Final Project for CSCI 964

Project 1: Prediction of Automobile Fuel Consumption (20 marks)

Please build up computational intelligent models to predict the car fuel consumption in miles per gallon by using the following attributes: cylinders, displacement, horsepower, weight, acceleration, model year, origin, car name.

Attribute Information:

1. car fuel consumption in miles per gallon: continuous
2. cylinders: multi-valued discrete
3. displacement: continuous
4. horsepower: continuous
5. weight: continuous
6. acceleration: continuous
7. model year: multi-valued discrete
8. origin: multi-valued discrete
9. car name: string (unique for each instance)

Comprehensive analysis requirements:

1. Model construction and parameter optimization; (10 marks)
2. Evaluate the prediction performance on the test set. (Divide the dataset into training and testing sets, and train:test = 7:3) (10 marks)

Project 2: Prediction of Abalone Age (20 marks)

Predicting the age of abalone from physical measurements. The age of abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope -- a boring and time-consuming task. Other measurements, which are easier to obtain, are used to predict the age. Further information, such as weather patterns and location (hence food availability) may be required to solve the problem.

Attribute Information:

Given is the attribute name, attribute type, the measurement unit and a brief description. The number of rings is the value to predict.

Name / Data Type / Measurement Unit / Description

Sex / nominal / -- / M, F, and I (infant)
Length / continuous / mm / Longest shell measurement
Diameter / continuous / mm / perpendicular to length

Height / continuous / mm / with meat in shell
Whole weight / continuous / grams / whole abalone
Shucked weight / continuous / grams / weight of meat
Viscera weight / continuous / grams / gut weight (after bleeding)
Shell weight / continuous / grams / after being dried
Rings / integer / -- / +1.5 gives the age in years

Comprehensive analysis requirements:

1. Model construction and parameter optimization; (10 marks)
2. Evaluate the prediction performance on the test set. (Divide the dataset into training and testing sets, and train:test = 7:3) (10 marks)

Project 3: Salary Prediction Task (20 marks)

Prediction task is to determine whether a person makes over 50K a year.

Attribute Information:

Listing of attributes:

>50K, <=50K.

1. age: continuous.
2. workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
3. fnlwgt: continuous.
4. education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
5. education-num: continuous.
6. marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
7. occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
8. relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
   race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
9. sex: Female, Male.
10. capital-gain: continuous.
11. capital-loss: continuous.
12. hours-per-week: continuous.
13. native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

Comprehensive analysis requirements:

1. Model construction and parameter optimization; (10 marks)
2. Evaluate the prediction performance on the test set. (10 marks)

Project 4: Cardiac Arrhythmia Analysis (20 marks)

Distinguish between the presence and absence of cardiac arrhythmia and classify it in one of the 16 groups. The aim is to distinguish between the presence and absence of cardiac arrhythmia and to classify it in one of the 16 groups. Class 01 refers to 'normal' ECG classes 02 to 15 refers to different classes of arrhythmia and class 16 refers to the rest of unclassified ones. For the time being, there exists a computer program that makes such a classification. However there are differences between the cardiolog's and the programs classification. Taking the cardiolog's as a gold standard we aim to minimise this difference by means of machine learning tools.

Attribute Information:

-- Complete attribute documentation:
1 Age: Age in years , linear
2 Sex: Sex (0 = male; 1 = female) , nominal
3 Height: Height in centimeters , linear
4 Weight: Weight in kilograms , linear
5 QRS duration: Average of QRS duration in msec., linear
6 P-R interval: Average duration between onset of P and Q waves in msec., linear
7 Q-T interval: Average duration between onset of Q and offset of T waves in msec., linear
8 T interval: Average duration of T wave in msec., linear
9 P interval: Average duration of P wave in msec., linear
Vector angles in degrees on front plane of:, linear
10 QRS
11 T
12 P
13 QRST
14 J

15 Heart rate: Number of heart beats per minute ,linear

Of channel DI:
Average width, in msec., of: linear
16 Q wave
17 R wave
18 S wave
19 R' wave, small peak just after R
20 S' wave

21 Number of intrinsic deflections, linear

22 Existence of ragged R wave, nominal
23 Existence of diphasic derivation of R wave, nominal
24 Existence of ragged P wave, nominal
25 Existence of diphasic derivation of P wave, nominal
26 Existence of ragged T wave, nominal
27 Existence of diphasic derivation of T wave, nominal

Of channel DII:
28 .. 39 (similar to 16 .. 27 of channel DI)
Of channels DIII:
40 .. 51
Of channel AVR:
52 .. 63
Of channel AVL:
64 .. 75
Of channel AVF:
76 .. 87
Of channel V1:
88 .. 99
Of channel V2:
100 .. 111
Of channel V3:
112 .. 123
Of channel V4:
124 .. 135
Of channel V5:
136 .. 147
Of channel V6:
148 .. 159

Of channel DI:
Amplitude , * 0.1 milivolt, of
160 JJ wave, linear
161 Q wave, linear
162 R wave, linear
163 S wave, linear
164 R' wave, linear
165 S' wave, linear
166 P wave, linear
167 T wave, linear

168 QRSA , Sum of areas of all segments divided by 10, (Area= width * height / 2), linear
169 QRSTA = QRSA + 0.5 * width of T wave * 0.1 * height of T wave. (If T is diphasic then the bigger segment is considered), linear

Of channel DII:
170 .. 179
Of channel DIII:
180 .. 189
Of channel AVR:
190 .. 199
Of channel AVL:
200 .. 209
Of channel AVF:
210 .. 219
Of channel V1:
220 .. 229
Of channel V2:
230 .. 239
Of channel V3:

240 .. 249
Of channel V4:
250 .. 259
Of channel V5:
260 .. 269
Of channel V6:
270 .. 279

Comprehensive analysis requirements:

1. Model construction and parameter optimization; (10 marks)
2. Evaluate the prediction performance on the test set. (Divide the dataset into training and testing sets, and train:test = 7:3) (10 marks)

Project 5: Characters Classification Task (20 marks)

This database has been artificially generated by using a first order theory which describes the structure of ten capital letters of the English alphabet and a random choice theorem prover which accounts for etherogeneity in the instances. The capital letters represented are the following: A, C, D, E, F, G, H, L, P, R. Each instance is structured and is described by a set of segments (lines) which resemble the way an automatic program would segment an image. Each instance is stored in a separate file whose format is the following:

CLASS OBJNUM TYPE XX1 YY1 XX2 YY2 SIZE DIAG

where CLASS is an integer number indicating the class as described below, OBJNUM is an integer identifier of a segment (starting from 0) in the instance and the remaining columns represent attribute values.

Attribute Information:

TYPE: the first attribute describes the type of segment and is always set to the string "line". Its C language type is char.

XX1,YY1,XX2,YY2: these attributes contain the initial and final coordinates of a segment in a cartesian plane. Their C language type is int.

SIZE: this is the length of a segment computed by using the geometric distance between two points A(X1,Y1) and B(X2,Y2). Its C language type is float.

DIAG: this is the length of the diagonal of the smallest rectangle which includes the picture of the character. The value of this attribute is the same in each object. Its C language type is float.

Comprehensive analysis requirements:

1. Model construction and parameter optimization; (10 marks)
2. Evaluate the prediction performance on the test set. (Divide the dataset into training and testing sets, and train:test = 7:3) (10 marks)

Project 6: Risky Evaluation of Automobile (20 marks)

This data set consists of three types of entities: (a) the specification of an auto in terms of various characteristics, (b) its assigned insurance risk rating, (c) its normalized losses in use as compared to other cars. The second rating corresponds to the degree to which the auto is more risky than its price indicates. Cars are initially assigned a risk factor symbol associated with its price. Then, if it is more risky (or less), this symbol is adjusted by moving it up (or down) the scale. Actuarians call this process "symboling". A value of +3 indicates that the auto is risky, -3 that it is probably pretty safe.

The third factor is the relative average loss payment per insured vehicle year. This value is normalized for all autos within a particular size classification (two-door small, station wagons, sports/speciality, etc...), and represents the average loss per car per year.

symboling: -3, -2, -1, 0, 1, 2, 3.

Attribute Information:

Attribute: Attribute Range

1. normalized-losses: continuous from 65 to 256.
2. make:
   alfa-romero, audi, bmw, chevrolet, dodge, honda,
   isuzu, jaguar, mazda, mercedes-benz, mercury,
   mitsubishi, nissan, peugot, plymouth, porsche,
   renault, saab, subaru, toyota, volkswagen, volvo
3. fuel-type: diesel, gas.
4. aspiration: std, turbo.
5. num-of-doors: four, two.
6. body-style: hardtop, wagon, sedan, hatchback, convertible.
7. drive-wheels: 4wd, fwd, rwd.
8. engine-location: front, rear.
9. wheel-base: continuous from 86.6 120.9.
10. length: continuous from 141.1 to 208.1.
11. width: continuous from 60.3 to 72.3.
12. height: continuous from 47.8 to 59.8.
13. curb-weight: continuous from 1488 to 4066.
14. engine-type: dohc, dohcv, l, ohc, ohcf, ohcv, rotor.
15. num-of-cylinders: eight, five, four, six, three, twelve, two.
16. engine-size: continuous from 61 to 326.
17. fuel-system: 1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi.
18. bore: continuous from 2.54 to 3.94.
19. stroke: continuous from 2.07 to 4.17.
20. compression-ratio: continuous from 7 to 23.
21. horsepower: continuous from 48 to 288.
22. peak-rpm: continuous from 4150 to 6600.
23. city-mpg: continuous from 13 to 49.
24. highway-mpg: continuous from 16 to 54.
25. price: continuous from 5118 to 45400.

Comprehensive analysis requirements:

1. Model construction and parameter optimization; (10 marks)

2. Evaluate the prediction performance on the test set. (Divide the dataset into training and testing sets, and train:test = 7:3) (10 marks)

Project 7: Breast Cancer Prediction Task (20 marks)

This is one of three domains provided by the Oncology Institute that has repeatedly appeared in the machine learning literature. (See also lymphography and primary-tumor.)

This data set includes 201 instances of one class and 85 instances of another class. The instances are described by 9 attributes, some of which are linear and some are nominal.

Class: no-recurrence-events, recurrence-events

Attribute Information:
\2. age: 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99.
\3. menopause: lt40, ge40, premeno.
\4. tumor-size: 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59.
\5. inv-nodes: 0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-39.
\6. node-caps: yes, no.
\7. deg-malig: 1, 2, 3.
\8. breast: left, right.
\9. breast-quad: left-up, left-low, right-up, right-low, central.
\10. irradiat: yes, no.

Comprehensive analysis requirements:

1. Model construction and parameter optimization; (10 marks)
2. Evaluate the prediction performance on the test set. (Divide the dataset into training and testing sets, and train:test = 7:3) (10 marks)

Project 8:Credit Approval Evaluation Task (20 marks)

This file concerns credit card applications. All attribute names and values have been changed to meaningless symbols to protect confidentiality of the data.

This dataset is interesting because there is a good mix of attributes -- continuous, nominal with small numbers of values, and nominal with larger numbers of values. There are also a few missing values.

Attribute Information:

A1: b, a.
A2: continuous.
A3: continuous.
A4: u, y, l, t.
A5: g, p, gg.
A6: c, d, cc, i, j, k, m, r, q, w, x, e, aa, ff.
A7: v, h, bb, j, n, z, dd, ff, o.
A8: continuous.
A9: t, f.

A10: t, f.
A11: continuous.
A12: t, f.
A13: g, p, s.
A14: continuous.
A15: continuous.
A16: +,- (class attribute)

Comprehensive analysis requirements:

1. Model construction and parameter optimization; (10 marks)
2. Evaluate the prediction performance on the test set. (Divide the dataset into training and testing sets, and train:test = 7:3) (10 marks)

Project 9:Forest Cover Type Prediction Task (20 marks)

Predicting forest cover type from cartographic variables only (no remotely sensed data). The actual forest cover type for a given observation (30 x 30 meter cell) was determined from US Forest Service (USFS) Region 2 Resource Information System (RIS) data. Independent variables were derived from data originally obtained from US Geological Survey (USGS) and USFS data. Data is in raw form (not scaled) and contains binary (0 or 1) columns of data for qualitative independent variables (wilderness areas and soil types).

This study area includes four wilderness areas located in the Roosevelt National Forest of northern Colorado. These areas represent forests with minimal human-caused disturbances, so that existing forest cover types are more a result of ecological processes rather than forest management practices.

Some background information for these four wilderness areas: Neota (area 2) probably has the highest mean elevational value of the 4 wilderness areas. Rawah (area 1) and Comanche Peak (area 3) would have a lower mean elevational value, while Cache la Poudre (area 4) would have the lowest mean elevational value.

As for primary major tree species in these areas, Neota would have spruce/fir (type 1), while Rawah and Comanche Peak would probably have lodgepole pine (type 2) as their primary species, followed by spruce/fir and aspen (type 5). Cache la Poudre would tend to have Ponderosa pine (type 3), Douglas-fir (type 6), and cottonwood/willow (type 4).

The Rawah and Comanche Peak areas would tend to be more typical of the overall dataset than either the Neota or Cache la Poudre, due to their assortment of tree species and range of predictive variable values (elevation, etc.) Cache la Poudre would probably be more unique than the others, due to its relatively low elevation range and species composition.

Attribute Information:

Given is the attribute name, attribute type, the measurement unit and a brief description. The forest cover type is the classification problem. The order of this listing corresponds to the order of numerals along the rows of the database.

Name / Data Type / Measurement / Description

Elevation / quantitative /meters / Elevation in meters
Aspect / quantitative / azimuth / Aspect in degrees azimuth
Slope / quantitative / degrees / Slope in degrees
Horizontal_Distance_To_Hydrology / quantitative / meters / Horz Dist to nearest surface water features
Vertical_Distance_To_Hydrology / quantitative / meters / Vert Dist to nearest surface water features
Horizontal_Distance_To_Roadways / quantitative / meters / Horz Dist to nearest roadway
Hillshade_9am / quantitative / 0 to 255 index / Hillshade index at 9am, summer solstice
Hillshade_Noon / quantitative / 0 to 255 index / Hillshade index at noon, summer soltice
Hillshade_3pm / quantitative / 0 to 255 index / Hillshade index at 3pm, summer solstice
Horizontal_Distance_To_Fire_Points / quantitative / meters / Horz Dist to nearest wildfire ignition points
Wilderness_Area (4 binary columns) / qualitative / 0 (absence) or 1 (presence) / Wilderness area designation
Soil_Type (40 binary columns) / qualitative / 0 (absence) or 1 (presence) / Soil Type designation
Cover_Type (7 types) / integer / 1 to 7 / Forest Cover Type designation

Comprehensive analysis requirements:

1. Model construction and parameter optimization; (10 marks)
2. Evaluate the prediction performance on the test set. (Divide the dataset into training and testing sets, and train:test = 7:3) (10 marks)

Project 10:Eryhemato-Squamous Disease Prediction Task (20 marks)

This database contains 34 attributes, 33 of which are linear valued and one of them is nominal.

The differential diagnosis of erythemato-squamous diseases is a real problem in dermatology. They all share the clinical features of erythema and scaling, with very little differences. The diseases in this group are psoriasis, seboreic dermatitis, lichen planus, pityriasis rosea, cronic dermatitis, and pityriasis rubra pilaris. Usually a biopsy is necessary for the diagnosis but unfortunately these diseases share many histopathological features as well. Another difficulty for the differential diagnosis is that a disease may show the features of another disease at the beginning stage and may have the characteristic features at the following stages. Patients were first evaluated clinically with 12 features. Afterwards, skin samples were taken for the evaluation of 22 histopathological features. The values of the histopathological features are determined by an analysis of the samples under a microscope.

In the dataset constructed for this domain, the family history feature has the value 1 if any of these diseases has been observed in the family, and 0 otherwise. The age feature simply represents the age of the patient. Every other feature (clinical and histopathological) was given a degree in the range of 0 to 3. Here, 0 indicates that the feature was not present, 3 indicates the largest amount possible, and 1, 2 indicate the relative intermediate values.

The names and id numbers of the patients were recently removed from the database.

Attribute Information:

Clinical Attributes: (take values 0, 1, 2, 3, unless otherwise indicated)
1: erythema
2: scaling
3: definite borders

4: itching
5: koebner phenomenon
6: polygonal papules
7: follicular papules
8: oral mucosal involvement
9: knee and elbow involvement
10: scalp involvement
11: family history, (0 or 1)
34: Age (linear)

Histopathological Attributes: (take values 0, 1, 2, 3)
12: melanin incontinence
13: eosinophils in the infiltrate
14: PNL infiltrate
15: fibrosis of the papillary dermis
16: exocytosis
17: acanthosis
18: hyperkeratosis
19: parakeratosis
20: clubbing of the rete ridges
21: elongation of the rete ridges
22: thinning of the suprapapillary epidermis
23: spongiform pustule
24: munro microabcess
25: focal hypergranulosis
26: disappearance of the granular layer
27: vacuolisation and damage of basal layer
28: spongiosis
29: saw-tooth appearance of retes
30: follicular horn plug
31: perifollicular parakeratosis
32: inflammatory monoluclear inflitrate
33: band-like infiltrate

Comprehensive analysis requirements:

1. Model construction and parameter optimization; (10 marks)
2. Evaluate the prediction performance on the test set. (Divide the dataset into training and testing sets, and train:test = 7:3) (10 marks)