet Process Mixtures of Generalized Linear Models Dirichlet process mixture regression (not models) CART Bayesian trees Nonparametric regression Gaussian processes Laplacian support vector machine LapSVM Manifold regularization (not approach) Linear regression LASSO Forward step-wise regression Supervised dictionary learning Additive noise (not models) Slow Feature Analysis Expectation-maximization (not algorithms) Full-rank unconstrained (not model) Optimal aggregation (not algorithm) Baesian Canonical Correlation Analysis Cascading Indian buffet (not process) Kruskal's (not algorithm) Independent Component Analysis ICA, FastICA, RobustICA, Supervised latent Dirichlet allocation Regularized (not regression) Regularized least squared (not (RLS) regression and classification) Least-squares support vector machine Ridge regression Greedy RLS Distance-based discriminant (not algorithm) Multidimensional scaling (not algorithm) Divisive Information Theoretic Feature Clustering (not model) Linear Bayesian Network (not models) SLIM (selected this due to wording) Sparse Linear Identifiable Multivariate (not modeling) SNIM (selected this due to wording good for training for new algos) Sparse Non-Linear Identifiable Multivariate CSLIM Correlated SLIM		

Linear ranking support vector machine RankSVM Moment generating function (technique) Restricted Boltzmann Machine Deep Belief Networks High-dimensional graphical (not models) $ k$ fold cross-validation Akaike information criterion Bayesian Information Criterion StARS Density (not modeling) Gaussian Process Latent variable Thermodynamic soft graph clustering State-space (not model) Reduced rank multivariate (not model) Rank constrained vector generalized linear (not model) Singular value penalized (not models) Special kernel based (not methods) Semiparametric (not model) Kernel based nearest neighbor (not approach) Kernel Induced Random Survival Forest Random Survival Forest Boltzmann Machine Normal means (not model) Iterative Detection Estimation Parametric kernel-based (not method) Nonparametric state-space (not model) Nonparametric classification Regularized kernel (not methods) Sparse linear regression Convergent optimization (not algorithm) Regularized least squares Supervised binary classification Bagging Online proximal (not algorithms) Multiple kernel learning Penalized regression (not method) Single Line Search (not (SLS) algorithm) Signed Single Line Search (not (SSLS) algorithm) Maximum likelihood joint tracking A bunch of different cross validation (not procedures) Standard linear (not model) Poisson-like (not model) Sparse Poisson-like (not model) Heteroscedastic, Homoscedastic (not model) Conditional random fields Cross-object domain matching CDOM (when followed by method) Bayesian modelling (not framework) Expectation propagation Continuous variable graphical (not models)		

Covariance decoupling (not techniques) Varying Structure (not VCVS graphical model) Temporally smoothed L1 regularized regression VCVS (before model) Proximal gradient (not method) Simultaneous Orthogonal Matching Pursuit (not S-OMP procedure) Ultra-high dimensional multi-task regression Adaptive lasso Greedy forward regression Orthogonal Matching Pursuit Multi-output regression ATR (followed by algorithm) Vector autoregressive (not VAR model) VAR (followed by model) Linear acyclic (not models) Structural equation (not model) Bayesian networks Iterative search (not algorithms) Kernel-free (not framework) Autoencoders Non-convolutional network Gaussian process classification Gaussian process (not model) GP (followed by model) Generalized additive (not models) Topic (not modeling) Variational inference (not algorithm) DILN topic (not model) Online inference (not algorithm) Distribution-free SR (not method) Sufficient component analysis Leas-squares independence regression Causal inference (algorithm) Auto-associative (models) Projection pursuit (algorithm) Principle component analysis Least-trimmed square regression Linear non-Gaussian structural equation (model) LiNGAM Sparse coding (algorithm) Gaussian noise (model) EM (algorithm) Univariate regression (model) Variable selection (method) Generalized linear (model) Independent screening (method) Single-index hazard rate (model) Penalized regression Multi-variate auto-regressive (process) Group LASSO (gLASSO) Greedy Bayesian EM Search (algo) Dirichlet Diffusion Tree		

Hierarchical clustering Pitman Yor Diffusion Tree Gaussian graphical model learning Adaptive gradient based (method) Dirchlet Homogenous Poisson Stochastic blockmodel Random dot product graph Latent position (model) Laplacian spectral clustering Sparse Bayesian learning Gaussian scale mixtures Real-value signal GSM (model) Bessel-K (model) Concave regularization (methods) Concave high dimensional sparse estimation (procedures) Boosted decision trees Friedman's gradient boosting Decision forests Gradient boosting Directed graphical (model) Directed (model) Occlusion based (model) Dirichlet Variable Length markov (model) Time Convolution restricted Boltzmann Machine Factor analysis (model) Bayesian (model) Didn't type the last 20		

prodigy db-out algo_data4 > /Users/sashaqanderson/Dropbox/USGS/NER_Work/algo_model4/algo_data_model4.jsonl

python -m spacy pretrain /Users/sashaqanderson/Dropbox/USGS/NER_Work/algo_model4/arxiv_train.jsonl en_vectors_web_lg /Users/sashaqanderson/Dropbox/USGS/NER_Work/algo_model4/pretrain_algo_model4/ --use-vectors

Prodigy ner.batch-train algo4 en_vectors_web_lg --init-tok2vec /Users/sashaqanderson/Dropbox/USGS/NER_Work/algo_model4/pretrain_algo_model4/model999.bin --output algo_model4 --eval-split 0.2 --label ALGO

46.8%

Prodigy ner.make-gold algo4 ./algo_model4 /Users/sashaqanderson/Dropbox/USGS/NER_Work/algo_model4/arxiv_train.jsonl --label ALGO

Not bad.. Keep making gold :-/…
try making gold on USGS data??? Think this will mess with it too much and I want to be sure to save it as something else

Prodigy ner.batch-train algo4 en_vectors_web_lg --init-tok2vec /Users/sashaqanderson/Dropbox/USGS/NER_Work/algo_model4/pretrain_algo_model4/model999.bin --output algo_model4 --eval-split 0.2 --label ALGO

65.2%

Hmm made it worse by training more on make gold damn!! Can't go back now 33% ☹

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

Prodigy ner.print-stream algo_model4 /Users/sashaqanderson/Dropbox/USGS/NER_Work/ner_text_train40.jsonl --label ALGO