Perturbation Validation: A New Heuristic to Validate **Machine Learning Models**

Jie M. Zhang ¹ Mark Harman ²¹ Benjamin Guedj ³¹ Earl T. Barr ¹ John Shawe-Taylor ¹

Abstract

This paper introduces Perturbation Validation (PV), a new heuristic to validate machine learning models. PV does not rely on test data. Instead, it perturbs training data labels, re-trains the model against the perturbed data, then uses the consequent training accuracy decrease rate to assess model fit. PV also differs from traditional statistical approaches, which make judgements without considering label distribution. We evaluate PV on 10 real-world datasets and 6 synthetic datasets. Our results demonstrate that PV is more discriminating about model fit than existing validation approaches and it accords well with widely-held intuitions concerning the properties of a good model fit measurement. We also show that PV complements existing validation approaches, allowing us to give explanations for some of the issues present in the recently-debated "apparent paradox" that high capacity (potentially "overfitted") models may, nevertheless, exhibit good generalisation ability.

1. Introduction

Model validation evaluates model performance as well as the match between data and learner. Insufficient model validation may threaten the effectiveness and robustness of machine learning applications.

In applied machine learning, out-of-sample validation is widely used to empirically validate models and reduce the threat of either overfitting or underfitting. It uses test data different from the training data to approximate unseen data. However, out-of-sample validation has the following limitations: 1) the validation or test samples may be insufficiently representative of the underlying unknown data distribution; 2) the samples are typically randomly selected from the collected data, and may therefore have similar bias as in

spondence to: Jie M. Zhang <jie.zhang@ucl.ac.uk>.

the training data, leading to an inflated validation score; 3) model optimisation, based on a fixed set of samples, may lead to overfitting to these samples (Werpachowski et al., 2019).

Previous work has shown that cross-validation may lead to the selection of overly complex models (Piironen & Vehtari, 2017; Gronau & Wagenmakers, 2019). The reliability of cross-validation depends largely on data quality and data sampling (Keevers, 2019).

In statistical machine learning, Vapnik-Chervonenkis (VC) dimension (Shawe-Taylor et al., 1998) and Rademacher complexity (Mohri & Rostamizadeh, 2009) have been used to measure the complexity of a model's hypothesis space. Both are theoretical tools that define generalisation error upper bounds. Although these bounds may help model validation, they can be loose and difficult to compute (Rosenberg & Bartlett, 2007).

Recently, experimental results from a number of papers have challenged traditional model complexity measurements. Zhang et al. (Zhang et al., 2016) found that deep hypothesis spaces can be large enough to memorise random labels. They discussed the limitations of Rademacher complexity in explaining the generalisation ability of large neural networks, and called for new measurements. Arpit et al. (Arpit et al., 2017) found that one reason for these limitations is that deep learning models that are large enough to memorise tend to learn patterns before they start to memorise. Their results reveal the importance of judging the circumstances under which models memorise. Frankle and Carbin (Frankle & Carbin, 2018) found that dense and randomly-initialised neural networks contain subnetworks ("winning tickets") with comparable test accuracy to the original network, indicating that large networks might explicitly contain simpler representations.

This paper presents Perturbation Validation (PV), a new heuristic for evaluating the degree of model fit that relies only upon training data. PV perturbs training data labels gradually and retrains the model using the perturbed labels, then measures the training accuracy decrease rate. The key intuition is that a model, if having truly learned the patterns in data, would be less likely to be 'fooled' by a small ratio of

¹University College London ²Facebook London ³Inria. Corre-

incorrect labels. Consequently, the accuracy decrease rate is expected to be high for such a good model. By contrast, an over-complex learner may have extra capacity to fit incorrect labels, and thus may retain good training accuracy despite the injected noise. Furthermore, an over-simple learner has poor learnability, and will have low training accuracy with or without the presence of incorrect labels. Thus, both over-complex and over-simple models tend to be insensitive to incorrect labels and exhibit small accuracy decrease rates, when PV injects label noise.

Unlike out-of-sample validation, PV does not split the data. It relates the changes in the data to changes in the training accuracy. Unlike VC-dimension and Rademacher complexity, PV does not assess the complexity of the hypothesis space. Rather, PV measures the degree to which the hypothesis space matches the available data. Arpit et al. (Arpit et al., 2017) found that training data plays an important role in determining the degree of memorisation, thus data-dependent measurements are required. VC-dimension is data-independent; Rademacher complexity uses random labels, and is thereby label-independent. By contrast, PV depends on both the instance and the label distributions, so it better captures training data distribution.

We evaluate PV on 10 open datasets (see Table 1), and 6 synthetic datasets with different known data distributions, using widely-adopted classifiers. We investigate whether PV is a complementary measure to out-of-sample validation for model selection and parameter tuning. We also investigate the influence of training data size on test accuracy and PV.

The results lead to the following primary observations. First, PV captures well the degree of match between decision boundaries and data patterns in model selection, and is more discriminating between models than cross validation and test accuracy. Second, PV is responsive to changes in capacity; its behaviour with respect to increasing capacity well matches natural intuitions concerning the expected degree of approximating fit between model and data. Third, when the learner is over-complex for the data based on PV, enriching training data can improve PV, even after test accuracy has plateaued.

The paper also discusses the 'apparent paradox' concerning hypothesis space complexity and model generalisation ability. We discussed the possible reasons and how PV helps to better understand this paradox.

Based on our findings, we propose PV as a complement to traditional out-of-sample validation. A good model fit is expected to lead to both high test accuracy and PV value. When test accuracy is high but PV value is low, it indicates that, for the current training data, there is a high degree of unnecessary capacity in the learner. The unnecessary capacity may increase cost, affect the learner efficiency, and

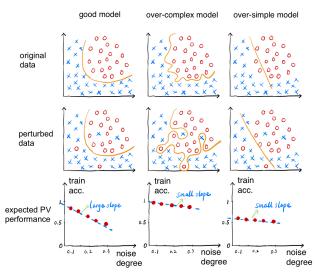


Figure 1. The intuition that underpins PV. Over-complex and over-simple models each tend to have a low accuracy decrease rate.

make the learner vulnerable to data attacks.

In summary, we make the following primary contributions:

- Heuristic: We introduce PV to validate the fit between
 the available data and a learner's hypothesis space.
 PV does not require test sets, but uses the accuracy
 decrease rate against per unit of label noise introduced.
 It can be adopted as a complement to existing model
 validation methods.
- **Approach**: PV is designed as an empirical practitioner toolbox. It is easy to use, and applicable for practitioners in machine learning.
- Empirical Evidence: We demonstrate PV's effectiveness in classification tasks on 16 datasets.

2. Perturbation Validation

Figure 1 illustrates the intuition that underpins PV. For a 'good' model that learns the real pattern in the data (first column of Figure 1), the learner is less likely to be 'fooled' by a small number of incorrect labels with the perturbed data. As a result, as the label noise degree increases, the training accuracy decreases.

An over-complex learner tends to fit noise in the training data (second column), thereby retaining good or even similar training accuracy on the perturbed datasets. As a result, the accuracy decrease rate may be small. In extreme cases, the training accuracy does not decrease despite the presence of incorrect label noise.

An over-simple learner has poor learnability. It has low training accuracy with or without incorrect labels, and the accuracy decrease rate is also small, as depicted by the sub-figures in the third column of Figure 1.

We also conjecture that, as depicted in the upper middle sub-figure of Figure 1, when there is sufficient noise-free training data, an over-complex model may still have high test accuracy. However, such a model would, nevertheless, have unnecessary capacity, increasing time and space costs and making the learner vulnerable to data attacks. We investigate this in more detail in Section 4.

2.1. Formalisation

Let S be a training sample. Let r be a label noise degree (ratio of perturbed labels for each class), $r_1, r_2, ..., r_m$ is a noise degree sequence. S_{r_i} is a perturbed training sample constructed by changing r_i proportion of the labels for each class in S. \mathcal{H} is the hypothesis set of the learner. Let $\widehat{\mathrm{Acc}}_S(h)$ be the empirical training accuracy of $h \in \mathcal{H}$ based on S. $\widehat{\mathrm{Acc}}(S) = \mathrm{argmax}_{h \in \mathcal{H}} \widehat{\mathrm{Acc}}_S(h)$ is the maximum empirical training accuracy of $h \in \mathcal{H}$ over training set S.

Definition 1 (PV) PV is the absolute value of linear regression coefficient when modeling the relationship between label noise degree r_i and training accuracy $\widehat{Acc}(S_{r_i})$:

$$PV = \left| \frac{\sum_{i=0}^{m} (r_i - \overline{r}) (\widehat{Acc}(S_{r_i}) - \widehat{\overline{Acc}(S_r)})}{\sum_{i=0}^{m} (r_i - \overline{r})^2} \right|. \tag{1}$$

PV uses a linear regression coefficient, because, as Figure 1 shows, we conjecture that the relationship between perturbed training accuracy and label noise degree tends to be linear. We present results that empirically confirm this conjecture in our appendix.

When perturbing S, we have two choices: we can either perturb globally disregarding the class, or with equal noise degree for each class. In this paper, we choose to perturb by each class to better preserve label distribution.

When $n=r_i*|S|$ labels are perturbed, the ideal case is that the learner is not 'fooled' by these incorrect labels, and retains all predictions for the original labels. For such a good model fit, its accuracy decrease with n perturbed labels would be $\frac{n}{|S|} = \frac{r_i*|S|}{|S|} = r_i$. That is, both noise degree and the accuracy decrease are r_i and the accuracy decrease rate is 1. There may be cases that a model's PV is larger than 1, indicating that the injected noise affects the predictions of unperturbed labels. To handle these cases, in this paper, we use 1 - |PV - 1| to 'punish' the extra accuracy decrease, to reflect PV's score around 1, making, for example, 1.2 equivalent to 0.8, so that the decrease rate is better up to 1, and then worse after it. According to our results, only two classifiers have PV larger than 1 when trained with 100 synthetic data points.

Dataset	abbr.	#training	#test	#class	#feature
synthetic-moon	moon	100 – 1e6	2,000	2	2
synthetic-moon (0.2 noise)	moon-0.2	100 - 1e6	2,000	2	2
synthetic-circle	circle	100 - 1e6	2,000	2	2
synthetic-circle (0.2 noise)	circle-0.2	100 - 1e6	2,000	2	2
synthetic-linear	linear	100 - 1e6	2,000	2	2
synthetic-linear (0.2 noise)	linear-0.2	100 - 1e6	2,000	2	2
Iris	iris	150	_	3	4
Wine	wine	178	-	3	13
Breast Cancer Wisconsin	cancer	569	-	2	9
Car Evaluation	car	1,728	_	4	6
Heart Disease	heart	303	_	5	14
Bank Marketing	bank	45,211	_	2	17

48,842

67,557

60,000

50,000

16,281

10,000

2

2

10

14

42

Table 1. Details of datasets

3. Experiments

Adult

Connect-4

CIFAR-10

MNIST

3.1. Research Question

We will answer the following questions to evaluate PV.

RQ1: What is the performance of PV in model selection?

RQ2: How do hyperparameters influence PV? RO3: How does training data size influence PV?

adult

connect

mnist

PV aims to access the fit between training data distribution and a learner's hypothesis space. To check whether PV accords with intuitions on the properties of a good model fit measurement, the first research question investigates whether PV suggests models whose decision boundaries match data patterns. The second research question fixes the training data, but adjusts a learner's hypothesis space by changing hyperparameters, and explores the response of PV scores. The third research question controls the hypothesis space, but adjusts training data size, then explores the response of PV assessment results.

3.2. Datasets

Table 1 shows the details of each dataset used to evaluate PV. To obtain datasets with known ground-truth decision boundaries, we first use synthetic datasets with three types of data distributions: moon, circle, and linearly-separable. To study the influence of noisy training data on PV, for each type of distribution, we create datasets with noise. We also generate different-size training data ranging from 100 to 1 million data points to study the influence of training data size. These synthetic datasets help check whether PV identifies the right model whose decision boundary matches the data distribution, with and without noise in the original dataset S. We do not expect such synthetic datasets to reflect real-world data, but the degree of control and interpretability they offer allows us to verify the behaviour of PV with a known ground-truth for model choice.

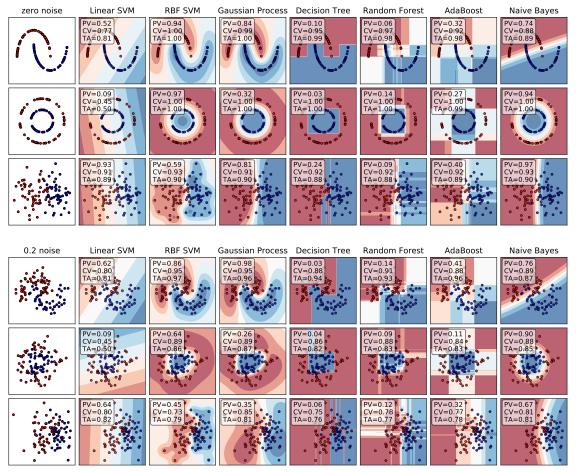


Figure 2. Performance of PV in model selection. PV, CV, TA denote three-fold PV, three-fold CV accuracy, and hold-out Test Accuracy. Red and blue points are the original training data without noise (top-three rows) and with 0.2 noise (bottom-three rows). Areas with different colours show the decision boundaries from each model. We observe that PV captures well the match between decision boundaries and data patterns, while the other metrics are less discriminating among multiple models, especially for zero-noise training data.

We also report results on a further 10 real-world widely-adopted datasets with different sizes and numbers of features. Eight of them are from the UCI machine learning repository, the remaining two are MNIST and CIFAR-10.

For each dataset, we use three-fold PV: we experiment with label noise degrees r of PV from 0.1 to 0.3, in steps of 0.1, yielding 3 perturbed training datasets. We randomly choose r labels from each label class to perturb.

3.3. RQ1: Effectiveness of PV in Model Selection

For the synthetic datasets, we use *scikit-learn* (Pedregosa et al., 2011) synthetic dataset distributions: 1) Moon: the data points are distributed with two interleaving half circles. 2) Circle: the data points are distributed in the form of a large circle containing a smaller circle in two dimensions. 3) Linearly-separable: the data points are linearly separable. This experiment uses the same settings as in the *Scikit-learn* tutorial (Classifier Comparison, 2019). Each dataset has

100 training data points for model selection. We generate another 2,000 points as hold-out test sets.

Figure 2 shows the training data points, decision boundaries of each classifier, and measurement values from PV, CV, and test accuracy on the 2,000 hold-out test set. When manually comparing data patterns and decision boundaries, we observe that PV tends to give high values for cases where the decision boundaries match well the data patterns. The *Scikit-learn* documentation mentions that Naive Bayes and Linear SVM are more suitable for linearly-separable data; accordingly, PV provides largest values for these two classifiers. We also observe that Gaussian Process has circle decision boundaries on zero-noise data, yet PV is low. This is because its decision boundaries are easily affected by noise, as shown by the noisy-data sub-graphs.

The figure also shows cases where cross validation and test accuracy have limitations in evaluating data-learner fit. For example, for the circle distribution, Decision Trees

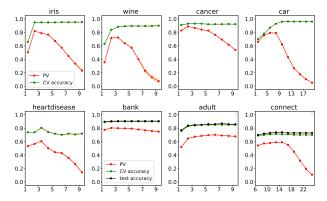


Figure 3. Changes in PV when increasing the maximum depth for Decision Trees. The x-axis ticks for car and connect differ to capture PV's inflection point. We can observe that, while CV and test accuracy agree with PV on the key influence point in most cases, they are less sensitive to depth than PV. This is evidence that PV may further help find the most suitable parameters.

and Random Forests (for which maximum depth is 10 for both classifiers) give obviously ill-fitted rectangle-shaped decision boundaries, yet the cross validation accuracy and hold-out test accuracy remain high.

Answer to **RQ1**: PV captures well intuitions about the match between model decision boundaries and data patterns on synthetic datasets. PV is also more discriminating than CV and test accuracy.

3.4. RQ2: Influence of Hyperparameters

For models whose capacity increases along with their hyperparameters, we expect their goodness of model fit to increase, then peak, before decreasing. We then assess whether PV matches this pattern.

We study capacity-related hyperparameters for three well-known algorithms: the maximum depth for a Decision Tree¹, the dropout rate for a Convolutional Neural Network (CNN), and C and gamma for a Support Vector Machine (SVM).

Maximum Depth for Decision Tree. Figure 3 shows how PV responds to increases in maximum depth of Decision Tree for eight real-world UCI datasets. For small datasets (smaller than 2,000), we do not split out hold-out test data, but use the whole data as training data. For the bank and connect datasets, we use 50% of the data as hold-out test set. For adult, we use its original test set. We repeat the experiments 10 times. The yellow shadow in the figure indicates the variance across runs.

From Figure 3, we make the following observations. First,

PV score first increases then decreases as maximum depth increases, and exhibits a single maximum in each curve. This is consistent with the pattern we expect a good measure to exhibit. We also observe that, for small datasets, large depths yield low PV scores, whereas large datasets do not. We explore the influence of training data size on PV further in Section 3.5.

We also observe that CV and test accuracy show similar validation scores for multiple parameters, yet PV is more responsive to changes to depth. This observation indicates that PV may provide further information to help tune maximum depth when CV and test accuracy are less able to distinguish multiple parameters.

In particular, if a developer uses grid search to select the best maximum-depth ranged between 5 and 10, we find that grid search suggests depths of 8, 6, 9 in three runs for the cancer dataset, which are over-complex and unstable. Similar results are observed for other small datasets. With PV, its decrease trend in this range indicates that there is a simpler model with comparable predictive accuracy but better robustness to label noise.

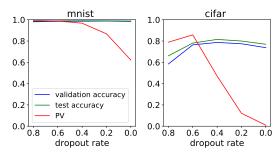


Figure 4. Influence of CNN's dropout rate on PV and validation/test accuracy. We see similar results to those for Decision Trees in Figure 3.

Dropout Rate for CNN. To check the influence of dropout rate for deep learning models, we use four dropout rates (0.2, 0.4, 0.6, 0.8) to classify the mnist and cifar datasets. We use CNN models coming from Keras documentation² for the two datasets. Validation accuracy is calculated with 60% training data and 40% validation data.

Figure 4 shows the results. For mnist, the PV scores for dropout rates 0.8, 0.6, 0.4 are high, approaching 1.0. For cifar, the scores are lower, and decrease dramatically when the dropout rate is 0.4. When the dropout rate is 0.6, PV has the largest value, where validation and test accuracy also witness a turning point. One can still observe that test accuracy increases very slowly after 0.6. We discuss the benefits and cost for the minor increase in Section 4 and Figure 8.

When the dropout rate is zero, test accuracy remains high,

¹Random Forests show similar results to Decision Trees, so we present only Decision Tree results for brevity.

²https://keras.io/

but PV gives low scores in our experiments, indicating that the model is easily biased by incorrect labels. This signals that the model contains unnecessary capacity, which is consistent with the recent 'winning lottery ticket' hypothesis (Frankle & Carbin, 2018) of neural networks: there may exist a subnetwork with comparable test accuracy. Previous work (Yu et al., 2018; Han et al., 2015; Hu et al., 2016) also found that deep neural networks can contain significant redundancy. The extra capacity may not seriously affect test accuracy under the current training data, yet it may make deep networks computationally expensive and memory intensive (Hu et al., 2016; Frankle & Carbin, 2018). Such over-capacity also elevates the risk of a successful data attack on the learners.

C and gamma for SVM. In SVM, the gamma parameter defines how far the influence of a single training example reaches; the C parameter decides the size of the decision boundary margin, behaving as a regularisation parameter. In a heat map of PV scores as a function of C and gamma, the expectation is that good models should be found close to the diagonal of C and gamma (Scikit-learn:SVM, 2020).

Figure 5 presents the heat map for cross validation and PV for two datasets. We do not use a hold out test set to ensure sufficient training data. The upper left triangle in each subfigure denotes small complexity; the bottom right triangle in each sub-figure denotes large complexity. In both cases, PV gives low scores.

When comparing CV and PV, PV is more responsive to hyperparameter value changes. With PV scores, it is more obvious that good models can be found along the diagonal of C and gamma. When C and gamma are both large, the CV score is high but PV score is low. As we observed for Decision Trees and CNNs, this is an indication that there exists a simpler model with similar test accuracy. In practice, as stated by Scikit-learn documentation, it is interesting to simplify the decision function with a lower value of C so as to favour models that use less memory and that are faster to predict (Scikit-learn:SVM, 2020).

Answer to **RQ2**: PV is responsive to hyperparameter value changes. When test accuracy and CV accuracy are less discriminating among different hyperparameters, PV can help to identify the existence of unnecessary capacity.

3.5. RQ3: Influence of Training Data

PV seeks to measure the degree of fit between available training data and a learner. In this section, we explore the influence of training data on PV. We expect that when a learner is over-complex for the data, adding extra training data will improve the learner's robustness to incorrect la-

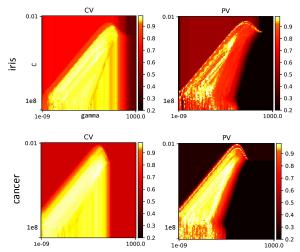


Figure 5. Influence of SVM parameters on CV and PV. The horizontal/vertical axis is gamma/C. Good models are expected be found close to the diagonal. As can be seen, CV has a broad high-valued (bright) region, while PV's high-valued region is narrower, showing that PV is more responsive to parameter changes.

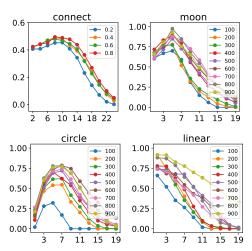


Figure 6. Influence of training data size on PV (vertical axis) when increasing maximum depths (horizontal axis) of Decision Trees. For a given depth, larger datasets tend to yield larger PV scores, indicating that data size plays an role in determining the trained model's robustness to incorrect training labels.

bels, thus the PV score ought also to tend to increase when training data size increases.

Indeed, previously from Figure 3, we have observed that models trained on large datasets (e.g., bank, adult, and connect) tend to be more robust to incorrect labels for large depths. In this section, we investigate this observation in more depth by gradually increasing the training data size to explore its influence on PV.

For real datasets, we experiment by randomly selecting different sizes of subsets from large datasets. From Figure 3, among the three large datasets, the PV values of bank and adult datasets are already stable, and thus we report only on the connect dataset. For synthetic datasets, we generate different sizes of data sets. Figure 6 shows the results.

We then investigate what would happen to PV should we deliberately use more training data than normally expected. That is, we go beyond the assumption that there is no need to increase data when the test accuracy becomes stable. As we can see from Figure 7, PV is more responsive to changes in training data size than test accuracy. For learners that are over-complex to the data, when adding more training data no longer increases test accuracy, PV continues to increase, indicating that the model's robustness to incorrect labels continues to increase.

In Figure 7, when comparing the left sub-figures (clean training data) and the right ones (0.2-noisy training data), we can observe that both test accuracy and PV scores are lower for noisy training data. We suspect that this is because the noisy data is comparatively more complex and difficult to learn, and thus more poorly fits the learner's hypothesis.

These observations indicate another possible value in the use of PV, as a complement to CV and test accuracy. Specifically, where CV or test accuracy is high yet PV is low, the ML engineer has two potential actions that he or she might choose to take to improve PV's assessment of the model fit: either to optimise the learner (e.g., search for smaller-capacity models with comparable test accuracy), or to optimise the data (e.g., increase data size to increase the model's robustness to incorrect labels).

Answer to **RQ3**: Enriching training data can increase PV, even when it no longer increases test accuracy. This indicates that PV is more responsive to changes in training data size, and highlights how PV complements existing model fit assessments.

4. Discussing the 'Apparent Paradox'

There has been recent recognition of an 'apparent paradox' that over-complex models may yet have good generalisation ability (Kawaguchi et al., 2017; Zhang et al., 2016; Arpit et al., 2017). Specifically, the conventional wisdom suggests that over-complex models that are large enough to memorise training data easily overfit the noise in training data, thereby failing to generalise well. However, recent studies show that deep learning models can be highly complex yet still enjoy a good test accuracy (Zhang et al., 2016; Arpit et al., 2017).

In this section, we discuss several possible reasons for this paradox, and how PV helps us to adopt a fresh angle on the explanations for it.

1) Risk may not always be triggered. Structural Risk Min-

imisation (SRM) has been used to provide boundaries of generalisation error. Let R be the generalisation error, E(S) be the training error, H be the capacity of the hypothesis space, and β be a constant. According to SRM, we have $R < E(S) + \beta H$. As the inequality reveals, when the boundary is large, R can be either large or small. This indicates that when a learner has a high degree of unnecessary capacity, the risk of a high R may not always be triggered.

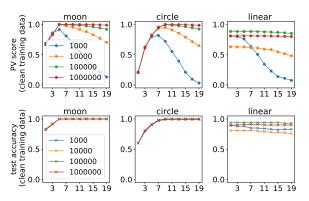
We conjecture that training data, *S*, plays a role in determining whether the risk in unnecessary capacity is triggered. The results in Figure 8 confirm our conjecture. When the training data does not contain injected noise, the test accuracy may keep increasing, albeit slowly, when PV is small (as shown by Figure 4). However, the more extra-capacity a learner has, the more dramatic the adverse effect noise has on its test accuracy.

Based on this observation, we suggest that it is important to be aware of unnecessary capacity because an over-complex model can be 1) easily biased by incorrect labels; 2) vulnerable to training data attack; 3) computationally and memory intensive (Hu et al., 2016; Frankle & Carbin, 2018; Han et al., 2015). Whether these problems matter in practice depends, of course, on the model's intended application domain. Our results indicate that PV can help the engineer to gain greater awareness of potentially unnecessary capacity.

2) Test accuracy may not always sufficiently measure model fit. As shown by Figure 2, test accuracy may sometimes be insufficient in measuring the fit between data patterns and decision boundaries. In addition, in statistical learning theory, there is an assumption that the test samples are drawn i.i.d. according to an underlying distribution (Mohri et al., 2018). However, this assumption might not hold in practice, where test samples can be insufficient or unevenly distributed, especially for image datasets, which can be quite diverse (Barbu et al., 2019).

PV does not rely on test data. It complements test accuracy and alleviates its limitations in assessing data-learner fit. Based on our results, when PV is high so is test accuracy and when test accuracy is low, so too is PV. This indicates that high test accuracy may be a necessary (but not sufficient) condition for a good data-learner fit (measured by PV). This observation highlights that PV is aligned with test accuracy yet also provides a complementary signal to the practitioners.

3) Data-independent capacity assessment criteria may have limitations. Figures 3 and 6 indicate that training data plays an important role in determining the data-learner fit, which is also discussed in previous work (Arpit et al., 2017). In other words, when calculating generalisation error, the fit between the hypothesis space and the training data matters more than the absolute size of the hypothesis space.



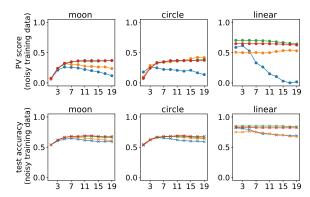


Figure 7. Performance of PV (first row) and test accuracy (second row) on training data of increasing size with different maximum depths (horizontal axis). Observations: 1) PV is more responsive to data size changes; 2) PV no longer decreases for large depths when the training data size is sufficiently large; 3) PV computed for noisy training data is lower than that computed for clean training data.

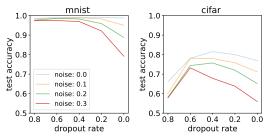


Figure 8. Changes of test accuracy with different training data noise. The more extra-capacity a learner has (according to PV as shown in Figure 4), the more dramatic the test accuracy would be affected by the noise in training data.

Previous complexity measurements typically judge the complexity of learner without considering training data distribution. For example, VC dimension considers neither feature distribution nor label distribution, while Rademacher Complexity does not consider label distribution.

By contrast, PV does consider label distribution. Although PV deliberately injects label noise, the aim is to perturb a relatively small proportion of labels under each class, so that the overall label distribution is preserved.

4) Perfect training accuracy does not always indicate over-fitting. The conventional wisdom suggests that as capacity increases, training error continues decreasing to zero, but test error would first decrease, then increase. There is belief that, when the training error approaches zero, the model overfits (Wyner et al., 2017). This might risk unfairly punishing good models that, indeed, do have almost perfect training accuracy and yet do not overfit.

PV does not use a single training accuracy, but rather the training accuracy decrease rate with several perturbed training data. Figure 9 shows the correlation between PV and training accuracy. Training accuracy approaching 1.0 may still lead to good robustness to incorrect labels as long as training data is appropriate. Indeed, as our experiments for

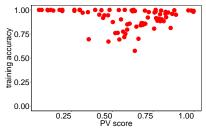


Figure 9. Difference between PV and training accuracy on real-world datasets (*p*-value=0.193). High training accuracy can correspond to large PV.

mnist demonstrate, we observe that a model with very high training accuracy can still have very high test accuracy and PV.

5. Conclusion

We introduced PV, a new approach to assess the fit between a learner's hypothesis space and the available training data. PV validates the fit via checking the learner's robustness to incorrect training labels (expressed as deliberately injected noise). We show that PV complements existing out-of-sample validations and is more responsive to model capacity and training data characteristics. Our results also demonstrate that PV accords well with traditional expectation for a good model fit assessment criterion.

There are many other exciting directions to explore, and we hope the present paper will serve as a starting point to future contributions.

References

- Arpit, D., Jastrzębski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M. S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y., et al. A closer look at memorization in deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 233–242. JMLR. org, 2017.
- Barbu, A., Mayo, D., Alverio, J., Luo, W., Wang, C., Gut-freund, D., Tenenbaum, J., and Katz, B. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in Neural Information Processing Systems*, pp. 9448–9458, 2019.
- Classifier Comparison. Classifier Comparison, 2019. URL https://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html.
- Frankle, J. and Carbin, M. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv* preprint *arXiv*:1803.03635, 2018.
- Gronau, Q. F. and Wagenmakers, E.-J. Limitations of bayesian leave-one-out cross-validation for model selection. *Computational brain & behavior*, 2(1):1–11, 2019.
- Han, S., Mao, H., and Dally, W. J. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *ICLR*, 2015.
- Hu, H., Peng, R., Tai, Y.-W., and Tang, C.-K. Network trimming: A data-driven neuron pruning approach towards efficient deep architectures. arXiv preprint arXiv:1607.03250, 2016.
- Kawaguchi, K., Kaelbling, L. P., and Bengio, Y. Generalization in deep learning. *arXiv preprint arXiv:1710.05468*, 2017.
- Keevers, T. L. Cross-validation is insufficient for model validation. Technical report, 2019.
- Mohri, M. and Rostamizadeh, A. Rademacher complexity bounds for non-iid processes. In *Advances in Neural Information Processing Systems*, pp. 1097–1104, 2009.
- Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of Machine Learning*. MIT press, 2018. Second edition.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V.,
 Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P.,
 Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.
 Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

- Piironen, J. and Vehtari, A. Comparison of bayesian predictive methods for model selection. *Statistics and Computing*, 27(3):711–735, 2017.
- Rosenberg, D. S. and Bartlett, P. L. The rademacher complexity of co-regularized kernel classes. In *Artificial Intelligence and Statistics*, pp. 396–403, 2007.
- Scikit-learn:SVM. RBF SVM parameters. https://scikit-learn.org/stable/auto_examples/svm/plot_rbf_parameters.html, 2020.
- Shawe-Taylor, J., Bartlett, P. L., Williamson, R. C., and Anthony, M. Structural risk minimization over datadependent hierarchies. *IEEE transactions on Information Theory*, 44(5):1926–1940, 1998.
- Werpachowski, R., György, A., and Szepesvári, C. Detecting overfitting via adversarial examples. 2019.
- Wyner, A. J., Olson, M., Bleich, J., and Mease, D. Explaining the success of adaboost and random forests as interpolating classifiers. *The Journal of Machine Learning Research*, 18(1):1558–1590, 2017.
- Yu, R., Li, A., Chen, C.-F., Lai, J.-H., Morariu, V. I., Han, X., Gao, M., Lin, C.-Y., and Davis, L. S. Nisp: Pruning networks using neuron importance score propagation. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.
- Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. *arXiv* preprint arXiv:1611.03530, 2016.