Programming Project 3: Calibrated Probability Estimation

Aman Raj, PID: A53247556

February 23, 2019

1 Dataset

We used the sentiment data provided for this project that contains 3000 samples with two class labels 0 and 1. We split this data into three different balanced datasets namely - *training set*(for learning SVM), *calibration set*(for learning monotonic transformations) and *test set*(for evaluating probability estimates). The data was pre-proccessed by applying standard techniques and then vectorized using bag-of-words with 5000 words in vocabulary.

2 Learning a Classifier: SVM

We train a soft-margin SVM using training data and select the optimum value of parameter C using 5-fold cross-validation. As we can see in Figure 1 the optimum value of C which gives lowest cross-validation error is *C=0.1131*. Next, we train the SVM classifier from scratch with optimum C and evaluate on *test set*, resulting in error of *0.2074*.

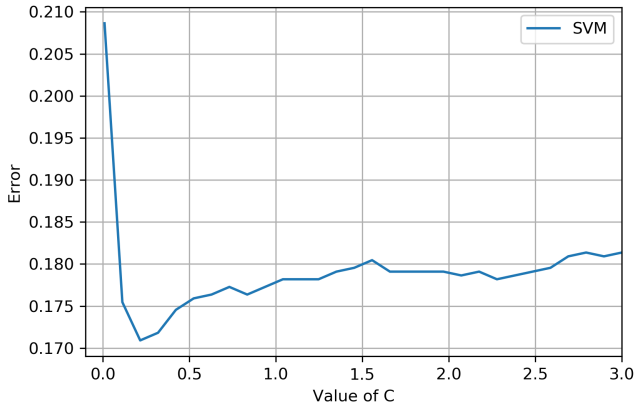


Figure 1: Variation of cross-validation error over C parameter for linear SVM.

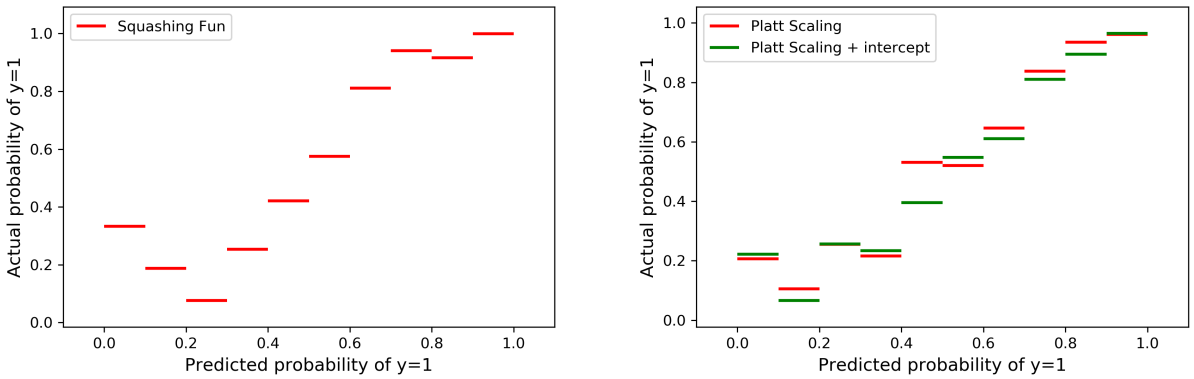
3 Learning a Monotonic Transformation

Now we want to learn a function that converts real-valued scores z obtained from SVM to probabilities. We investigated with following three choices:

1. **Squashing function**: Maps the z score to probabilities using the function:

$$1/(1+e^{-z})$$

Reliability diagram for such kind of mapping is shown in Figure 2a. As we can observe from the diagram this is not a calibrated probability estimate.



(a) Squashing Function

(b) Plat Scaling Comparison

Figure 2: Reliability diagram for squash function and platt scaling

For example, let's take a general interval $[p-\delta, p+\delta]$, now for $[0.0, 0.1]$ where p=*0.05*, and $\delta = 0.05$, actual probability of y=*1* is around *0.35* which is very high than value of p, hence predicted probability is an overestimate. Similarly, for most of the intervals in the graph predicted probability is either overestimate or underestimate.

2. **Platt scaling**: In this case we learn the mapping from z score to the probability values by:

$$1/(1 + e^{-(az+b)})$$

using the calibration set C. We tried two approaches, first when we also learn the intercept term b along with a and second when we only learn the parameter a. The reliability diagram for both methods is shown in the Figure 2b.

As we can observe from the graph platt scaling with intercept term b in general performs better than without b platt. For example the overestimate for interval $[0.4, 0.5]$ is decreased. Similar, it has improved the estimate for other intervals except few. Platt scaling with zero intercept forces the decision boundary to pass through the origin which is not true in general hence by using intercept term we provide flexibility to the algorithm to learn an optimum decision boundary.

Also, Platt scaling is better calibrated than squash function because in platt we are learning the mapping using a dataset called calibration set rather than plainly mapping z score to probability values. The same can be observed from reliability diagram, e.g. for interval $[0.0, 0.1]$ platt scaling gives better estimate of true probability than squash function.

3. **Isotonic regression**: In isotonic regression the idea is to learn a general monotonic function from z to calibrated probability score. Reliability graph of this method is shown in Figure 3. Compared with platt scaling it gives better calibrated probabilities for some intervals such as estimate for $[0.0, 0.1]$ is now close to true probability but at the same time the estimate for some intervals have worsen.

4 Critical Evaluation

Platt scaling is better suited than isotonic regression for cases when the size of calibration dataset is small as it gives better approximation of calibrated probabilities than isotonic regression. It is particularly effective for max-margin methods such as SVMs as we can see in our experimentation. Hence, the choice of algorithm between two depends heavily on the application, the type of linear classifier used for generating z score and the size of calibration set. In general we should choose isotonic regression when enough calibration dataset is available and or else Platt scaling.

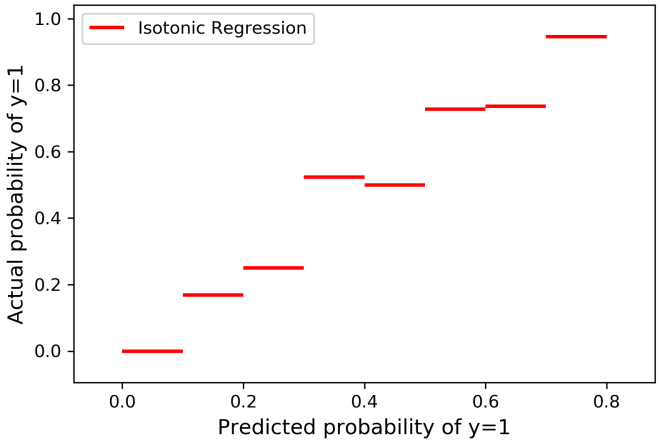


Figure 3: Reliability diagram for Isotonic Regression

In case of multiclass setting, the calibrated probability for each class can be calculated by using one versus rest methodology, this way creating a binary probability calibration problem that we have already addressed above. Hence, for each class we can learn this binary probability calibration problem using either isotonic regression or platt scaling.

From our experiment we realized that platt scaling tried to fit a sigmoid function and is suitable if we have high confidence that appropriate calibration is of that form whereas isotonic regression fits a piece-wise constant function. Also, we know that platt scaling is a parametric approach while isotonic regression is a non-parametric approach.