genre_model_upper_bound_to_accuracy

June 1, 2020

1 Calculate the upper bound for accuracy of any model trained on our training data.

The data is of the form (X, y) with $X_i \in \{0,1\}^{\times p}$ $(p = 1494)$, and $y \in \{0,1\}$. There are 12376 training samples. Let $\{\bar{X}_a\}_{1 \leq a \leq 6230}$ be unique representatives of the inputs in the training set; That is, for all i there exists a such that $X_i = \bar{X}_a$. For each $\bar{X}_a$ the number of female artists $(\mathrm{fem}\left(\bar{X}_a\right))$ and male artists $\left(\mathrm{mal}(\bar{X}_a)\right)$ with $X_i = \bar{X}_a$ are calculated. Define a classifier on the set of training data $f_0 : \{X_i\}_{i=1}^{12376} \to \{0,1\}$ as

$$f(X_i) = \mathrm{argmax}_{\{\text{male,female}\}} \left\{\mathrm{mal}(\bar{X}_a), \mathrm{fem}(\bar{X}_a)\right\} \text{ if } X_i = \bar{X}_a$$

Then extend f_0 to $f : \{0,1\}^{\times p} \to \{0,1\}$. When f is only used on the training data, the extension from f_0 to f is irrevelant, and f_0 gives rise to an optimal classifier. However, to generalize to data which includes points in $\{0,1\}^{\times p}$ that were not in the training set, a rule is needed to make the extension.

This notebook shows that even on the training data f_0 has an expected error of 26.8%, or an accuracy of 73.2%.

Questions: - for the DNN classifier the 1-fold CV accuracy has a mean of 76% with std 1%. How? - for the DNN classifier the training accuracy can be close to 80%. How? Is it memorizing the order and particular

[4]:

```
import numpy as np
import pandas as pd

import matplotlib.pyplot as plt
import seaborn as sns; sns.set()

import re
```

Import the cleaned data:

[5]:

```
#%ls -lt ../../data/genre_lists/data_ready_for_model/
```

[6]:

```
%store -r now
now
#now = '2020-05-11-14-35'
```