Sample

February 3, 2020

1 Geron, Appendix B: ML Project Check List

- Define the problem
- Describe the Data
  - where does it come from
- Exploratory data analysis
  - visualize, gain insights
  - find potentially useful features
- Data cleaning
  - problems with the data and how we fixed them
- Data transformation
  - pre-processing data
- Experiments
  - describe a hypotheses, experiment and result
  - iterate

[1]:
```
from IPython.core.interactiveshell import InteractiveShell
InteractiveShell.ast_node_interactivity = "all"

# Reload all modules imported with %aimport
%load_ext autoreload
%autoreload 1

%matplotlib notebook
```

[2]:
```
import matplotlib.pyplot as plt
import numpy as np
```

[3]:
```
def gen_data(num):
    """
    Function to generate random data

    Parameters
    ----------
    num: Integer.  Number of pairs to generate

    Returns
```

```
    -------
    Tuple (X,Y):
    - X NumPy ndarray, shape (num, 1)
    - Y NumPy ndarray, shape (num, 1)
    """
    v = 2
    t = np.arange(0, num).reshape(-1,1)
    d= (v * t) + np.random.normal(0, np.sqrt(num), (num,1))

    return t,d

X, Y = gen_data(50)
```

[4]:
```
# np.<TAB>
# np.random.<TAB>
```

[10]:
```
gen_data?
```

[11]:
```
gen_data??
```

2 Problem Description

This is the problem I'm trying to solve. Here are the key points to know: - First point - Second point

3 Data

The data was obtained from scraping the web.

3.1 Exploratory Data Analysis

The distribution of the data is:

[7]:
```
print("X: mean={:3.2f}, std={:3.2f}".format(X.mean(), X.std()))
print("Y: mean={:3.2f}, std={:3.2f}".format(Y.mean(), Y.std()))
```

```
X: mean=24.50, std=14.43
Y: mean=49.03, std=29.38
```

[8]:
```
fig = plt.figure()

ax_x = fig.add_subplot(1,2, 1)
_ = ax_x.hist(X)
_ = ax_x.set_xlabel("Time")
```

```
_ = ax_x.set_ylabel("Count")
_ = ax_x.set_title("X")

ax_y = fig.add_subplot(1,2, 2)
_ = ax_y.hist(Y)
_ = ax_y.set_xlabel("Distance")
_ = ax_y.set_ylabel("Count")
_ = ax_y.set_title("Y")
```

```
<IPython.core.display.Javascript object>
```

```
<IPython.core.display.HTML object>
```

3.2 Data cleaning

There were a number of observations that were incomplete, i.e., had at least one feature missing Furthermore, there were features with clearly incorrect values (e.g., negative or zero Price).

We cleaned the data as follows: - eliminate rows in which any feature is missing - eliminate rows in which any feature has a clearly incorrect

We considered replacing clearly incorrect values with various proxies but rejected that approach b/c …

3.3 Data transformation

Because different features had widely different ranges (i.e., min and max) we first transformed the data as follows: - standardized (mean 0, unit variance) variables …

3.4 Experiments

We started by examining the hypothesis that Price was linear in Widget color .. Here is a Linear model with associated accuracy.