Report : Big Data Assignment 1

Dataset

Source : https://www.kaggle.com/dansbecker/aer-credit-card-data

Description : A small credit card dataset for simple econometric analysis,

Content

- **card:** Dummy variable, 1 if application for credit card accepted, 0 if not
- **reports:** Number of major derogatory reports
- **age:** Age n years plus twelfths of a year
- **income:** Yearly income (divided by 10,000)
- **owner:** 1 if owns their home, 0 if rent
- **selfempl:** 1 if self employed, 0 if not.
- **months:** Months living at current address

Methods

Imputation methods used :

1. **1 Nearest Neighbour**: Finds the nearest neighbour and imputes the value with neighbour's value.

2. **K Nearest Neighbour** (KNN where k=7) : find the nearest K (here K=7) neighbours and impute with average value of the K neighbours in case of numerical attribute. But, in case of categorical attribute impute with the **Mode** of class.

3. **Weighted KNN** – here I have used distance weighted KNN where if have modified the calculation for categorical feature to incorporate its contributions as well.

Distance measure used:

1. **Euclidian Distance (numerical) :** The distance between two points defined as the square root of the sum of the squares of the differences between the corresponding coordinates of the points;

2. **Chi-squared Distance (numerical) :** The distance between two points defined as the half of sum of the squares of the sum between the corresponding coordinates of the points divided by difference between the corresponding coordinates.

3. **Hamming Distance (categorical):** The distance between two points defined as the sum of the absolute differences between the corresponding coordinates of the points;

Feature Scaling methods used:

1. **Z-score Scaling :** $z = (X - \mu) / \sigma$
   where z is the z-score, X is the value of the element, μ is the population mean, and σ is the standard deviation

2. **Min-Max Scaling:** y= (x-min)/(max-min)
   where min and max are the minimum and maximum values in X, where X is the set of observed values of x.

Accuracy measure:

1. **For Numerical Attributes :** R^2 (coefficient of determination) regression score function.

   Best possible score is 1.0 and it can be negative (because the model can be arbitrarily worse). A constant model that always predicts the expected value of y, disregarding the input features, would get a R^2 score of 0.0.

2. **For Categorical Attributes :** Accuracy classification score.

   In multilabel classification, this function computes subset accuracy: the set of labels predicted for a sample must exactly match the corresponding set of labels in y_true.

Tools and Library:

Programming Language : Python 3.7
IDE : Jupyter Labs/ Jupyter notebook
Libraries: Numpy, Pandas, Scikit learn, Scipy

Results

Accuracy output for unscaled features

Columns

Methods

			card	reports	age	income	owner	selfemp	months
5% Missing	K=1	Euclidian	0.651515	-0.701473	-1.677208	-1.458000	0.484848	0.878788	-1.852147
		Hamming	0.696970	-0.572790	-1.673214	-1.205211	0.439394	0.893939	-1.869624
		Chi Square	0.696970	-2.217071	-1.258416	-0.893644	0.484848	0.893939	-1.117754
	K=7	Euclidian	0.757576	-0.423414	-0.686513	-0.025474	0.515152	0.939394	-0.315987
		Hamming	0.757576	-0.221051	-0.584000	-0.036608	0.545455	0.939394	-0.322893
		Chi Square	0.712121	-0.465875	-0.502183	-0.060203	0.439394	0.939394	-0.228356
	Weighted	Euclidian	0.772727	-0.071062	-0.543245	-0.090693	0.560606	0.939394	-0.231012
		Hamming	0.772727	-0.053341	-0.573477	-0.075928	0.575758	0.939394	-0.206262
		Chi Square	0.727273	-0.445153	-0.922091	-0.258225	0.515152	0.924242	-0.425157
10% Missing	K=1	Euclidian	0.696970	-0.593057	-1.703559	-0.674203	0.553030	0.886364	-1.013810
		Hamming	0.727273	-0.376130	-1.756284	-0.457344	0.545455	0.893939	-0.992117
		Chi Square	0.689394	-1.643796	-1.375845	-0.587657	0.507576	0.893939	-0.993401
	K=7	Euclidian	0.757576	-0.319702	-0.597577	-0.159445	0.507576	0.946970	-0.332576
		Hamming	0.750000	-0.352500	-0.619180	-0.148756	0.507576	0.946970	-0.309511
		Chi Square	0.689394	-0.868941	-0.746573	-0.197682	0.454545	0.946970	-0.193802
	Weighted	Euclidian	0.742424	-0.030262	-0.650815	-0.095542	0.590909	0.946970	-0.141303
		Hamming	0.742424	-0.051500	-0.680965	-0.107458	0.560606	0.946970	-0.124621
		Chi Square	0.742424	-0.564115	-1.017960	-0.259990	0.522727	0.939394	-0.217987
20% Missing	K=1	Euclidian	0.651515	-0.921923	-0.847226	-0.976188	0.515152	0.901515	-1.886306
		Hamming	0.651515	-0.584130	-0.856760	-0.878566	0.500000	0.897727	-1.973947
		Chi Square	0.643939	-0.921923	-0.860392	-0.919987	0.518939	0.878788	-2.463650
	K=7	Euclidian	0.742424	-0.142255	-0.325531	-0.270511	0.526515	0.939394	-0.634215
		Hamming	0.761364	-0.131246	-0.335584	-0.266712	0.488636	0.939394	-0.587740
		Chi Square	0.734848	-0.285704	-0.297867	-0.263862	0.492424	0.939394	-0.550112
	Weighted	Euclidian	0.768939	-0.074496	-0.311341	-0.124576	0.564394	0.939394	-0.260506
		Hamming	0.765152	-0.069332	-0.314610	-0.134083	0.564394	0.939394	-0.250164
		Chi Square	0.685606	-0.335159	-0.446534	-0.369830	0.511364	0.928030	-0.863873

Accuracy output with Z-score Scaling

Columns

Methods

			card	reports	age	income	owner	selfemp	months
5% Missing	**K=1**	**Euclidian**	0.696970	-1.845321	-2.789064	-0.911683	0.409091	0.909091	-1.579102
		Hamming	0.696970	-2.231369	-2.387223	-0.939483	0.409091	0.909091	-1.432679
		Chi Square	0.742424	-0.286828	-2.082352	-0.833151	0.636364	0.878788	-2.133153
	K=7	**Euclidian**	0.727273	-0.878175	-0.586820	-0.315867	0.439394	0.939394	-0.333468
		Hamming	0.742424	-0.949792	-0.528348	-0.263201	0.439394	0.939394	-0.316769
		Chi Square	0.742424	-0.388702	-0.626321	-0.090323	0.575758	0.939394	-0.419762
	Weighted	**Euclidian**	0.712121	-1.064645	-0.721001	-0.403442	0.500000	0.939394	-0.403825
		Hamming	0.712121	-1.133276	-0.677886	-0.343997	0.515152	0.939394	-0.458769
		Chi Square	0.757576	-0.344021	-0.728207	-0.284039	0.590909	0.939394	-0.561087
10% Missing	**K=1**	**Euclidian**	0.651515	-1.711586	-1.301599	-0.608955	0.507576	0.893939	-0.951710
		Hamming	0.666667	-2.199671	-1.245872	-0.624314	0.553030	0.901515	-1.066737
		Chi Square	0.674242	-1.325185	-1.388390	-0.501973	0.484848	0.886364	-1.961310
	K=7	**Euclidian**	0.674242	-0.838223	-0.465398	-0.175896	0.537879	0.946970	-0.419783
		Hamming	0.674242	-0.851935	-0.505592	-0.186654	0.530303	0.946970	-0.352388
		Chi Square	0.727273	-0.297162	-0.535503	-0.258734	0.469697	0.946970	-0.212413
	Weighted	**Euclidian**	0.681818	-0.902449	-0.458162	-0.223252	0.553030	0.946970	-0.399022
		Hamming	0.674242	-0.976475	-0.635089	-0.192826	0.553030	0.946970	-0.503911
		Chi Square	0.727273	-0.403517	-0.633503	-0.288940	0.469697	0.946970	-0.347306
20% Missing	**K=1**	**Euclidian**	0.666667	-1.078006	-0.565095	-0.949810	0.511364	0.901515	-1.722505
		Hamming	0.647727	-1.331933	-0.663667	-0.893368	0.515152	0.897727	-1.916606
		Chi Square	0.625000	-0.803113	-0.978113	-0.473771	0.522727	0.863636	-1.759033
	K=7	**Euclidian**	0.689394	-0.446333	-0.202359	-0.236532	0.500000	0.939394	-0.551791
		Hamming	0.678030	-0.425800	-0.165121	-0.233559	0.473485	0.939394	-0.563870
		Chi Square	0.738636	-0.090429	-0.234336	-0.161968	0.488636	0.939394	-0.290337

Accuracy using Min Max Scaling

Columns

Methods

			card	reports	age	income	owner	selfemp	months
5% Missing	**K=1**	**Euclidian**	0.681818	-1.616551	-1.670180	-1.275929	0.469697	0.909091	-1.149999
		Hamming	0.666667	-2.102686	-1.716083	-1.338062	0.469697	0.909091	-0.970179
		Chi Square	0.696970	-1.759532	-1.754554	-1.766074	0.530303	0.924242	-0.908887
	K=7	**Euclidian**	0.696970	-0.837568	-0.414757	-0.232329	0.515152	0.939394	-0.307345
		Hamming	0.712121	-1.003332	-0.387350	-0.217529	0.515152	0.939394	-0.257814
		Chi Square	0.712121	-0.807122	-0.436045	-0.190086	0.484848	0.939394	-0.299000
	Weighted	**Euclidian**	0.712121	-0.959411	-0.579451	-0.320996	0.484848	0.924242	-0.313594
		Hamming	0.712121	-1.020607	-0.623153	-0.306009	0.469697	0.924242	-0.380142
		Chi Square	0.696970	-1.002305	-0.535927	-0.265148	0.484848	0.939394	-0.321282
10% Missing	**K=1**	**Euclidian**	0.674242	-1.542112	-1.320322	-0.743742	0.545455	0.886364	-0.957064
		Hamming	0.674242	-1.623459	-1.239395	-0.686261	0.522727	0.893939	-1.047817
		Chi Square	0.689394	-1.474322	-1.412463	-0.702016	0.537879	0.878788	-0.957433
	K=7	**Euclidian**	0.704545	-0.528020	-0.547506	-0.198818	0.560606	0.946970	-0.377037
		Hamming	0.689394	-0.612773	-0.512321	-0.213056	0.583333	0.946970	-0.355359
		Chi Square	0.704545	-0.503469	-0.579411	-0.200880	0.545455	0.946970	-0.375768
	Weighted	**Euclidian**	0.674242	-0.632104	-0.546114	-0.211440	0.560606	0.939394	-0.387861
		Hamming	0.666667	-0.708299	-0.558968	-0.197302	0.560606	0.939394	-0.481031
		Chi Square	0.674242	-0.512251	-0.574895	-0.243186	0.568182	0.946970	-0.393418
20% Missing	**K=1**	**Euclidian**	0.696970	-1.210794	-0.559600	-0.979430	0.530303	0.878788	-1.788611
		Hamming	0.693182	-1.415799	-0.659350	-0.921369	0.496212	0.882576	-1.863247
		Chi Square	0.708333	-1.292330	-0.612116	-1.329382	0.534091	0.863636	-1.955068
	K=7	**Euclidian**	0.731061	-0.537302	-0.248798	-0.311168	0.537879	0.939394	-0.576034
		Hamming	0.734848	-0.521866	-0.225425	-0.311445	0.522727	0.939394	-0.573689
		Chi Square	0.723485	-0.502668	-0.239997	-0.285488	0.511364	0.939394	-0.578843
	Weighted	**Euclidian**	0.742424	-0.560182	-0.292027	-0.332229	0.556818	0.939394	-0.640154
		Hamming	0.731061	-0.554777	-0.283570	-0.361580	0.503788	0.935606	-0.717313
		Chi Square	0.731061	-0.574812	-0.248218	-0.348957	0.507576	0.931818	-0.743424

Observation/Conclusions

- Weighted KNN tends to give more accuracy as compared to KNN.
- When we apply scaling accuracy increases for each feature.