Documentation: Coding

**Overall take-aways**

EM algorithm already converges after about 5 iterations. More iterations do not really improve the overall result.

1. **Hard EM real data unbalanced**

**Hyperparameter**

* Number bins = 5
* Iterations = 5

**Results**

**Label Distribution, unbalanced**

|  |  |
| --- | --- |
| Label | Count |
| 0 | 26891 (~45%) |
| 2 | 18429 (~31%) |
| 1 | 14285 (~24%) |
| Total | 59605 |

**Classification Report**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 0.456984 | 0.981775 | 0.623670 | 5432.000000 |
| 1 | 0.000000 | 0.000000 | 0.000000 | 2839.000000 |
| 2 | 0.370518 | 0.025479 | 0.047680 | 3650.000000 |
| accuracy | 0.455163 | 0.455163 | 0.455163 | 0.455163 |
| macro avg | 0.275834 | 0.335751 | 0.223783 | 11921.000000 |
| weighted avg | 0.321678 | 0.455163 | 0.298784 | 11921.000000 |

Interpretation:

* Very high recall for class 0
* Very low recall for class 1 and 2
* About 45% accuracy (about 45% of labels are 0)
* Model mainly predicts class 0 and in very rare cases class 2.

1. **Hard EM real data balanced**

**Hyperparameter**

* Number bins = 5
* Iterations = 5

**Label Distribution, balanced**

|  |  |
| --- | --- |
| Label | Count (after balancing) |
| 0 | 14285 |
| 1 | 14285 |
| 2 | 14285 |

**Classification Report**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 0.351668 | 0.474407 | 0.403919 | 2911.000000 |
| 1 | 0.336308 | 0.293555 | 0.313481 | 2824.000000 |
| 2 | 0.343277 | 0.263752 | 0.298305 | 2836.000000 |
| accuracy | 0.345117 | 0.345117 | 0.345117 | 0.345117 |
| macro avg | 0.343751 | 0.343905 | 0.338568 | 8571.000000 |
| weighted avg | 0.343831 | 0.345117 | 0.339175 | 8571.000000 |

Interpretation:

* Precision and recall are better distributed among all labels
* Accuracy is slightly better than guessing (guessing would be 33%)

1. **Soft EM real data unbalanced**

**Hyperparameter**

* Number bins = 5
* Iterations = 5

**Results**

**Classification Report**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 0.455666 | 1.000000 | 0.626059 | 5432.000000 |
| 1 | 0.000000 | 0.000000 | 0.000000 | 2839.000000 |
| 2 | 0.000000 | 0.000000 | 0.000000 | 3650.000000 |
| accuracy | 0.455666 | 0.455666 | 0.455666 | 0.455666 |
| macro avg | 0.151889 | 0.333333 | 0.208686 | 11921.000000 |
| weighted avg | 0.207632 | 0.455666 | 0.285274 | 11921.000000 |

Interpretation: See 1)

1. **Soft EM real data balanced**

**Hyperparameter**

* Number bins = 5
* Iterations = 5

**Results**

**Classification Report**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 0.000000 | 0.000000 | 0.000000 | 2911.000000 |
| 1 | 0.329483 | 1.000000 | 0.495656 | 2824.000000 |
| 2 | 0.000000 | 0.000000 | 0.000000 | 2836.000000 |
| accuracy | 0.329483 | 0.329483 | 0.329483 | 0.329483 |
| macro avg | 0.109828 | 0.333333 | 0.165219 | 8571.000000 |
| weighted avg | 0.108559 | 0.329483 | 0.163310 | 8571.000000 |

Interpretation: See 1) but here only class 2 is predicted.

* Since balanced data is used but the model still overfits one specific class, the problem might be the initialization of the EM algorithm
* Use better initialization methods

1. **Soft EM real data balanced with KMeans Initialization**

**Hyperparameter**

* Number of bins = 5
* Iterations = 5

**Results**

**Classification Report**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 0.404972 | 0.173480 | 0.242905 | 2911.000000 |
| 1 | 0.339497 | 0.741147 | 0.465680 | 2824.000000 |
| 2 | 0.383952 | 0.156911 | 0.222778 | 2836.000000 |
| accuracy | 0.355034 | 0.355034 | 0.355034 | 0.355034 |
| macro avg | 0.376140 | 0.357179 | 0.310455 | 8571.000000 |
| weighted avg | 0.376444 | 0.355034 | 0.309646 | 8571.000000 |

Interpretation:

* Predictions are more distributed as without KMeans initialization
* Accuracy is slightly better than guessing (guessing would be 33%)