extracredit

March 21, 2024

1 Supervised Learning Coding Extra Credit

1.0.1 1. Take a random sample of 100 rows from teacher_utterance_labels. (df.sample(50))

[ ]:
```
import pandas as pd
import numpy as np
```

[ ]:
```
df = pd.read_excel("/Users/giovanni-lunetta/uconn_masters/s2/epsy_5643/
↪Assignment-2/Assignment-2/teacher_utterance_labels.xlsx")
df.sample()
```

[ ]:
```
      transcript_id  utterance_id  teacher_utterance_number  \
3983            107         11578                        44

                                                   text         gold_standard
3983  oh I see thumbs going up so fast liam one don'…  classroom management
```

[ ]:
```
df["math_instruction"] = np.where(df.gold_standard == "math instruction", 1, 0)
df.math_instruction.value_counts()
```

[ ]:
```
math_instruction
1    8564
0    2238
Name: count, dtype: int64
```

[ ]:
```
sampled_df = df.sample(100, random_state=2896849)
```

[ ]:
```
sampled_df["math_instruction"].value_counts()
```

[ ]:
```
math_instruction
1    79
0    21
Name: count, dtype: int64
```

1.0.2 2. Export the utterance_id and text to an excel file (.to_excel). Do not export the other columns.

[ ]:
```
export_df = sampled_df[['utterance_id', 'text']]
```

[ ]:
```
# export_df.to_excel('/Users/giovanni-lunetta/uconn_masters/s2/epsy_5643/
↪Assignment-2/Assignment-2-Extra-Credit/utterances_sample.xlsx', index=False)
```

1.0.3 3. Create a new column called math_instruction_gold. Then, in each row, classify the utterance by hand according to your codebook. 1 = math instruction

Column created in excel file

1.0.4 4. Calculate accuracy, recall, and precision between your hand codes and the original "gold_standard" I provided.

[ ]:
```
df_gold_standard = pd.read_excel('/Users/giovanni-lunetta/uconn_masters/s2/
↪epsy_5643/Assignment-2/Assignment-2-Extra-Credit/utterances_sample.xlsx')
```

[ ]:
```
sampled_df = sampled_df.set_index('utterance_id')
df_gold_standard = df_gold_standard.set_index('utterance_id')
```

[ ]:
```
merged_df = df_gold_standard.merge(sampled_df['math_instruction'],␣
↪left_index=True, right_index=True)

merged_df
```

[ ]:
```
                                                          text  \
utterance_id
11524          Could you count that? Where would you count fi…
6437           Nolan, tell me how you did it. Oh, okay. Would…
1679           So let's go ahead and read it all again. Jane …
213            That's okay, too, because this is our first we…
1327           it's telling you to add. And what if you're lo…
…                                                            …
10921          These two numbers are representing what  60 an…
1173           That's fine. There's word problems on the back…
7338           By ones the way Elise did? Or by 10s? The way …
5727           Yeah, yeah. Is everyone listening? Yeah. If yo…
6382           Good. So if you started if you were doing that…

               math_instruction_gold  math_instruction
utterance_id
11524                              1                 1
6437                               1                 1
1679                               1                 1
213                                0                 1
```

```
1327                                1                   1
…                                 …                   …
10921                               1                   1
1173                                0                   0
7338                                1                   1
5727                                0                   0
6382                                1                   1

[100 rows x 3 columns]
```

[ ]:
```
TP = ((merged_df['math_instruction_gold'] == 1) &␣
 ↪(merged_df['math_instruction'] == 1)).sum()
FP = ((merged_df['math_instruction_gold'] == 1) &␣
 ↪(merged_df['math_instruction'] == 0)).sum()
TN = ((merged_df['math_instruction_gold'] == 0) &␣
 ↪(merged_df['math_instruction'] == 0)).sum()
FN = ((merged_df['math_instruction_gold'] == 0) &␣
 ↪(merged_df['math_instruction'] == 1)).sum()

accuracy = (TP + TN) / (TP + TN + FP + FN)
recall = TP / (TP + FN)
precision = TP / (TP + FP)

print(f"Accuracy: {accuracy}")
print(f"Recall: {recall}")
print(f"Precision: {precision}")
```

```
Accuracy: 0.95
Recall: 0.9367088607594937
Precision: 1.0
```

Results:

- With a precision of 1, we can say that there were no instances where we said it math instruction and you said it wasnt.
- With a recall of ~94%, we can say that there were a few instances where we said it was classroom manegeament and you said it was math instruction.
- With an accuracy of 95%, we can say that there we agreed on classificaiton 95% of the time.