knn-higgs-accuracy-vs-k-sklearn

September 2, 2021

[1]:
```
import pandas as pd
```

[2]:
```
full_higgs_data = pd.read_csv('HIGGS.csv.gz')
```

[3]:
```
full_higgs_data.to_hdf('higgs.hdf5', 'higgs')
```

[4]:
```
full_higgs_data = pd.read_hdf('higgs.hdf5', 'higgs')
```

[5]:
```
n_samples, n_features = full_higgs_data.shape
```

[6]:
```
n_samples, n_features
```

[6]:
```
(10999999, 29)
```

[7]:
```
train_samples, test_samples = 100000, 50000
```

[8]:
```
train_data, train_labels = full_higgs_data.iloc[0:train_samples, 1:],␣
 ↪full_higgs_data.iloc[0:train_samples, 0]
```

[9]:
```
train_data.shape, train_labels.shape
```

[9]:
```
((100000, 28), (100000,))
```

[10]:
```
test_data, test_labels = full_higgs_data.iloc[train_samples:(train_samples +␣
 ↪test_samples), 1:], full_higgs_data.iloc[train_samples:(train_samples +␣
 ↪test_samples), 0]
```

[11]:
```
test_data.shape, test_labels.shape
```

[11]:
```
((50000, 28), (50000,))
```

[12]:
```
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import accuracy_score
```

[13]:
```
njobs = 8
```

[14]:

```
def handle_k(k: int) -> float:
    classifier = KNeighborsClassifier(n_neighbors = k, algorithm = 'brute',␣
 ↪n_jobs = njobs)
    classifier.fit(train_data, train_labels)
    result = classifier.predict(test_data)
    return accuracy_score(test_labels, result)
```

[15]:

```
ks = list(range(1, 10)) + list(range(10, 110, 10))
rs = []

for k in ks:
    res = handle_k(k)
    print(f'K: {k}, Accuracy: {res}')
    rs.append(res)
```

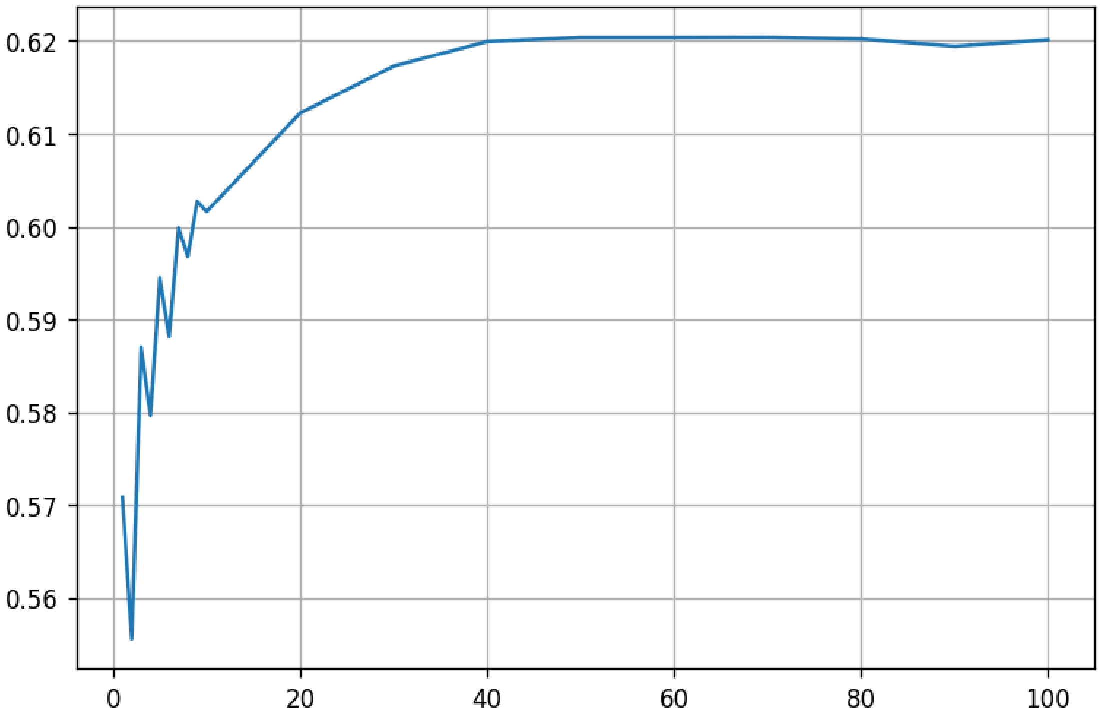
```
K: 1, Accuracy: 0.57084
K: 2, Accuracy: 0.55558
K: 3, Accuracy: 0.58704
K: 4, Accuracy: 0.57966
K: 5, Accuracy: 0.59452
K: 6, Accuracy: 0.58814
K: 7, Accuracy: 0.59986
K: 8, Accuracy: 0.5968
K: 9, Accuracy: 0.60274
K: 10, Accuracy: 0.60162
K: 20, Accuracy: 0.61224
K: 30, Accuracy: 0.6173
K: 40, Accuracy: 0.61994
K: 50, Accuracy: 0.62038
K: 60, Accuracy: 0.62038
K: 70, Accuracy: 0.6204
K: 80, Accuracy: 0.62022
K: 90, Accuracy: 0.61944
K: 100, Accuracy: 0.62012
```

[16]:

```
import matplotlib.pyplot as plt
```

[17]:

```
plt.figure(figsize = (7.5, 5), dpi =  120)
plt.plot(ks, rs)
plt.grid()
plt.savefig('higgs-accuracy-knn-vs-k.pdf')
```



```
[18]: print(rs[-1])
```

```
0.62012
```

```
[ ]:
```