Jiusheng Chen MAT_602_HW_1 1/11/2017

Problrm 1: Write a Python program that accepts input for two 2 2 matrices of integers and multiplies them (matrix multiplication). The program should output the resulting matrix. Show how the program works using 2 different examples.

In [15]:
```
import numpy as np
# Create matrices a and b
a = np.array([[1,2],[3,4]])
b = np.array([[12,13],[2,1]])
# Multiply the matrix a and b
np.dot(a,b)
```

Out[15]:
```
array([[16, 15],
       [44, 43]])
```

In [12]:
```
# Multiply the matrix b and a
np.dot(b,a)
```

Out[12]:
```
array([[51, 76],
       [ 5,  8]])
```

In [13]:
```
# Create the matrices c and d
c = np.array([[10,10],[11,11]])
d = np.array([[3,4],[9,6]])
# Multiply the matrix c and d
np.dot(c,d)
```

Out[13]:
```
array([[120, 100],
       [132, 110]])
```

In [14]:
```
# Multiply the matrix d ang c
np.dot(d,c)
```

Out[14]:
```
array([[ 74,  74],
       [156, 156]])
```

To a conclusion, Matrix a times b is different from matrix b times a, and Matrix c times d is different from d time c.

Problem 2: Review datasets available for our course. These are at http://archive.ics.uci.edu/ml/ (http://archive.ics.uci.edu/ml/) and on our course syllabus. For five datasets that interest you, document names, the number of features and samples as well as versions available for training and testing (if any). Note any issues or errors.

Dataset #1:

Amazon book reviews Data Set http://archive.ics.uci.edu/ml/datasets/Amazon+book+reviews (http://archive.ics.uci.edu/ml/datasets/Amazon+book+reviews) Number of Instances: 213335, Number of Attributes: 4.

Data Set Information: Gone Girl: 41.974; The Girl on the Train: 37.139 ; The Fault in our Stars: 35.844; Fifty Shades of Grey: 32.977; Unbroken: 25.876; The hunger games: 24.027; The Goldfinch: 22.861; The Martian: 22.571

Attribute Information:

1. review score
2. tail of review url ([Web Link])
3. review title
4. HTML of review text

Dataset #2

Air Quality Data Set http://archive.ics.uci.edu/ml/datasets/Air+Quality (http://archive.ics.uci.edu/ml/datasets/Air+Quality) Number of Instances: 9358, Number of Attributes: 15.

Data Set Information: The dataset contains 9358 instances of hourly averaged responses from an array of 5 metal oxide chemical sensors embedded in an Air Quality Chemical Multisensor Device. The device was located on the field in a significantly polluted area, at road level,within an Italian city. Data were recorded from March 2004 to February 2005 (one year) representing the longest freely available recordings of on field deployed air quality chemical sensor devices responses. Ground Truth hourly averaged concentrations for CO, Non Metanic Hydrocarbons, Benzene, Total Nitrogen Oxides (NOx) and Nitrogen Dioxide (NO2) and were provided by a co-located reference certified analyzer. Evidences of cross-sensitivities as well as both concept and sensor drifts are present as described in De Vito et al., Sens. And Act. B, Vol. 129,2,2008 (citation required) eventually affecting sensors concentration estimation capabilities. Missing values are tagged with -200 value. This dataset can be used exclusively for research purposes. Commercial purposes are fully excluded.

Attribute Information:

1. Date (DD/MM/YYYY)
2. Time (HH.MM.SS)
3. True hourly averaged concentration CO in mg/m^3 (reference analyzer)
4. PT08.S1 (tin oxide) hourly averaged sensor response (nominally CO targeted)
5. True hourly averaged overall Non Metanic HydroCarbons concentration in microg/m^3 (reference analyzer)
6. True hourly averaged Benzene concentration in microg/m^3 (reference analyzer)
7. PT08.S2 (titania) hourly averaged sensor response (nominally NMHC targeted)
8. True hourly averaged NOx concentration in ppb (reference analyzer)
9. PT08.S3 (tungsten oxide) hourly averaged sensor response (nominally NOx targeted)
10. True hourly averaged NO2 concentration in microg/m^3 (reference analyzer)
11. PT08.S4 (tungsten oxide) hourly averaged sensor response (nominally NO2 targeted)
12. PT08.S5 (indium oxide) hourly averaged sensor response (nominally O3 targeted)
13. Temperature in Â°C
14. Relative Humidity (%)
15. AH Absolute Humidity

Dataset #3

Wine Data Set http://archive.ics.uci.edu/ml/datasets/Wine (http://archive.ics.uci.edu/ml/datasets/Wine) Number of Instances: 178, Number of Attributes: 13.

Data Set Information: These data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines.

I think that the initial data set had around 30 variables, but for some reason I only have the 13 dimensional version. I had a list of what the 30 or so variables were, but a.) I lost it, and b.), I would not know which 13 variables are included in the set.

The attributes are (dontated by Riccardo Leardi, riclea '@' anchem.unige.it) 1) Alcohol 2) Malic acid 3) Ash 4) Alcalinity of ash 5) Magnesium 6) Total phenols 7) Flavanoids 8) Nonflavanoid phenols 9) Proanthocyanins 10) Color intensity 11) Hue 12) OD280/OD315 of diluted wines 13) Proline

In a classification context, this is a well posed problem with "well behaved" class structures. A good data set for first testing of a new classifier, but not very challenging.

Attribute Information:

All attributes are continuous

No statistics available, but suggest to standardise variables for certain uses (e.g. for us with classifiers which are NOT scale invariant)

NOTE: 1st attribute is class identifier (1-3)

Dataset #4

Car Evaluation Data Set http://archive.ics.uci.edu/ml/datasets/Car+Evaluation (http://archive.ics.uci.edu/ml/datasets/Car+Evaluation) Number of Instances: 1728, Number of Attributes: 6.

Data Set Information: Car Evaluation Database was derived from a simple hierarchical decision model originally developed for the demonstration of DEX, M. Bohanec, V. Rajkovic: Expert system for decision making. Sistemica 1(1), pp. 145-157, 1990.). The model evaluates cars according to the following concept structure:

car acceptability verall price buying price maint price of the maintenance technical characteristics comfort number of doors persons capacity in terms of persons to carry the size of luggage boot estimated safety of the car

Attribute Information:

Class Values: unacc, acc, good, vgood

Attributes: buying: vhigh, high, med, low. maint: vhigh, high, med, low. doors: 2, 3, 4, 5more. persons: 2, 4, more. lug_boot: small, med, big. safety: low, med, high.

Dataset #5

Forest Fires Data Set http://archive.ics.uci.edu/ml/datasets/Forest+Fires (http://archive.ics.uci.edu/ml/datasets/Forest+Fires) Number of Instances: 517, Number of Attributes: 13.

Data Set Information: In [Cortez and Morais, 2007], the output 'area' was first transformed with a ln(x+1) function. Then, several Data Mining methods were applied. After fitting the models, the outputs were post-processed with the inverse of the ln(x+1) transform. Four different input setups were used. The experiments were conducted using a 10-fold (cross-validation) x 30 runs. Two regression metrics were measured: MAD and RMSE. A Gaussian support vector machine (SVM) fed with only 4 direct weather conditions (temp, RH, wind and rain) obtained the best MAD value: 12.71 +- 0.01 (mean and confidence interval within 95% using a t-student distribution). The best RMSE was attained by the naive mean predictor. An analysis to the regression error curve (REC) shows that the SVM model predicts more examples within a lower admitted error. In effect, the SVM model predicts better small fires, which are the majority.

Attribute Information:

1. X - x-axis spatial coordinate within the Montesinho park map: 1 to 9
2. Y - y-axis spatial coordinate within the Montesinho park map: 2 to 9
3. month - month of the year: 'jan' to 'dec'
4. day - day of the week: 'mon' to 'sun'
5. FFMC - FFMC index from the FWI system: 18.7 to 96.20
6. DMC - DMC index from the FWI system: 1.1 to 291.3
7. DC - DC index from the FWI system: 7.9 to 860.6
8. ISI - ISI index from the FWI system: 0.0 to 56.10
9. temp - temperature in Celsius degrees: 2.2 to 33.30
10. RH - relative humidity in %: 15.0 to 100
11. wind - wind speed in km/h: 0.40 to 9.40
12. rain - outside rain in mm/m2 : 0.0 to 6.4
13. area - the burned area of the forest (in ha): 0.00 to 1090.84