EP2420 - Project 1: Task IV

André Silva

November 13, 2020

1 Task IV

Since we are analysing *VoD Flashcrowd*, we are going to convert our target values into an histogram of 31 labels, with midpoints 0, 1, 2, ..., 30.

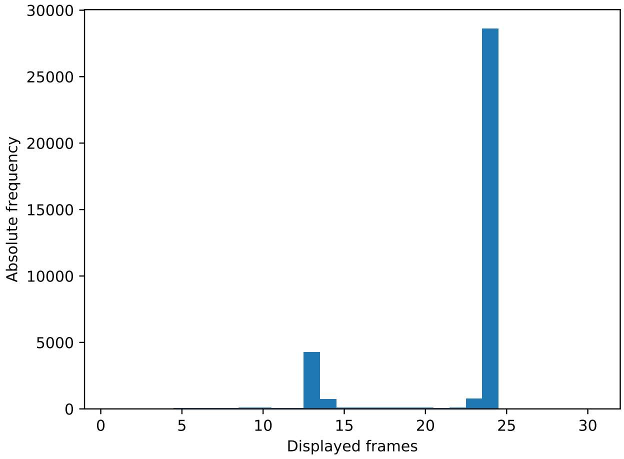

Figure 1: Histogram of the video frame rate on the interval $y \in [-0.5, 30.5]$, with a bin size of 1

We then take each bin of the histogram as a class and fit a *RandomForestClassifier* [1] to the data, which has been target of standardization by column and outlier removal.

The calculated NMAE for this model was approximatly 7.3%. In *Task II*, we had been able to train a model which got 8.7% NMAE, which was then marginally improved in *Task III*.

We can conclude by these results that the discretization of our target values allows us to better predict them. The loss of information as a result of this process is negligible, given the nature of the target values.

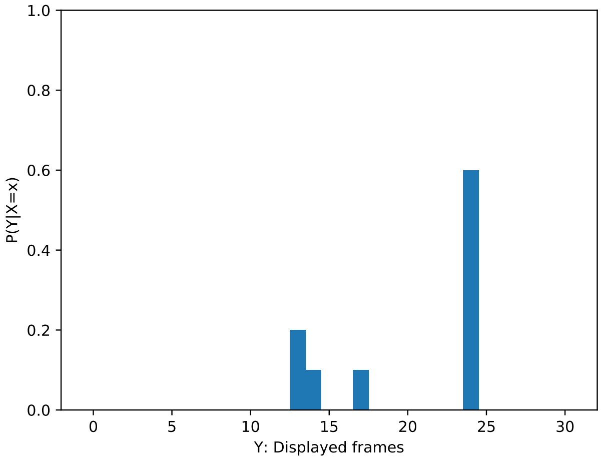

Figure 2: Predicted histogram for a randomly selected sample of the test set. The measured target value was 24.0

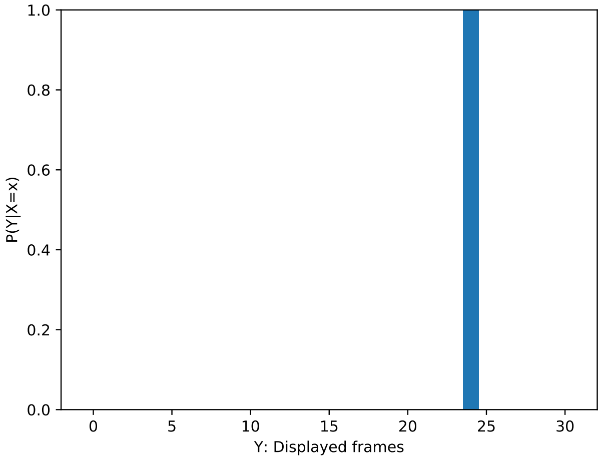

Figure 3: Predicted histogram for a randomly selected sample of the test set. The measured target value was 24.0

In Figures 2 and 3, we can see the predicted histograms for two randomly selected samples of the test set, both with a measured value of 24.0.

In the first case, $P(Y|X)$ is distributed between four classes, 13, 14, 17 and 24, whereas in the second case it is concentrated entirely on class 24. We can extract from this that, when the system is in a state equivalent to the first sample, it is more unstable relative to the second, and that we can expect some disturbances in the video frame rate registered by the client.

References

[1] *sklearn.ensemble.randomforestclassifier*. `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html#sklearn.ensemble.RandomForestClassifier`. accessed: 2020-11-13.