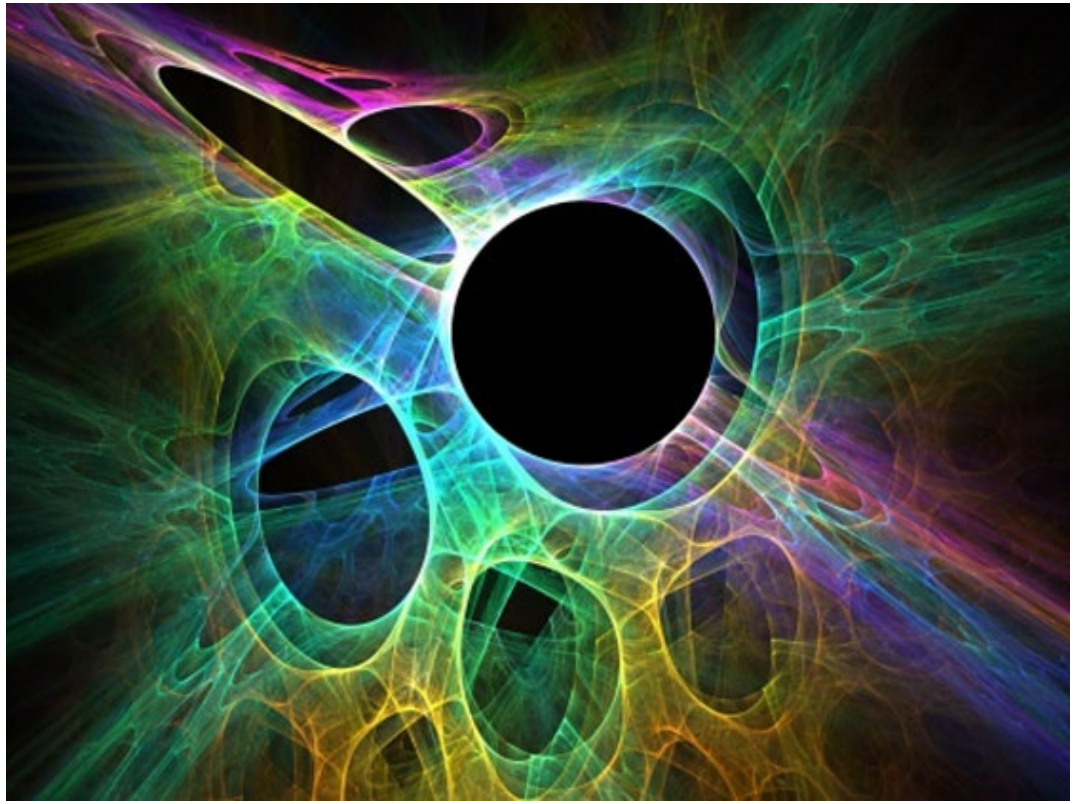
Evaluating Applicability Domain Methods



Scott Kolmar

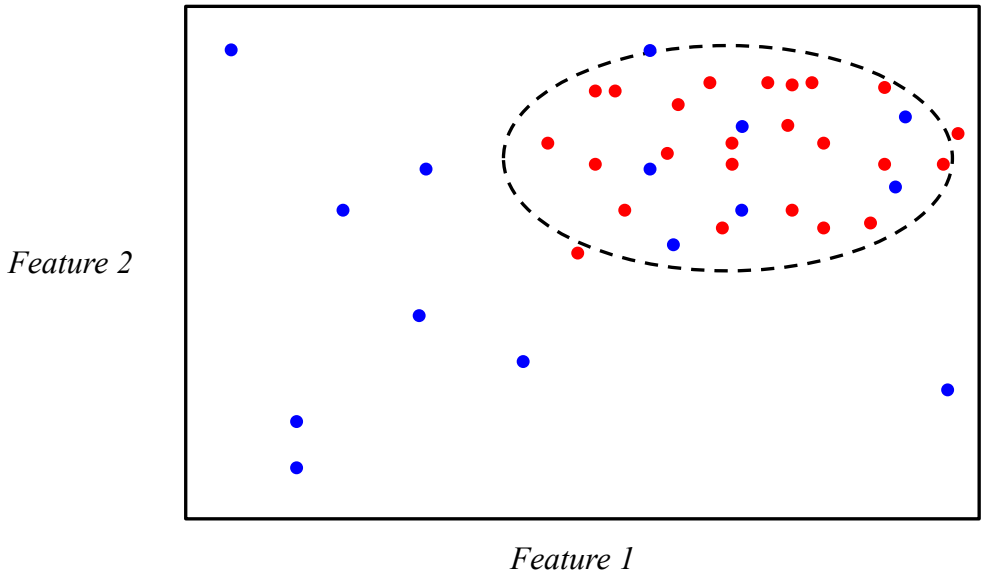
February 11th, 2022

ORD-CCTE-CCED-CCCB

This work does not reflect EPA policy.

Applicability Domain

- *Applicability Domain:*
  "…the response and chemical structure space in which the model makes predictions with a given reliability."

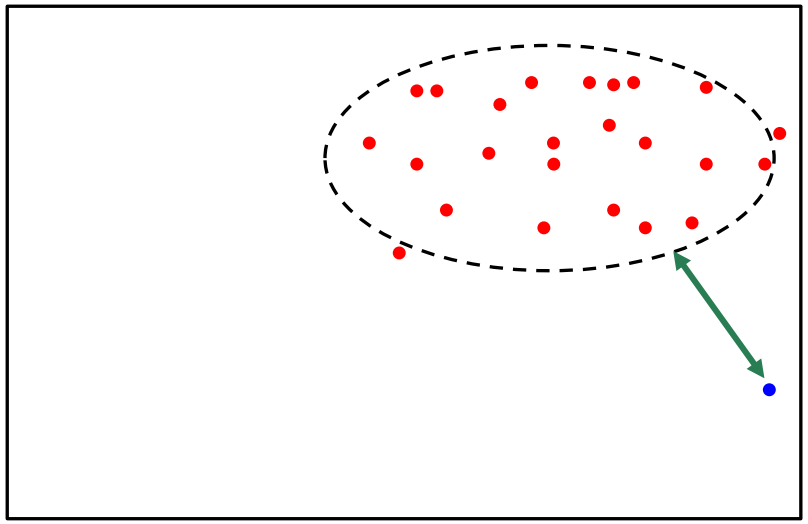


- Red: Training set
- Blue: Test set
- **Black dash:** Training set feature space

Project Goal:

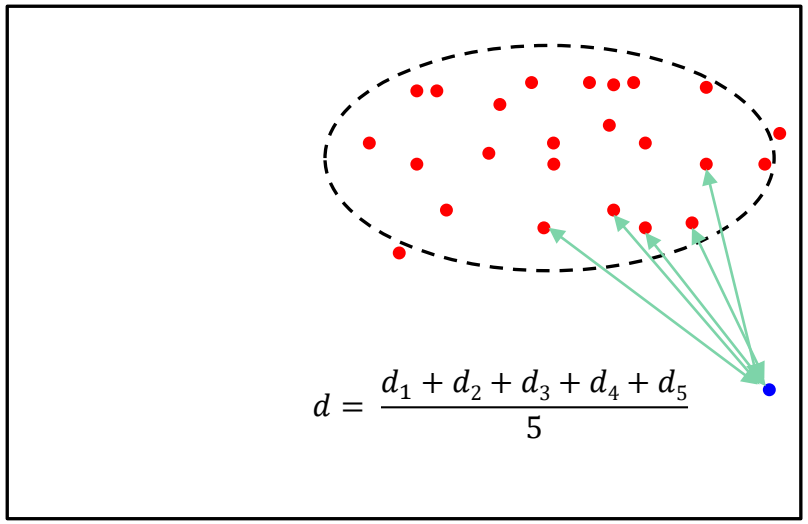
- Quantitatively evaluate Applicability Domain (AD) methods by measuring the correlation between AD metrics and prediction error

Applicability Domain Methods



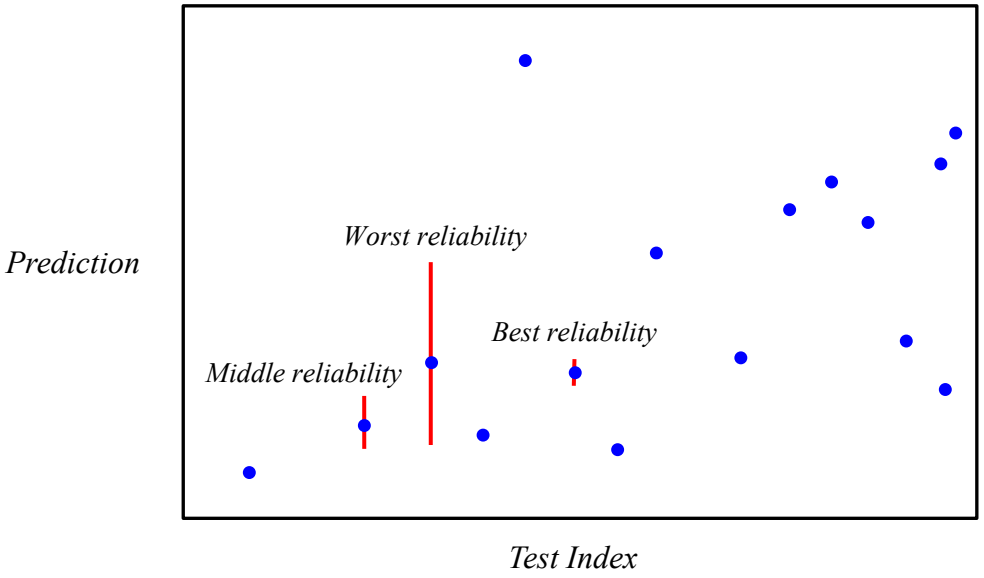
- *Boundary methods*
  - *Approach*
    - Determine distance between test set point and training set boundary
  - *Examples*
    - One Class SVM
    - Robust Covariance
    - Isolation Forest
    - Local Outlier Factor
  - *Anchor Distance*
    - The boundary of the training set space is set to 0

Applicability Domain Methods



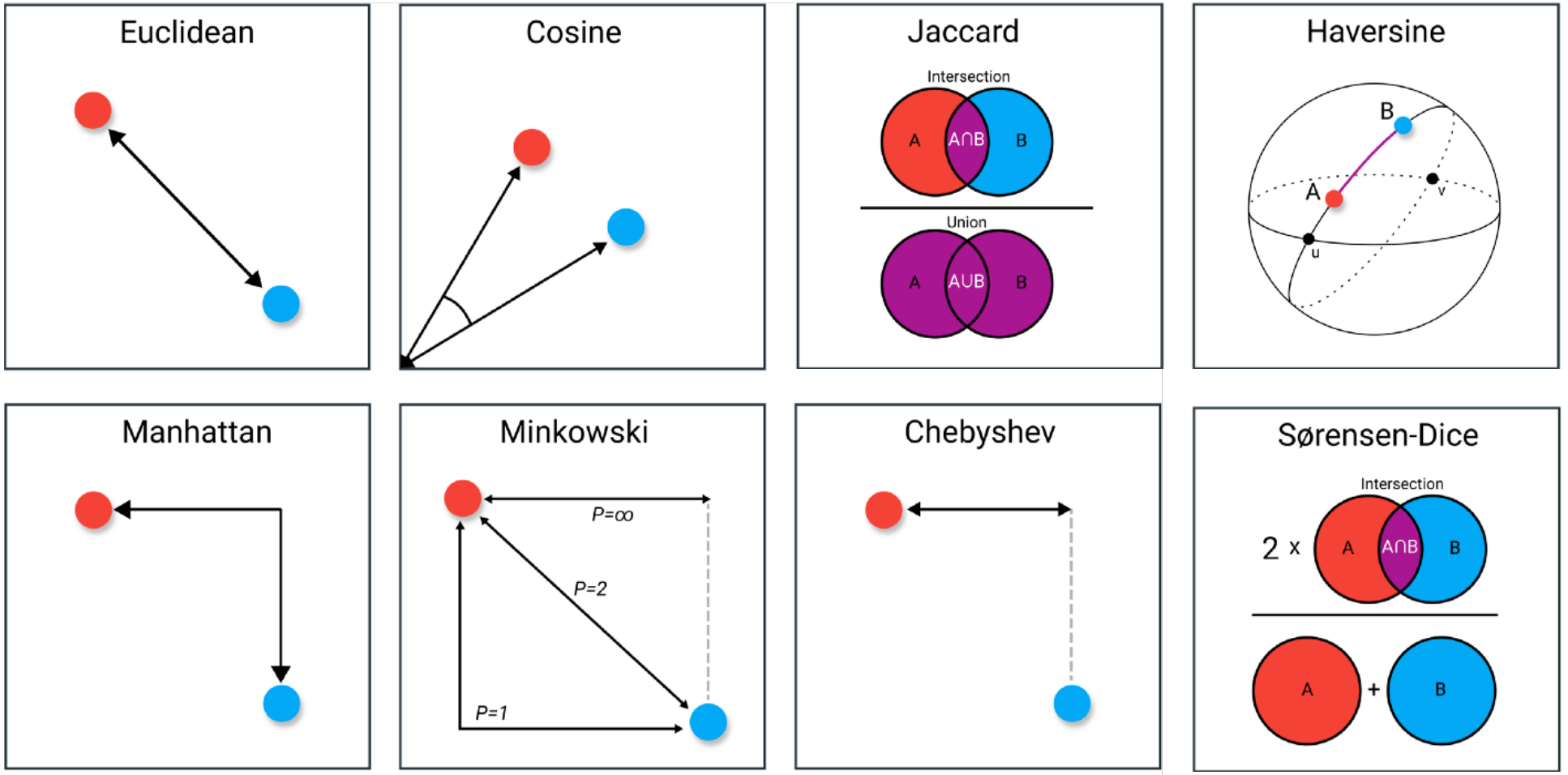
- *kNN method ("distance")*
  - *Approach*
    - Determine average distance between a test point and its *k* nearest training set neighbors
  - *Threshold Distance*
    - The average of:
      the average distance between each training set point and its *k* nearest neighbors

Applicability Domain Methods



- *Confidence Interval methods*
  - *Approach*
    - Use an algorithms internal calculation of prediction uncertainty
  - *Examples*
    - *Random Forest* has a prediction for each individual tree, from which prediction uncertainty is calculated
    - *Gaussian Process* uses a kernel to calculate similarity between compounds which gives prediction uncertainty
    - *Consensus methods* have predictions from multiple models, from which prediction uncertainty is calculated
  - *Threshold measure*
    - Average uncertainty for training set predictions

Distance Measures



Random Forest

	Scaled				Filtered				Filtered and Scaled			
	Pearson R	p-val	R^2	Slope	Pearson R	p-val	R^2	Slope	Pearson R	p-val	R^2	Slope
cosine_distance	0.02	0.82371	0.00	8.478e+02	0.00	0.98900	0.00	1.555e+03	0.10	0.15472	0.01	4.967e+03
euclidean	0.17	0.01827	0.03	8.896e+01	-0.04	0.53397	0.00	-8.312e-04	0.21	0.00245	0.05	1.483e+02
cityblock	0.29	0.00002	0.09	1.199e+01	-0.04	0.54852	0.00	-4.424e-04	0.24	0.00078	0.06	1.009e+01
minkowski	0.17	0.01827	0.03	8.896e+01	-0.04	0.53397	0.00	-8.312e-04	0.21	0.00245	0.05	1.483e+02
chebyshev	0.03	0.68103	0.00	3.587e+01	-0.04	0.56905	0.00	-9.886e-04	0.17	0.01814	0.03	3.380e+02
sorensen	-0.03	0.69711	0.00	-3.110e-02	-0.02	0.79939	0.00	-4.682e+03	-0.00	0.96485	0.00	-6.168e-03
gower	0.29	0.00002	0.09	1.731e+04	-0.04	0.54852	0.00	-4.384e-01	0.24	0.00078	0.06	1.003e+04
sorgel	-0.01	0.91230	0.00	-6.043e-01	-0.02	0.77561	0.00	-3.038e+03	0.04	0.59895	0.00	6.878e-01
kulczynski	0.04	0.56583	0.00	1.240e+01	-0.02	0.80771	0.00	-1.814e+03	0.03	0.69228	0.00	2.688e+01
canberra	0.12	0.08631	0.01	1.365e+01	0.10	0.15366	0.01	2.740e+01	0.17	0.01803	0.03	1.542e+01
lorentzian	0.31	0.00001	0.10	2.835e+01	0.15	0.03946	0.02	1.387e+01	0.23	0.00120	0.05	2.024e+01
czekanowski	-0.03	0.69711	0.00	-3.110e-02	-0.02	0.79939	0.00	-4.682e+03	-0.00	0.96485	0.00	-6.168e-03
ruzicka	-0.01	0.91230	0.00	-6.043e-01	-0.02	0.77561	0.00	-3.038e+03	0.04	0.59895	0.00	6.878e-01
tanimoto	-0.01	0.91230	0.00	-6.043e-01	-0.02	0.77561	0.00	-3.038e+03	0.04	0.59895	0.00	6.878e-01
one_class_svm	-0.16	0.02250	0.03	-4.059e+01	-0.06	0.43006	0.00	-4.145e+00	-0.15	0.03648	0.02	-3.686e+01
isolation_forest	-0.25	0.00031	0.06	-3.736e+04	-0.20	0.00430	0.04	-4.786e+04	-0.13	0.06950	0.02	-1.927e+04
local_outlier_factor	-0.14	0.04698	0.02	-2.383e+03	-0.12	0.08740	0.01	-1.242e+01	-0.21	0.00313	0.04	-6.401e+03

Support Vector Regressor

	Scaled				Filtered				Filtered and Scaled			
	Pearson R	p-val	R^2	Slope	Pearson R	p-val	R^2	Slope	Pearson R	p-val	R^2	Slope
cosine_distance	-0.30	0.00001	0.09	-1.995e+05	0.17	0.01404	0.03	2.509e+06	-0.42	0.00000	0.18	-2.882e+05
euclidean	0.23	0.00100	0.05	1.514e+03	0.11	0.10849	0.01	1.951e-02	0.25	0.00036	0.06	2.428e+03
cityblock	0.39	0.00000	0.15	1.945e+02	0.10	0.14578	0.01	9.763e-03	0.25	0.00034	0.06	1.499e+02
minkowski	0.23	0.00100	0.05	1.514e+03	0.11	0.10849	0.01	1.951e-02	0.25	0.00036	0.06	2.428e+03
chebyshev	0.07	0.33251	0.00	1.038e+03	0.11	0.11510	0.01	2.489e-02	0.12	0.08522	0.01	3.452e+03
sorensen	-0.02	0.79205	0.00	-2.589e-01	0.11	0.12811	0.01	2.550e+05	0.01	0.83940	0.00	3.964e-01
gower	0.39	0.00000	0.15	2.809e+05	0.10	0.14578	0.01	9.675e+00	0.25	0.00034	0.06	1.492e+05
sorgel	-0.20	0.00484	0.04	-1.881e+02	0.10	0.17821	0.01	1.306e+05	0.03	0.71707	0.00	6.623e+00
kulczynski	0.01	0.85723	0.00	4.778e+01	0.12	0.10433	0.01	1.101e+05	-0.34	0.00000	0.12	-4.559e+03
canberra	-0.26	0.00027	0.07	-3.518e+02	0.16	0.02065	0.03	4.037e+02	-0.42	0.00000	0.18	-5.400e+02
lorentzian	0.40	0.00000	0.16	4.511e+02	0.22	0.00156	0.05	1.929e+02	0.21	0.00278	0.04	2.616e+02
czekanowski	-0.02	0.79205	0.00	-2.589e-01	0.11	0.12811	0.01	2.550e+05	0.01	0.83940	0.00	3.964e-01
ruzicka	-0.20	0.00484	0.04	-1.881e+02	0.10	0.17821	0.01	1.306e+05	0.03	0.71707	0.00	6.623e+00
tanimoto	-0.20	0.00484	0.04	-1.881e+02	0.10	0.17821	0.01	1.306e+05	0.03	0.71707	0.00	6.623e+00
one_class_svm	-0.42	0.00000	0.18	-1.306e+03	-0.22	0.00167	0.05	-1.489e+02	-0.43	0.00000	0.18	-1.497e+03
isolation_forest	-0.57	0.00000	0.33	-1.045e+06	-0.51	0.00000	0.26	-1.100e+06	-0.55	0.00000	0.30	-1.147e+06
local_outlier_factor	-0.16	0.02661	0.02	-3.265e+04	-0.03	0.68246	0.00	-2.720e+01	-0.20	0.00514	0.04	-8.480e+04

k-Nearest Neighbors

	Scaled				Filtered				Filtered and Scaled			
	Pearson R	p-val	R^2	Slope	Pearson R	p-val	R^2	Slope	Pearson R	p-val	R^2	Slope
cosine_distance	0.01	0.93912	0.00	7.996e+02	0.11	0.11513	0.01	6.616e+05	0.05	0.47986	0.00	3.581e+03
euclidean	0.29	0.00003	0.09	4.304e+02	0.48	0.00000	0.23	3.347e-02	0.24	0.00064	0.06	2.414e+02
cityblock	0.42	0.00000	0.18	4.747e+01	0.34	0.00000	0.12	1.328e-02	0.27	0.00009	0.08	1.700e+01
minkowski	0.29	0.00003	0.09	4.304e+02	0.48	0.00000	0.23	3.347e-02	0.24	0.00064	0.06	2.414e+02
chebyshev	0.13	0.07746	0.02	4.223e+02	0.54	0.00000	0.29	4.883e-02	0.18	0.01087	0.03	5.274e+02
sorensen	-0.05	0.51292	0.00	-1.438e-01	0.06	0.41154	0.00	5.644e+04	0.00	0.99976	0.00	6.122e-05
gower	0.42	0.00000	0.18	6.855e+04	0.34	0.00000	0.12	1.316e+01	0.27	0.00009	0.08	1.692e+04
sorgel	-0.05	0.49942	0.00	-1.019e+01	0.00	0.95775	0.00	2.108e+03	-0.17	0.01630	0.03	-4.521e+00
kulczynski	0.04	0.57719	0.00	3.313e+01	0.14	0.05587	0.02	5.296e+04	0.07	0.35253	0.00	9.141e+01
canberra	0.06	0.42291	0.00	1.760e+01	0.10	0.14168	0.01	1.053e+02	0.08	0.26099	0.01	1.068e+01
lorentzian	0.43	0.00000	0.18	1.075e+02	0.23	0.00135	0.05	7.999e+01	0.28	0.00007	0.08	3.582e+01
czekanowski	-0.05	0.51292	0.00	-1.438e-01	0.06	0.41154	0.00	5.644e+04	0.00	0.99976	0.00	6.122e-05
ruzicka	-0.05	0.49942	0.00	-1.019e+01	0.00	0.95775	0.00	2.108e+03	-0.17	0.01630	0.03	-4.521e+00
tanimoto	-0.05	0.49942	0.00	-1.019e+01	0.00	0.95775	0.00	2.108e+03	-0.17	0.01630	0.03	-4.521e+00
one_class_svm	-0.31	0.00001	0.10	-2.179e+02	-0.60	0.00000	0.36	-1.663e+02	-0.20	0.00432	0.04	-7.254e+01
isolation_forest	-0.39	0.00000	0.15	-1.571e+05	-0.22	0.00221	0.05	-1.913e+05	-0.23	0.00121	0.05	-4.934e+04
local_outlier_factor	-0.24	0.00048	0.06	-1.141e+04	-0.20	0.00477	0.04	-7.604e+01	-0.18	0.01033	0.03	-8.074e+03

Gaussian Process

	Scaled				Filtered				Filtered and Scaled			
	Pearson R	p-val	R^2	Slope	Pearson R	p-val	R^2	Slope	Pearson R	p-val	R^2	Slope
cosine_distance	-0.22	0.00190	0.05	-1.679e+06	-0.18	0.00927	0.03	-2.960e+07	-0.29	0.00003	0.09	-2.337e+06
euclidean	-0.24	0.00054	0.06	-1.862e+04	-0.11	0.10669	0.01	-2.186e-01	-0.40	0.00000	0.16	-4.517e+04
cityblock	-0.41	0.00000	0.16	-2.379e+03	-0.14	0.04937	0.02	-1.467e-01	-0.41	0.00000	0.17	-2.880e+03
minkowski	-0.24	0.00054	0.06	-1.862e+04	-0.11	0.10669	0.01	-2.186e-01	-0.40	0.00000	0.16	-4.517e+04
chebyshev	-0.05	0.47670	0.00	-8.928e+03	-0.08	0.23234	0.01	-2.106e-01	-0.24	0.00075	0.06	-7.770e+04
sorensen	-0.06	0.40377	0.00	-9.589e+00	-0.17	0.01637	0.03	-4.465e+06	0.09	0.21500	0.01	2.813e+01
gower	-0.41	0.00000	0.16	-3.436e+06	-0.14	0.04937	0.02	-1.454e+02	-0.41	0.00000	0.17	-2.865e+06
sorgel	0.24	0.00056	0.06	2.683e+03	-0.16	0.02124	0.03	-2.480e+06	0.06	0.37399	0.00	1.887e+02
kulczynski	-0.12	0.09666	0.01	-5.146e+03	-0.17	0.01788	0.03	-1.782e+06	-0.44	0.00000	0.19	-6.832e+04
canberra	-0.23	0.00128	0.05	-3.653e+03	-0.29	0.00003	0.08	-7.936e+03	-0.15	0.02842	0.02	-2.322e+03
lorentzian	-0.43	0.00000	0.19	-5.689e+03	-0.24	0.00049	0.06	-2.364e+03	-0.40	0.00000	0.16	-5.714e+03
czekanowski	-0.06	0.40377	0.00	-9.589e+00	-0.17	0.01637	0.03	-4.465e+06	0.09	0.21500	0.01	2.813e+01
ruzicka	0.24	0.00056	0.06	2.683e+03	-0.16	0.02124	0.03	-2.480e+06	0.06	0.37399	0.00	1.887e+02
tanimoto	0.24	0.00056	0.06	2.683e+03	-0.16	0.02124	0.03	-2.480e+06	0.06	0.37399	0.00	1.887e+02
one_class_svm	0.15	0.03818	0.02	5.310e+03	0.12	0.08094	0.02	9.295e+02	0.21	0.00277	0.04	8.512e+03
isolation_forest	0.21	0.00337	0.04	4.392e+06	0.25	0.00046	0.06	5.936e+06	0.20	0.00515	0.04	4.795e+06
local_outlier_factor	0.06	0.37400	0.00	1.540e+05	0.14	0.04749	0.02	1.462e+03	0.25	0.00031	0.06	1.261e+06

Supplemental

Distance Measures

- *Euclidean*
  - NOT scale invariant
  - Behaves poorly in higher dimensions
- *Cosine distance*
  - Does not take magnitude of vectors into account
- *Manhattan (cityblock)*
  - Less intuitive than direct distance measures
- *Chebyshev*
  - Typically used for very specific cases
- *Minkowski*
  - Has a hyperparameter *p* which should be varied to fit the test case
- *Jaccard (~Tanimoto)*
  - Dependent on dataset size
- *Haversine*
  - Assume points lie on a sphere
- *Sorensen-Dice*
  - Weights sets inversely proportional to their size, giving potentially undue importance to some

Boundary Measures

- Define a contour around the training space (*in p dimensions, for an (n x p) training set)*
- Provide a *decision function* which gives a distance to boundary
- Any observations outside the contour (*dec. func. < 0*) are outside of AD
  - *One Class SVM*
    - Kernel method to draw a hypersphere around the training space
  - *Robust Covariance*
    - Assumes the distribution of the training data and fits an ellipse based on this assumption
  - *Isolation Forest*
    - Random forest type method which uses average path length to a training sample as a proxy for a decision function
  - *Local Outlier Factor*
    - Determines local density around training points to determine if a test point lies outside the training space