آماده‌سازی مسئله و شناخت داده ها (ورودی الگوریتم‌های یادگیری ماشین)

نادیه آرمین

گروه مهندسی کامپیوتر

دانشگاه فردوسی

پاییز ۹۹

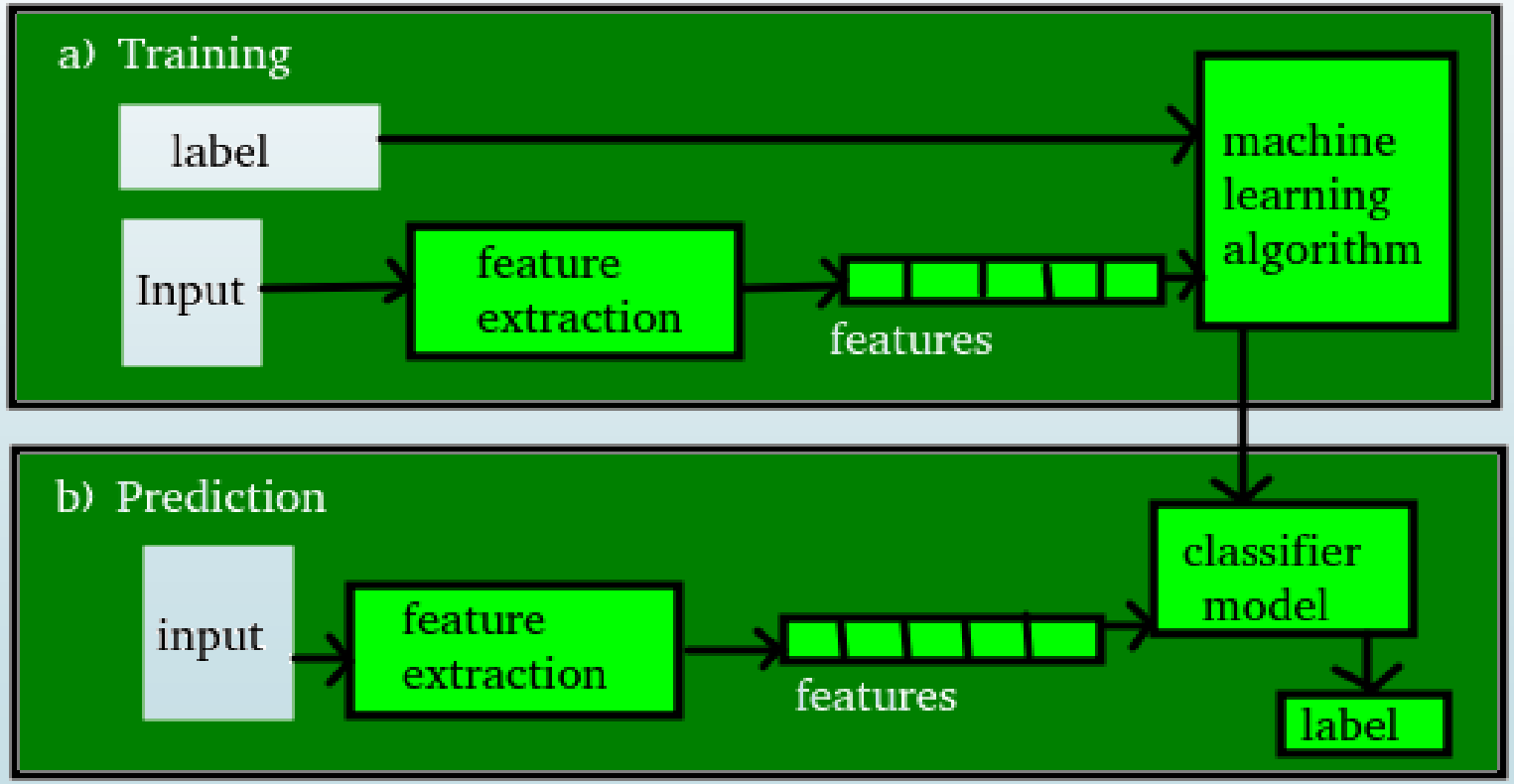
سر فصل

- قالب پروژه یادگیری ماشین
- آماده سازی مسئله:
  - بارگذاری داده ها
- شناخت داده ها :
  - با آمار توصیفی
  - با مصور سازی

پروژه یادگیری ماشین

ساخت مدل

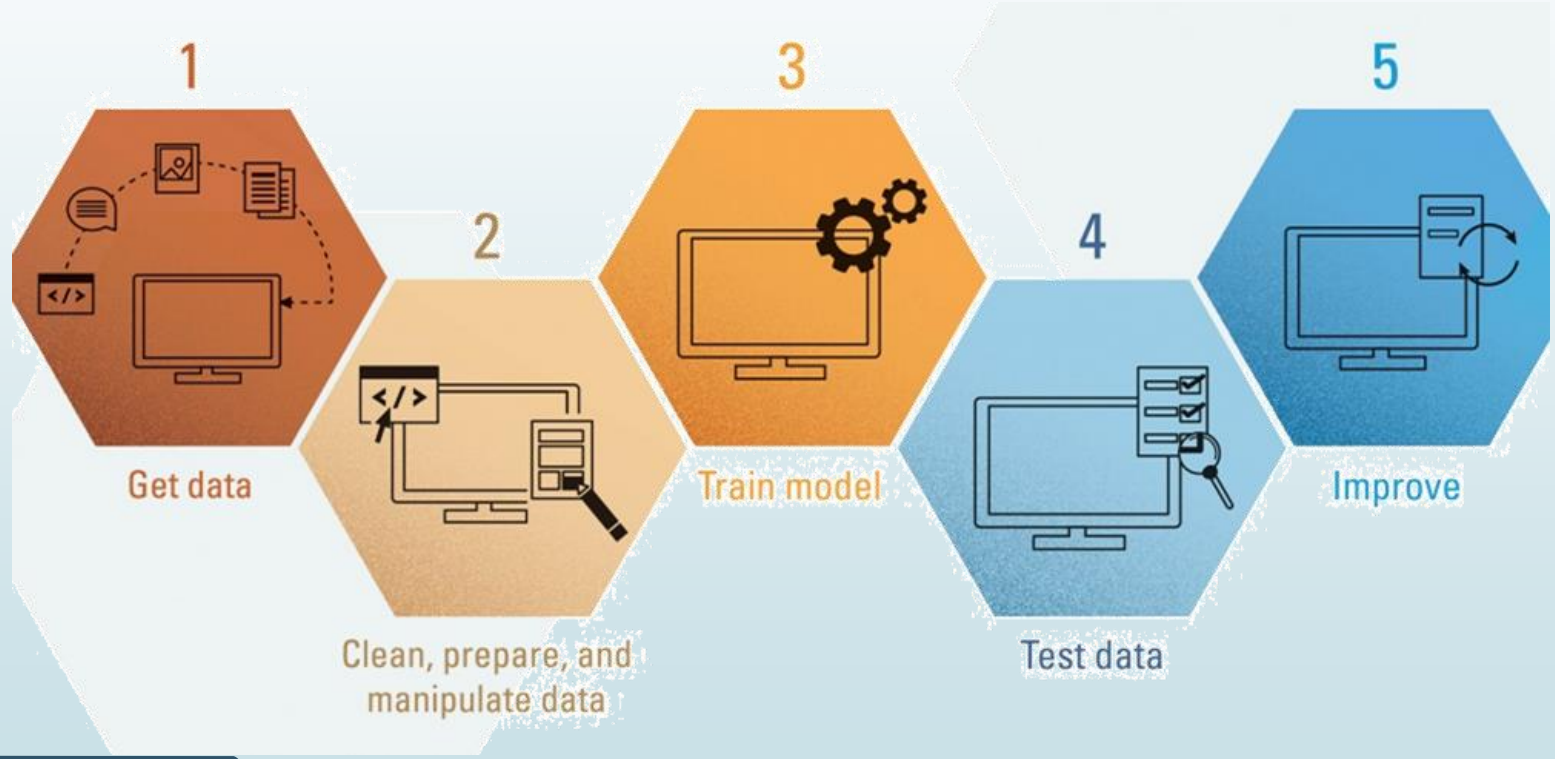
1. آموزش
2. پیش بینی



پروژه یادگیری ماشین

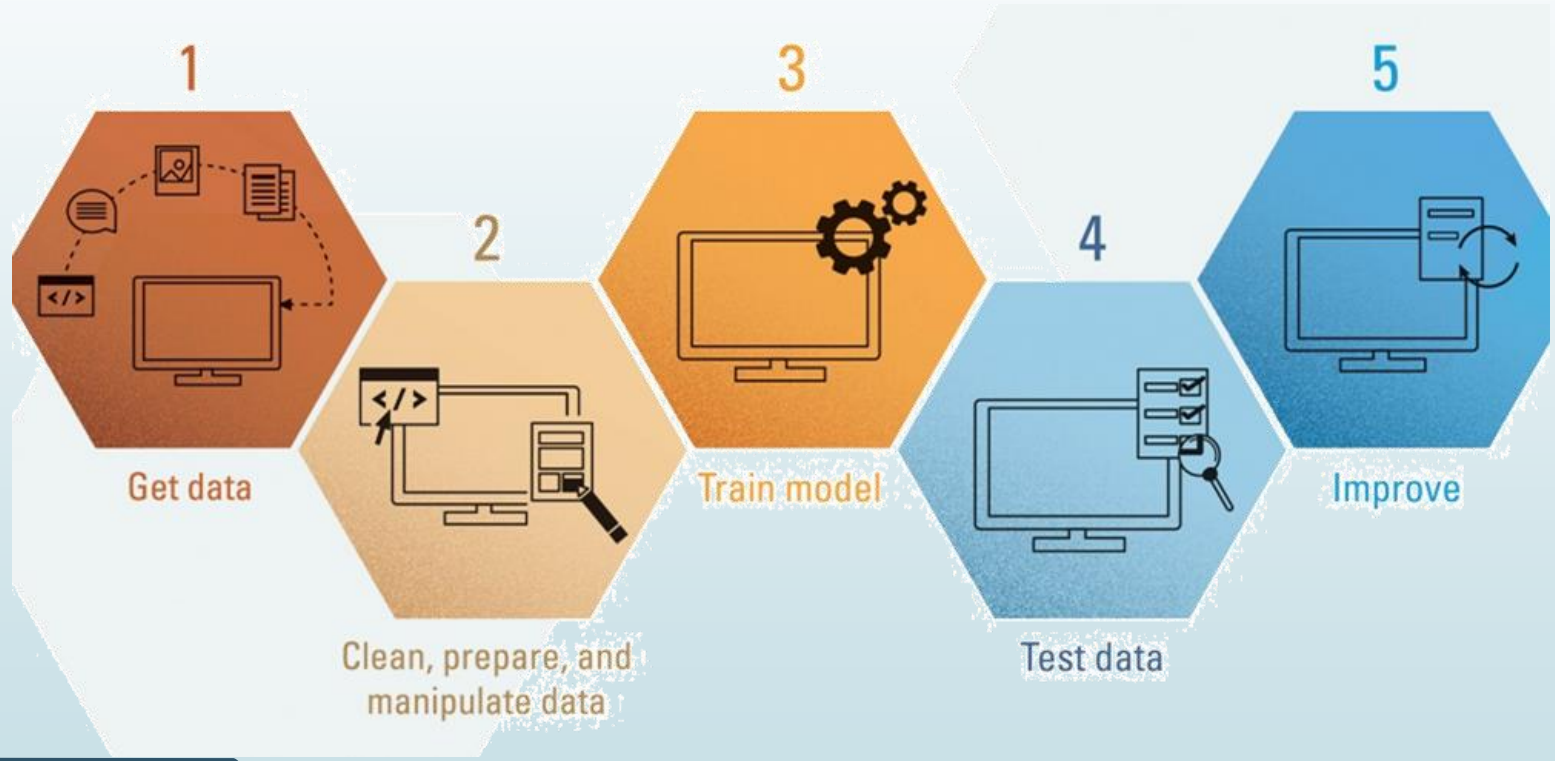
ساخت مدل

1. آموزش
2. پیش بینی



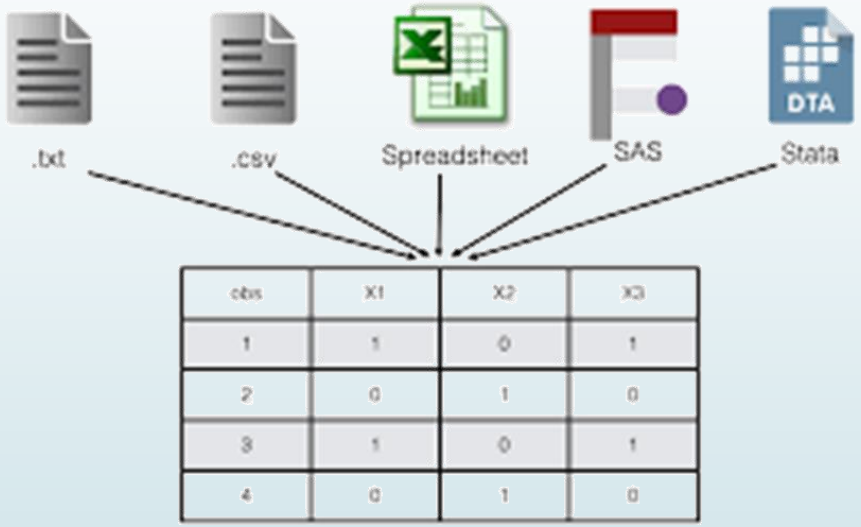
پروژه یادگیری ماشین

1. آماده سازی مسئله
2. شناخت داده ها
3. آماده سازی داده ها
4. ساخت و ارزیابی مدل
5. بهبود دقت
6. نهایی‌کردن مدل



قالب پروژه یادگیری ماشین

1. **آماده سازی مسئله:**
   a) **بارگذاری کتابخانه ها**
   b) **بارگذاری مجموعه داده**
2. شناخت داده ها
3. آماده سازی داده ها
4. ساخت و ارزیابی مدل
5. بهبود دقت
6. نهایی کردن مدل



آماده سازی مسئله

- بارگذاری مجموعه داده

```
#Load CSV Using Python Standarad Library

import csv
import numpy
filename='pima-indians-diabetes.csv'
raw_data=open(filename,'rt')
reader=csv.reader(raw_data,delimiter=',',quoting=csv.QUOTE_NONE)
x=list(reader)
data=numpy.array(x).astype('float')
print(data)
print (data.shape)
```

آماده سازی مسئله

- بارگذاری مجموعه داده

```
#Load CSV Using Numpy

from numpy import loadtxt
filename='pima-indians-diabetes.csv'
raw_data=open(filename,'rt')
data=loadtxt(raw_data,delimiter=',')
print (data.shape)
```

آماده سازی مسئله

- بارگذاری مجموعه داده

```
#Load CSV Using Pandas

from pandas import read_csv
filename='pima-indians-diabetes.csv'
names=['preg','plas','pres','skin','test','mass','pedi','age','class']
data=read_csv(filename,names=names)
print (data.shape)
```

قالب پروژه یادگیری ماشین

1. آماده سازی مسئله
2. **شناخت داده ها:**
   a) **با آمار توصیفی**
   b) با مصورسازی
3. آماده سازی داده ها
4. ساخت و ارزیابی مدل
5. بهبود دقت
6. نهایی کردن مدل

شناخت داده ها:با آمار توصیفی

- ورانداز داده‌های خام

```
from pandas import read_csv
filename = 'pima-indians-diabetes.csv'
names = ['preg', 'plas', 'pres', 'skin', 'test', 'mass', 'pedi', 'age', 'class']
data = read_csv(filename, names=names)

# view first 20 rows
peek = data.head(20)
print(peek)
```

- ابعاد داده

```
# Dimensions of your data
shape = data.shape
print(shape)
```

- نوع داده هر ستون

```
# data type for each attribute
types = data.dtypes
print(types)
```

شناخت داده ها:با آمار توصیفی

```
# Statistical Summary
from pandas import read_csv
from pandas import set_option

filename = 'pima-indians-diabetes.csv'
names = ['preg', 'plas', 'pres', 'skin', 'test', 'mass', 'pedi', 'age', 'class']
data = read_csv(filename, names=names)

set_option('precision', 3)
description = data.describe()
print(description)
```

- تعداد
- میانگین
- انحراف معیار
- مقدار کمینه
- صدک بیست و پنجم
- صدک پنجاهم (میانه)
- صدک هفتاد وپنجم
- مقدار بیشینه

شناخت داده ها:با آمار توصیفی

- توزیع کلاس (دسته بندی)

```
# Class Distribution
from pandas import read_csv

filename = 'pima-indians-diabetes.csv'
names = ['preg', 'plas', 'pres', 'skin', 'test', 'mass', 'pedi', 'age', 'class']
data = read_csv(filename, names=names)

class_count = data.groupby('class').size()
print(class_count)
```

شناخت داده ها:با آمار توصیفی

- همبستگی بین صفات

```
# Pairwise pearson correlations
from pandas import read_csv
from pandas import set_option

filename = 'pima-indians-diabetes.csv'
names = ['preg', 'plas', 'pres', 'skin', 'test', 'mass', 'pedi', 'age', 'class']
data = read_csv(filename, names=names)
set_option('precision', 3)

correlation = data.corr(method='pearson')
print(correlation)
```

شناخت داده ها:با آمار توصیفی

- چوله توزیع های یک متغییره

```
# Skew for each attribute
from pandas import read_csv

filename = 'pima-indians-diabetes.csv'
names = ['preg', 'plas', 'pres', 'skin', 'test', 'mass', 'pedi', 'age', 'class']
data = read_csv(filename, names=names)
skew = data.skew()
print(skew)
```

قالب پروژه یادگیری ماشین

1. آماده سازی مسئله
2. **شناخت داده ها:**
   a) با آمار توصیفی
   b) با مصورسازی
3. آماده سازی داده ها
4. ساخت و ارزیابی مدل
5. بهبود دقت
6. نهایی کردن مدل

شناخت داده ها:با مصور سازی

- هیستوگرام

```
from matplotlib import pyplot
from  pandas import read_csv

filename = 'pima-indians-diabetes.csv'
names = ['preg', 'plas', 'pres', 'skin', 'test', 'mass', 'pedi', 'age', 'class']
data = read_csv(filename, names=names)

# Univariate Histograms
data.hist()
pyplot.show()
```

شناخت داده ها:با مصور سازی

- نمودارهای چگالی

```
from matplotlib import pyplot
from  pandas import read_csv

filename = 'pima-indians-diabetes.csv'
names = ['preg', 'plas', 'pres', 'skin', 'test', 'mass', 'pedi', 'age', 'class']
data = read_csv(filename, names=names)

# Univariate Density Plots
data.plot(kind='density', subplots=True, layout=(3, 3), sharex=False)
pyplot.show()
```

شناخت داده ها:با مصور سازی

- نمودارهای جعبه و خط

```
from matplotlib import pyplot
from  pandas import read_csv

filename = 'pima-indians-diabetes.csv'
names = ['preg', 'plas', 'pres', 'skin', 'test', 'mass', 'pedi', 'age', 'class']
data = read_csv(filename, names=names)

# Box and Whisker Plots
data.plot(kind='box', subplots=True, layout=(3, 3), sharex=False)
pyplot.show()
```

شناخت داده ها:با مصور سازی

- نمودارهای ماتریس همبستگی

```
data = read_csv(filename, names=names)
#Plot Correlation Matrix
correlation=data.corr()
fig=pyplot.figure()
ax=fig.add_subplot(111)
cax=ax.matshow(correlation,vmin=-1,vmax=1)
fig.colorbar(cax)
ticks=numpy.arange(0,9,1)
ax.set_xticks(ticks)
ax.set_yticks(ticks)
ax.set_xticklabels(names)
ax.set_yticklabels(names)
pyplot.show()
```

شناخت داده ها:با مصور سازی

- نمودارهای ماتریس همبستگی

```
from matplotlib import pyplot
from  pandas import read_csv
filename = 'pima-indians-diabetes.csv'
names = ['preg', 'plas', 'pres', 'skin', 'test', 'mass', 'pedi', 'age', 'class']
data = read_csv(filename, names=names)
 # Corelation Matrix Plot (generic)
correlation=data.corr()
#plot correlation matrix
fig=pyplot.figure()
ax=fig.add_subplot(111)
cax=ax.matshow(correlation,vmin=-1,vmax=1)
fig.colorbar(cax)
pyplot.show()
```

شناخت داده ها:با مصور سازی

- ماتریس نمودار پراکنشی

```
from matplotlib import pyplot
from  pandas import read_csv
filename = 'pima-indians-diabetes.csv'
names = ['preg', 'plas', 'pres', 'skin', 'test', 'mass', 'pedi', 'age', 'class']
data = read_csv(filename, names=names)

#Scatterplot Matrix
from pandas.plotting import scatter_matrix
scatter_matrix(data)
pyplot.show()
```

خلاصه

- ✓قالب پروژه یادگیری ماشین
- ✓آماده سازی مسئله: بارگذاری داده ها
- ✓شناخت داده ها :
  - ✓ با آمار توصیفی:
    ورانداز داده‌های، ابعاد داده، انواع داده، توزیع کلاس، خلاصه داده‌ها، همبستگی‌ها، چولگی
  - ✓با مصور سازی:
    هیستوگرام، نمودارهای چگالی، جعبه و خط، ماتریس همبستگی و ماتریس نمودار پراکنشی