Proceedings of Machine Learning Research 1–5

A Mini-Replication of CNM06

John Cooper JLCOOPER@UCSD.EDU

Abstract

Drawing from the CNM06 paper, I attempt to perform a mini-scale replication of its methods and results, using only three algorithms and five scoring metrics (some of which are not used in the original paper). I compare KNN, SVM, and Logistic Regression.

Keywords: *CNM06* - Caruana Paper, *COV* - Cover Type Data, *Letter OPos* - Letter Data "O" positive classification, *Letter AMPos* - Letter Data "AM" positive classification

1. Introduction

This project presents a small-scale analysis of a select subset of data sets, algorithms, and performance metrics found CNM06. I attempt to compare the binary-classification performances of SVM, KNN, and Logit across four data sets and five scoring metrics.

By undertaking a small-scale analysis (voluntarily) constrained by four data sets and classifiers with the sole intention of comparing performance, bluntly using a single performance metric alone (accuracy) did not seem like an appropriate method to capture the differences between these classifiers. As two of the four data sets analyzed were quite imbalanced (ADULT, LetterOPos), more metrics were needed to compensate for this, including Precision, Recall, F1, and ROC-AUC.

To address certain classifiers, such as SVM, which are not meant to predict probabilities (Caruana, 2006), I did **not** calibrate with Platt Scaling or Isotonic Regression as CNM06 does. Instead, the decision outputs were mapped to probabilities through an **untrained** logit function. The pitfalls of this post-scaling are not addressed, and so roc-performance comparisons between SVM and the other classifiers should be taken hesitantly.

The results of this analysis were surprisingly consistent with the CNM06 paper. Table 2 (Mini-Replication) shows the performances over metrics for each algorithm, where SVM scores slightly higher than KNN and significantly higher than LOGIT under all metrics. This is consistent with the Table 2 in the CNM06 paper (6-6). The difference in per-metric-scoring means between the SVM and the other classifiers were also found to be statistically significant.

By problem, SVM outscored KNN on all data sets and LOGIT on all but one. Interestingly, the scores of KNN on LetterOPos and LetterAMPos were statistically indistinguishable from greater SVM scores on those data sets. This is relatively consistent with CNM06, wherein Table 3 (CNM06) KNN actually outperforms SVM on few metrics, although not by much (6-6).

2. Methods

For the algorithms selected, their hyperparameter selection follows those presented in CNM06 almost verbatim, while the performance metrics slightly differ.

2.1. Classifiers

SVMs: Kernels used were radial basis function, linear, and polynomial kernels. For the rbf, gammas used were $\{.001, .005, .01, .05, .1, .5, 1, 2\}$, and for the polynomial kernel, degrees $\{2, 3\}$ were used. For each of these, their regularization parameter varied from 10^{-7} to 10^{4}.

KNN: Hyperparameters varied were the number of neighbors and the types of weights. Number of neighbors took equally spaced (20-step) values between 1 and 500 (exclusive). Weights searched were $\{uniform, distance\}$.

Logit: This classifier was initialized with solver set to Limited-memory BFGS and max iterations set to 1000, the latter to encourage convergence. The penalization term took values in $\{l2, none\}$ and the regularization term ranged from 10^{-8} to 10^{5}.

2.2. Metrics

The performance metrics used in this analysis are Accuracy, ROC, Precision, Recall, and F1.

Accuracy is used as the base measure for comparison across algorithms and data sets, although as mentioned in Section 1 this metric fails to capture to the true performance of algorithms on imbalanced data sets. Even a poor classifier will produce high accuracy on a data set with extreme class imbalance, since by simply guessing at the over-represented classification, the proportion of correct guesses should roughly parallel the over-represented class. For this reason, in negative-class imbalanced data sets, we turn to ROC, Precision, Recall, and F1.

ROC and Precision operate irrespective of class distribution, and measure the efficiency with which classifiers order positive cases before negative cases (Caruana, 2006). Precision provides information on how well our classifier is able to correctly predict positive cases from the pool of all existing positive labels. Recall, in the case of negative imbalance, will tend to be small if the classifier is leans towards negative classification, as the trade-off between true positive and false negative will favor the latter. F1 is used to incorporate both these ideas. All four will provide us a measure of how poorly our classifier is actually doing despite high accuracy.

2.3. Comparisons Across Metrics

Unlike CNM06, calibration is not used due to (1) the computational load of five-fold grid searches nested in multiple three-fold cross-validated calibrations (Platt/Iso), and (2) time constraints. Both KNN and Logit are equipped with simple probability mappings, whereas SVM outputs decisions which *can* be mapped to probabilities through a sigmoid function (Platt)(6-1)

$$P(y = 1|x) = \frac{1}{1 + e^{Af(x)+B}}$$

whose parameters $\{A, B\}$ are tuned through a k-fold cross validation on the training set. As mentioned in Section 1, this procedure was not followed. The decisions of the SVM were simply mapped to probabilities through

$$\frac{e^x}{1 + e^x},$$

which is clearly not best practice, especially since we are training highly non-linear SVMs. Aside from SVM performance, performance on other metrics and algorithms are not calibrated. Instead,

their probability prediction functions are used, which internally seem to use a sigmoid for mapping (6-3).

2.4. Data

The four data sets used were LETTER1(OPos), LETTER2 (AMPos), COVTYPE, and ADULT (6-7). For each data set, all categorical features were converted using one-hot-encoding.

For the ADULT data set, the target variable (income) was converted to binary classification through assigning the mapping $(> 50k) \rightarrow 1$ and $(<= 50k) \rightarrow 0$.

For COV, the target variable ("cov") was converted to binary by assigning the largest class to 1 and the rest to 0.

For LETTER, the target variable ("lettr") was split into two cases as in CNM06. First, "O" was mapped to 1 and the rest of the letters were mapped to 0. Next, "A-M" was mapped to 1, and the rest of the letters were mapped to 0. Classification was done on both targets.

3. Training and Testing Framework

First a data set is chosen and randomly sampled from three times, each time breaking the data into a training set (5000) and a testing set (# samples - 5000).

A classifier is then selected.For each of train/test splits, a five-fold-cv gridsearch with this classifier is run to obtain the best parameters. The classifier with the best parameters is then retrained on the entire train set and used to make predictions on the test set, yielding a test score.

This train/test process is repeated on the two remaining train-test pairs, yielding two more test scores.

From these we obtain three test scores for a single classifier over a single data set. This entire process is repeated for every classifier-dataset pair. Thus, under a performance single metric, one classifier will have a total of **12** test scores to report as well as **12** validation scores.

Table 1 lists the set-up of each problem: the data set, the number of features, the train size, test size, and train size percentage of total data

Note:
The actual implementation differs in structure from what is described above, but only superficially. The implementation runs a for loop to simulate trials, and within each loop splits the data with a random seed corresponding to the trial number before the classifier is trained. The random seed acts to preserve consistency in splitting (as if the actual implementation followed the above).

4. Performance

Tables are presented which summarize performances.

In Table 2, the rows list the three classifiers, while the columns list the metrics. For each classifier-data combo, three scoring performances are produced. This produces nine scoring performances (each performance encapsulating the five metrics used) in total for each classifier. The entries in this table represent the *averaged* scores per classifier, by metric. Boldfaced scores represent a score (under a specific metric and algorithm) whose population mean has been determined to be larger than the others under a 95% confidence process.

In Table 3, the rows list the three classifiers used, while the columns list the data sets. The entries represent the averages of each classifier across all five metrics on a

single data set. Since there are three trials, each metric corresponds to three performance instances over those three trials. Given that there are five metrics, we sum over the five metrics entry-wise (by trial) adn divide by the number of metrics to obtain an average over the five metrics per trial. After this we sum these averages and divide by the number of trials.

Note:
Within the testing/training framework, the grid search selects the optimal classifier based on the classifier's internal scoring method - for SVM, Logit, and KNN, this scoring method is ACC. So when we tune our parameters, we are optimizing based on this metric. Thus when we compare classifiers with metrics other than ACC, we are actually comparing how well classifiers, which are optimized on accuracy, perform under criterion not inherent to the optimization process. This is why there are no validation errors for metrics other than mean accuracy in Table 4.

5. Discussion-Conclusion

5.1. Test Performance Summary

The best classifier on the LetterOPos and LetterAMPos data sets is SVM, whereas on the COV and ADULT data sets, SVM and LOGIT show the best performances, respectively. LOGIT surprisingly outperforms SVM and KNN on the ADULT data set. In the case of LOGIT on the ADULT data set, there is no other score that rivals it with statistical significance, whereas in the case of LetterOPos and LetterAMPos, SVM and KNN perform at statistically indistinguishable levels (Table 3, Table 6).

Over scoring metrics, SVM uniformly outperforms the other classifiers, with LOGIT performing uniformly worse across all metrics against the other two classifiers. While this seems to show that SVM performs better on all metrics, by problem, we see that some uniformly-best classifier does not exist (Table 2, Table 5).

5.2. Validation Performance

One concerning aspect of this analysis was that validation performance and test performance seemed to run in parallel - sometimes they were equal, other times test performance actually exceeded validation performance (Table 2, Table 4). This is referring specifically to accuracy, as those are the only validation metrics which were saved. In the case of testing scoring on non-accuracy metrics outperforming validation scores computed internally with an accuracy, this shouldn't be too much of a concern. Since the classifiers are optimized on accuracy, it makes that their testing performance under different metrics would not naturally produce a score bounded by accuracy validation scores.

5.3. Non-accuracy metrics

It was necessary to include other metrics to gauge classifier performance, especially on the imbalanced data sets COV and ADULTOPos. Table 3 shows the difference between performance metrics, with ROC and ACC showing the best performances respectively, followed by REC, PREC, and F1, not exactly in that order. The introduction of these metrics was the only way for us to see how poorly LOGIT actually performs relative to the other classifiers on most data sets. Furthermore, accuracy would have generally inflated our view of all classifiers performances, especially on imbalanced data (Table 2), in some cases widening

and other cases shrinking the performance margins.

5.4. Calibration

As mentioned, calibration was not used in this analysis. Without a uniform mapping of prediction (or decision functions) to probabilities, especially between SVM and the other classifiers, the ROC score comparison between SVM and the other classifiers is most likely flawed.

6. Citations and Bibliography

1. "Platt Scaling" Wikipedia, Wikimedia Foundation, 6 Dec. 2020, en.wikipedia.org/wiki/Platt_scaling
2. Scikit-learn.org. 2020. Sklearn.Neighbors.Kneighborsclassifier — Scikit-Learn 0.23.2 Documentation. [online] Available at: https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html [Accessed 15 December 2020].
3. GitHub. 2020. Scikit-Learn/Scikit-Learn. [online] Available at: https://github.com/scikit-learn/scikit-learn/blob/95d4f0841/sklearn/neural_network/_multilayer_perceptron.py [Accessed 15 December 2020].
4. Scikit-learn.org. 2020. Sklearn.Svm.SVC — Scikit-Learn 0.23.2 Documentation. [online] Available at: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html [Accessed 15 December 2020].
5. Scikit-learn.org. 2020. Sklearn.LinearModel.Logisticregression— Scikit-Learn 0.23.2 Documentation. [online] Available at: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html[Accessed 15 December 2020].
6. Caruana, R. and Niculescu-Mizil, A., 2020. An Empirical Comparison Of Supervised Learning Algorithms. [online] Cs.cornell.edu. Available at: https://www.cs.cornell.edu/~caruana/ctp/ct.papers/caruana.icml06.pdf
7. Archive.ics.uci.edu. 2020. UCI Machine Learning Repository. [online] Available at: https://archive.ics.uci.edu/ml/index.php [Accessed 15 December 2020].

Tables

Table 1: Problem descriptions

	#ATTR	TRAIN SIZE	TEST SIZE	%POZ
ADULT	106	5000	40222	24%
COV	54	5000	576012	49%
LETTEROPos	16	5000	15000	4%
LETTERAMPos	16	5000	15000	50%

Table 2: Performance over metrics (over four problems)

	ACC	PREC	REC	ROC	F1	MEAN
KNN	.885	.826	.777	.892	.795	.835
SVM	**.899**	**.850**	**.804**	**.936**	**.824**	**.863**
LOGIT	.82	.546	.527	.842	.534	.706

Table 3: Performance by problem (averaged over four metrics)

	COV	ADULT	LetterOPos	LetterAMPos	MEAN
KNN	.777	.696	.914*	.749*	.784
SVM	**.812**	.732	**.937**	**.753**	.807
LOGIT	.769	**.743**	.359	.437	.577

Table 4: Validation performance over single metric (over four problems)

	MEAN ACC
KNN	.882
SVM	.897
LOGIT	.823

Tables Cont.

Table 5: P-Vals for Table 2 (SVM compared with Logit/KNN)

	Pval-Logit	Pval-KNN
SVM-ACC	.021	.0004
SVM-PREC	.019	.0025
SVM-REC	.025	.030
SVM-ROC	.002	.0002
SVM-F1	.023	.0005

Table 6: P-Vals for Table 3(SVM versus rest over data sets)

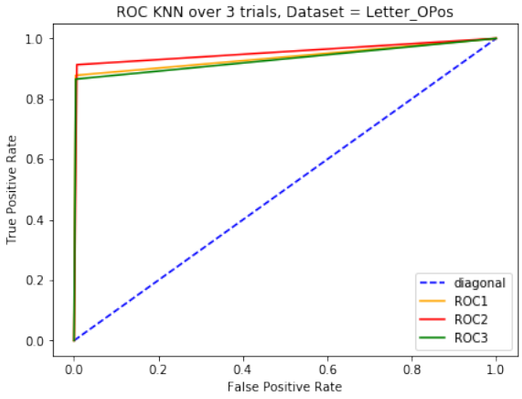
	SVM
COV (KNN)	.045
COV(LOGIT)	.008
ADULT(KNN)	.006
ADULT(LOGIT)	.023
LETTER-O(KNN)	.08
LETTER-O(LOGIT)	5.36e(-5)
LETTER-AM (KNN)	.310
LETTER-AM (LOGIT)	.0001

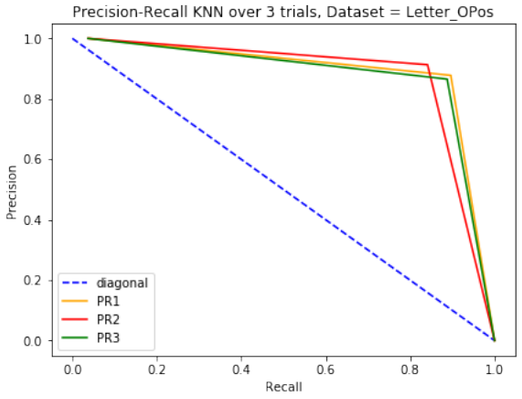
Table 7: Validation scores over data sets (ordered tuple represents score by trial)

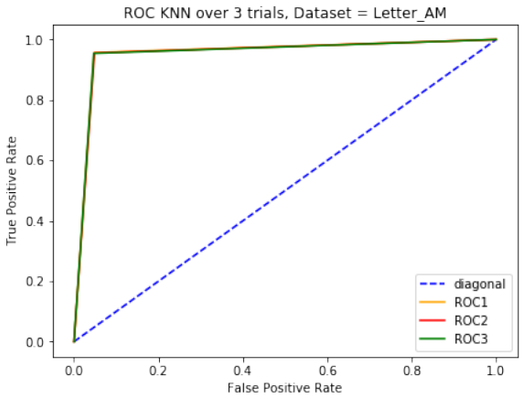
	KNN	SVM	LOGIT
COV	(0.764,0.771 ,0.762)	(0.799,0.793,0.786)	(0.759,0.762,0.753)
ADULT	(0.827,0.835,0.828)	(0.845,0.846,0.844)	(0.841,0.847,0.842)
LETTER-O	(0.988,0.989,0.991)	(0.993,0.992,0.992)	(0.966, 0.962 ,0.961)
LETTER-AM	(0.947,0.943,0.944)	(0.960,0.957,0.954)	(0.736,0.721,0.731)

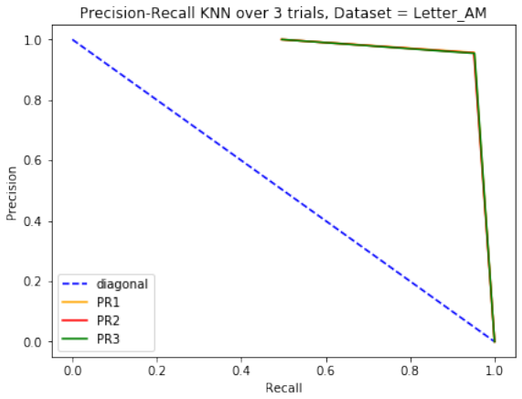
Figures

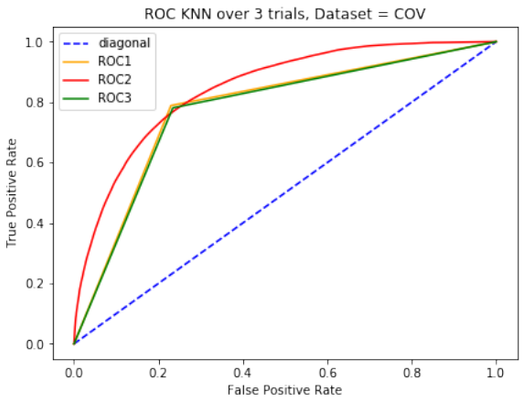
(KNN)

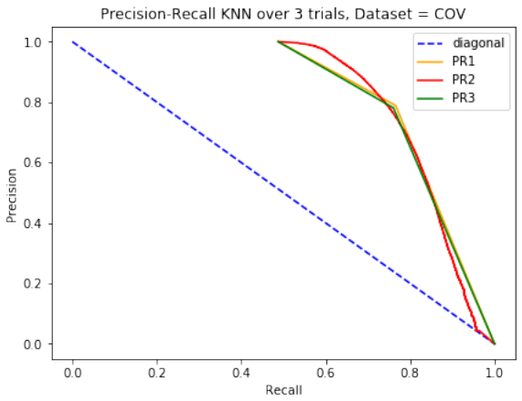

ROC KNN over 3 trials, Dataset = Letter_OPos
True Positive Rate
False Positive Rate
diagonal
ROC1
ROC2
ROC3

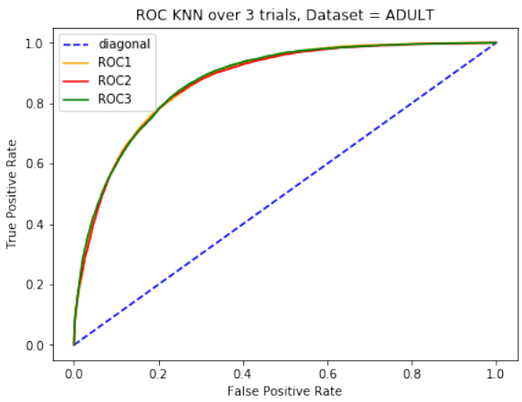

Precision-Recall KNN over 3 trials, Dataset = Letter_OPos
Precision
Recall
diagonal
PR1
PR2
PR3

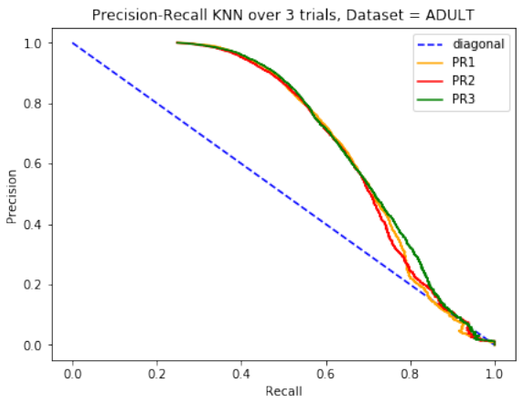

ROC KNN over 3 trials, Dataset = Letter_AM
True Positive Rate
False Positive Rate
diagonal
ROC1
ROC2
ROC3


Precision-Recall KNN over 3 trials, Dataset = Letter_AM
Precision
Recall
diagonal
PR1
PR2
PR3


ROC KNN over 3 trials, Dataset = COV
True Positive Rate
False Positive Rate
diagonal
ROC1
ROC2
ROC3

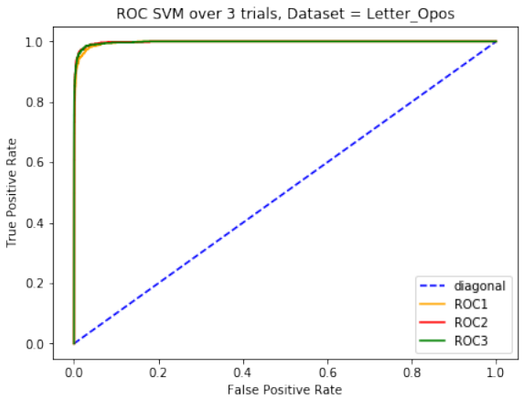

Precision-Recall KNN over 3 trials, Dataset = COV
Precision
Recall
diagonal
PR1
PR2
PR3

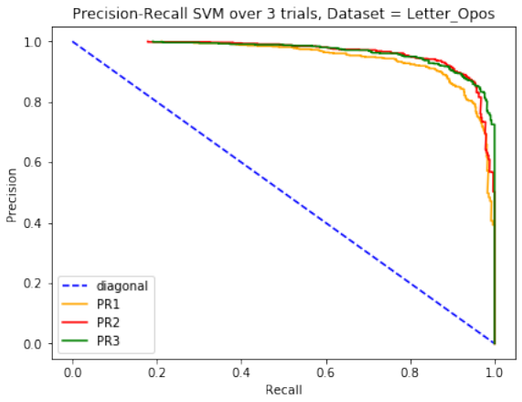

ROC KNN over 3 trials, Dataset = ADULT
True Positive Rate
False Positive Rate
diagonal
ROC1
ROC2
ROC3

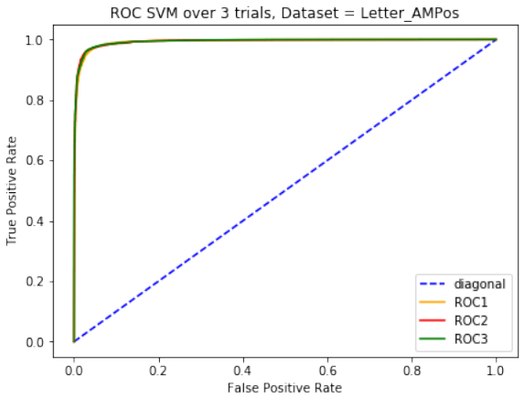

Precision-Recall KNN over 3 trials, Dataset = ADULT
Precision
Recall
diagonal
PR1
PR2
PR3

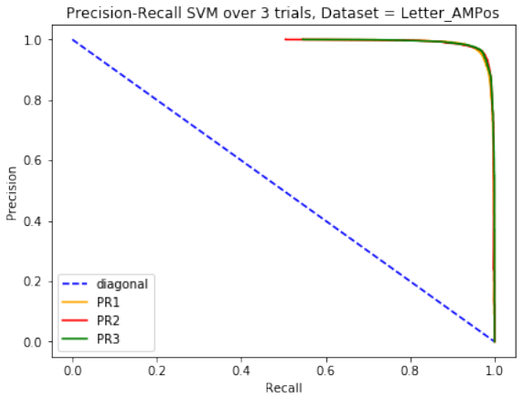
Figures Cont.

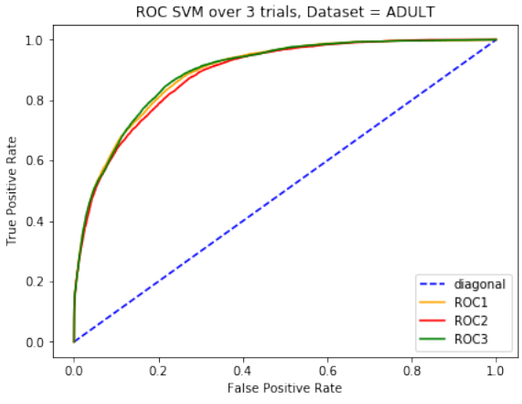
(SVM)

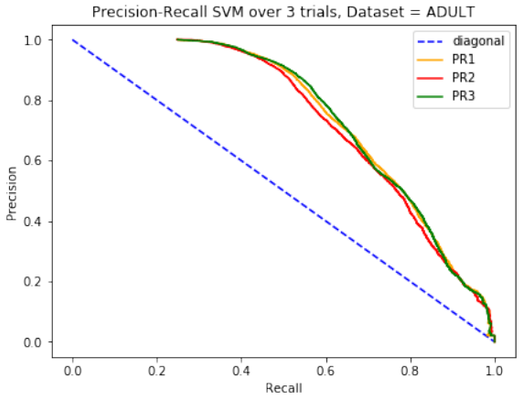


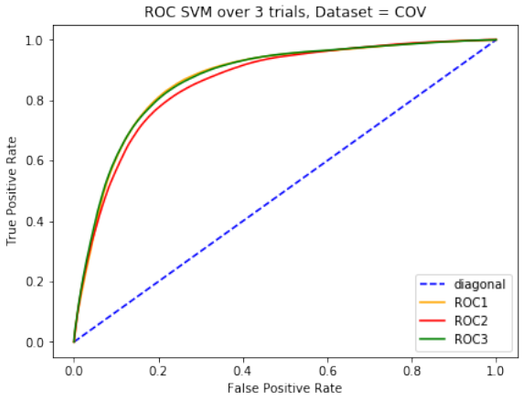


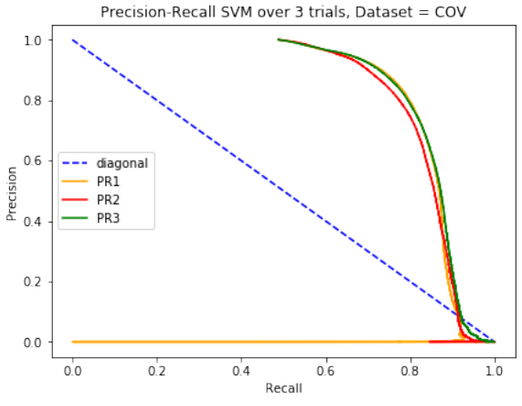












Figures Cont.

(LOGIT)