Model calibration

Model calibration is used to align the output of a classification model to probability prediction, e.g., instances with .05 model score has approximately 5% of chance of being positive. Summaries from the two articles listed in references are as follows:

- In practice, model calibration should be conducted on out-of-sample set to avoid overfitting.
- There are two major algorithms for model calibration
  - Platt/sigmoid scaling is equivalent of training another logistic regression model on the (raw score, label) pairs, i.e., estimate A and B in the following

$$p(y = 1|x) = \frac{1}{1 + \exp\left[Af(x) + B\right]}$$

  - Fit an isotonic regression model on $f^{(i)}(x), y^{(i)}$ pairs
- Reliability curve is often used to measure how well the probabilistic predictions of a classifier is calibrated.
  - Note that instead of decile/ventile the whole sample, we divide the sample into pre-set score bucket, i.e., $[0.0 - 0.1], ..., [0.9 - 1.0]$ and plot the mean `dep_var` with count/frequency at the bottom (see an scikit-learn example).
- In practice, the larger the calibration set, the better the results. Platt scaling works better when the calibration set is small. When the set is large, isotonic regression is often preferred as it is non-parametric (easy to fit) and works better.
- If the negative class is downsampled with probability w before training, the raw model scores need to be calibrated by the following formula:

$$q = \frac{p}{p + (1 - p)/w}$$

References

- (In Chinese) Zhihu - model calibration by LLL
- scikit-learn - 1.16. Probability calibration