dem unüberwachten Lernverfahren wird ein Clustering gelernt. Das gelernte Clustering partitioniert die Trainingsdaten derart in Untergruppen, dass Datensätze innerhalb einer Untergruppe untereinander sehr ähnlich sind, Datensätze unterschiedlicher Gruppen jedoch nur ein geringes Maß an Ähnlichkeit aufweisen. Während bei der oben beschriebenen Aufgabe der Klassifizierung mit überwachtem Lernen die Klassenzugehörigkeit der Trainingsdaten schon gegeben ist, muss diese im Fall des unüberwachten Lernens erst gelernt werden. Zum Einsatz kommen derartige Verfahren u.a. für die Kategorisierung von Kunden aber auch in der Bildsegmentierung oder in der nichtlinearen Vektorquantisierung.

### 1.3.5 Externe Referenzen zur Vorbereitung

Grundlagen in

- Python: `http://www.hdm-stuttgart.de/~maucher/Python/html/index.html`
- Pandas: `http://pandas.pydata.org/pandas-docs/stable/10min.html`.
- Numpy: `http://www.scipy.org/Tentative_NumPy_Tutorial`.
- Scipy:
  - `http://docs.scipy.org/doc/scipy/reference/spatial.distance.html`
  - `http://docs.scipy.org/doc/scipy/reference/cluster.html`.
- Matplotlib: `http://matplotlib.org/users/pyplot_tutorial.html`
- Scikit Learn:
  - `http://scikit-learn.org/stable/index.html`
  - `http://www.hdm-stuttgart.de/~maucher/Python/SklearnIntro/html/index.html`

## 1.4 Vor dem Versuch zu klärende Fragen

Eine Untermenge der im Folgenden aufgeführten Fragen wird zu Beginn des Versuchs im Rahmen eines Gruppenkolloqs abgefragt. Auf jede Frage sollte von mindestens einem Gruppenmitglied eine Antwort geliefert werden und jedes Gruppenmitglied muss mindestens eine der gestellten Fragen beantworten können.

**Aus der in Kapitel 1.3 beschriebenen Theorie**

1. Erklären Sie den Sinn der *Transformation* innerhalb der Data Mining Prozesskette.
2. Worin besteht der Unterschied zwischen überwachtem und unüberwachtem Lernen.
3. Beschreiben Sie die Unterschiede zwischen Klassifikation, Regression und Clustering. Nennen Sie für diese 3 verschiedenen Verfahren je ein Anwendungsbeispiel.

**Python Allgemein:**

1. Was ist eine Python List-Comprehension?
2. Wie importiert man Daten aus einem Textfile?
3. Wie speichert man Daten aus Python in ein Textfile?
4. Wie hängt man an eine Python-Liste die Elemente einer zweiten Liste an?

**Numpy:**

1. Nennen Sie zwei verschiedene Möglichkeiten ein Numpy-Array zu erzeugen. arange(x), array([1,2,3])
2. Wie legt man ein $(3, 4)$-Array mit ausschließlich 0-Einträgen an? zeros( (3,4) )
3. Wie ruft man die Anzahl der Dimensionen, die Anzahl der Elemente pro Dimension und den Datentyp der Arrayelemente ab? a.ndim; a.shape; a.dtype
4. Wie wandelt man ein $(3, 4)$-Array in ein $(2, 6)$-Array um? a = a.reshape(2,6)
5. Wie transponiert man ein zweidimensionales Array? a.transpose()
6. Wie multipliziert man zwei Arrays elementweise? a1*a2
7. Wie führt man eine Matrixmultiplikation zweier zweidimensionaler Arrays $A$ und $B$ aus? Welche Bedingungen müssen $A$ und $B$ erfüllen, damit überhaupt eine Matrixmultiplikation durchgeführt werden kann? dot(a1,a2)
8. Wie greift man auf das Element $(2, 3)$ in einem $(4, 4)$-Array $A$ zu? Wie greift man auf die erste Spalte, wie auf die erste Zeile dieses Arrays zu? a[2-1][3-1]
9. Wie berechnet man die Quadratwurzel aller Elemente eines Arrays? sqrt(a)
10. Wie legt man eine flache Kopie, wie eine tiefe Kopie eines Arrays an? a.ravel(); a.reshape(x,y)?

**Pandas:**

1. Wie wird ein Numpy-Array in einen Pandas-Dataframe geschrieben? Wie legt man dabei die Spaltenbezeichnungen und einen Index an? df = pd.Dataframe(np.random.randn(4,4), index=pd.date_range('20131014',periods=4), columns=('Eins','Zwei',...)
2. Wie kann auf einzelne Spalten, wie auf einzelne Zeilen eines Pandas Dataframes zugegriffen werden? df['Eins]; df[4]
3. Wie können Pandas Dataframes sortiert werden? df.sort(columns='Zwei')
4. Wie kann zu einem bestehenden Dataframe eine neue Spalte hinzugefügt werden? df.append(...)
5. Wie werden Daten aus einem .csv File in einen Pandas Dataframe geschrieben? df = pd.read_csv('foo.csv')
6. Wie wird ein Pandas Dataframe in einem .csv File abgelegt? df.to_csv('foo.csv')

**Matplotlib:**

1. Wie erzeugt man mit Matplotlib einen Plot, wie er in Abbildung 4 dargestellt ist?
2. Wie kann man mehrere Graphen in einen Plot eintragen?
3. Wie erzeugt man mit Matplotlib ein Bild, das 12 Subplots in 3 Zeilen und 4 Spalten geordnet, enthält.
4. Wie erzeugt man mit Matplotlib ein Histogramm?
5. Wie erzeugt man mit Matplotlib einen Boxplot?

**Scipy:**

- Geben Sie kurz die Schritte an, die für die Durchführung eines hierarchischen Clustering mit Scipy notwendig sind.

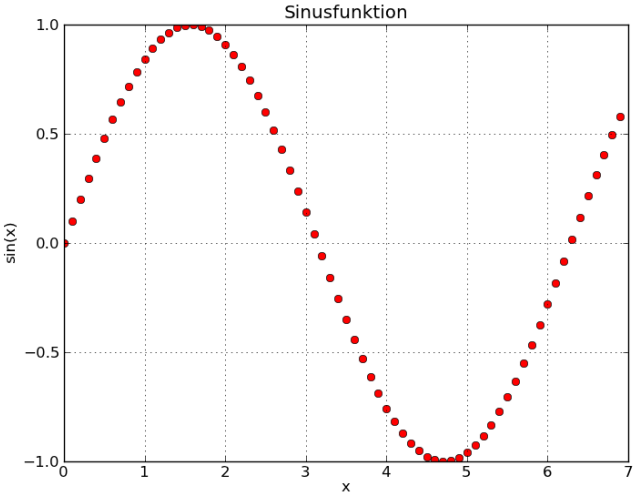


Abbildung 4: Sinusfunktion mit Matplotlib

**Scikit Learn:**

- Sklearn stellt u.a. eine umfassende Bibliothek von Klassen für das überwachte Lernen bereit. Mit welchem Methodenaufruf werden diese Kassen trainiert? Mit welchem Methodenaufruf können die trainierten Modelle auf neue Eingabedaten angewandt werden um den entsprechenden Ausgabewert zu berechnen?
- Mit welchem Leistungsmaß kann die Qualität eines Regressionsmodells bewertet werden? Wie wird dieses Leistungsmaß mit Sklearn berechnet?
- Mit welchem Leistungsmaß kann die Qualität eines Klassifikationsmodells bewertet werden? Wie wird dieses Leistungsmaß mit Sklearn berechnet?
- Was versteht man unter *x-facher Kreuzvalidierung* und wie wird diese mit Sklearn durchgeführt?