Uncertainty-aware INVASE: Enhanced Breast Cancer Diagnosis Feature Selection

Jia-Xing Zhong,[1] **Hongbo Zhang** [2]
[1] School of Electronic and Computer Engineering, Peking University
[2] Department of Computer and Information Sciences, Virginia Military Institute
jxzhong@pku.edu.cn, hbzhang@vt.edu

Abstract

In this paper, we present an uncertainty-aware INVASE to quantify predictive confidence of healthcare problem. By introducing learnable Gaussian distributions, we leverage their variances to measure the degree of uncertainty. Based on the vanilla INVASE, two additional modules are proposed, i.e., an uncertainty quantification module in the predictor, and a reward shaping module in the selector. We conduct extensive experiments on UCI-WDBC dataset. Notably, our method eliminates almost all predictive bias with only about 20% queries, while the uncertainty-agnostic counterpart requires nearly 100% queries. The open-source implementation with a detailed tutorial is available at https://github.com/jx-zhong-for-academic-purpose/Uncertainty-aware-INVASE/blob/main/tutorial_invase%2B.ipynb.

Introduction

Breast cancer is an increasing health problem (Howell et al. 2014). One in Eight U.S. women will develop invasive breast cancer in her life time. Early diagnosis of Breast cancer is important. Among them, conventional global feature based machine learning method has only achieved limited successes (Wang et al. 2016). High dimension instance-wise feature selection is an emerging machine learning approach, on which the relevant subset of features should be discovered *for each individual data sample*. To address that problem, researchers have proposed several valuable models (Shrikumar, Greenside, and Kundaje 2017; Yoon, Jordon, and van der Schaar 2019; Chen et al. 2018; Lundberg and Lee 2017). Among them, learning to explain (Chen et al. 2018) has built the foundation of instance feature selection and explanation of the features as well using a mutual information model. As one of the state-of-the-art algorithms, INVASE (Yoon, Jordon, and van der Schaar 2019) further extends the learning to explain using a baseline network and a predictor to train a selector in the actor-critic manner, which allows variable-size feature selection.

Existing instance-wise feature selectors are devised to achieve high performance in target tasks. However, they ignore another important goal: *capture the confidence of their outputs*. The lack of accurate confidence interval will lead to deviated estimate. In practice, that may lead to over-confident yet incorrect predictions, which is likely dangerous particularly in the application scenario of healthcare (Tonekaboni et al. 2019) such as false positive and false negative. Since medical services are extremely complex and hardly fault-tolerance, a confidence-agnostic algorithm is undesirable. Thus, an uncertainty-aware approach to mitigate the over confidence problem should be introduced to instance-wise feature selection, for the purpose of *avoiding potential error decisions*.

In this paper, we enhance INVASE to *quantify its predictive confidence by learning an uncertainty estimation*. The vanilla INVASE optimizes the predictor by treating data points as samples from a set of distributions with Dirac delta probability density functions, whereas our model regards data as samples from learnable uncertainty-aware distributions. To be specific, we establish our model with series of Gaussian distributions, of which the corresponding *variances measure the degree of uncertainty*. Our model is completely consistent with the vanilla INVASE in an extreme condition, *i.e.*, the model "thinks" that every prediction is absolutely certain. In our work, two modules are added to the vanilla INVASE, *viz.*, uncertainty quantification and reward shaping. The former estimates the uncertainty of selected features for the predictor, while the latter assists in improving such estimation via the selector.

To demonstrate the efficacy of our presented extension for INVASE, we conduct extensive experiments on a real-world medical dataset *Wisconsin Diagnostic Breast Cancer* (UCI-WDBC) (Dua and Graff 2017). Experimental results show the superiority of our certainty-aware approach: at only about 20% query rates, our model correct almost all predictive bias. To achieve an equal performance gain, certainty-agnostic counterparts require to query about nearly 100% testing data.

In summary, the contribution of this paper is three-fold:

- Based on INVASE, we put forward an uncertainty-aware extension. To the best of our knowledge, our model is the first instance-wise feature selector to quantify predictive uncertainty. That is beneficial to discover potential mistakes of the model output.
- Theoretically, the vanilla version of INVASE can be considered as a particular case of ours. As a seamlessly

backward-compatible extension, our implementation only needs two modules: uncertainty quantification for the predictor and reward shaping for the selector.

- Experimentally, we evaluate our model on the UCI-WDBC dataset from two aspects, *i.e.*, overall performance of the intact model, along with in-depth studies of every component module. For reproducible research, the source codes are provided online.

Related Work

Instance-wise feature selectors have only been studied recently (Shrikumar, Greenside, and Kundaje 2017; Yoon, Jordon, and van der Schaar 2019; Chen et al. 2018; Lundberg and Lee 2017), unlike the well-developed global ones (Lin et al. 2015; Candes et al. 2018; Lin et al. 2012). Different from prior work, our uncertainty-aware INVASE is able to quantify the confidence for instance-wise predictions.

Uncertainty quantification has two main categories from the perspective of Bayesian modeling (Kendall and Gal 2017; McDermott and Wikle 2019; Postels et al. 2019; Tagasovska and Lopez-Paz 2019), *i.e.*, *aleatoric* and *epistemic* uncertainty. Aleatoric uncertainty (a.k.a. data uncertainty) originates from the information bias of datasets, *e.g.*, noisy observations in the data. Epistemic uncertainty (a.k.a. model uncertainty) stems from the unseen inputs of a model, *e.g.*, insufficient training samples. Under such a taxonomy, the uncertainty studied in this paper can be viewed as a type of aleatoric uncertainty.

Methodology

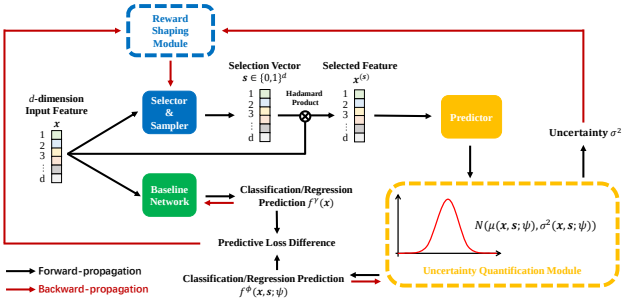


Figure 1: *Framework of uncertainty-aware INVASE.* Our baseline network is exactly the same as the vanilla INVASE, whereas we make changes in the selector and the predictor. The two new modules over the vanilla INVASE are denoted with *dotted boxes*. Based on the predictor, we apply an uncertainty quantification module to estimate the predictive uncertainty for selected features. Based on the selector, we use a reward shaping module to guide the process of uncertainty exploration toward a more precise result.

Problem Statement

Given a continuous label space $\mathcal{Y} = R$ or a proper subset of R (or its discrete c-class counterpart $\mathcal{Y} = \{1, 2, .., c\}$), we denote a label as $Y \in \mathcal{Y}$. $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times ... \times \mathcal{X}_d$ represents a d-dimension input feature space, and let $X = (X_1, X_2, ..., X_d) \in \mathcal{X}$ be a random variable. Following the notations of INVASE, a selection vector $\mathbf{s} = \{0, 1\}^d$ indicates that the i^{th}-dimension variable is selected if $\mathbf{s}_i = 1$, otherwise it is not selected. In the formulation of *instance-wise feature selection*, we are required to obtain an optimal selection $\mathbf{s}$ for a certain realization $\mathbf{x} \in \mathcal{X}$ of X. The corresponding suppressed feature vector for the i^{th} dimension is defined as:

$$\mathbf{x}_i^{(\mathbf{s})} = \begin{cases} \mathbf{x}_i & (\mathbf{s}_i = 1) \\ * & (\mathbf{s}_i = 0) \end{cases}, \tag{1}$$

where $*$ refers to that the i^{th} dimension is not chosen. We define a selection function $\mathcal{S} : \mathcal{X} \mapsto \{0, 1\}^d$ for the d-dimension instance-wise feature:

$$(Y|X^{S(\mathbf{x})} = \mathbf{x}^{S(\mathbf{x})}) \overset{dis.}{=\!=} (Y|X = \mathbf{x}), \tag{2}$$

where $\overset{dis.}{=\!=}$ means distributional equality and $S(\mathbf{x})$ is minimal in accordance with the equality.

Recapitulation of the Vanilla INVASE

In the vanilla INVASE, Kullback-Leibler (KL) divergence is leveraged to measure the "difference" between the two distributions in Equation (2). To minimize the KL divergence, INVASE defines a loss estimator $\hat{l}(\mathbf{x}, \mathbf{s})$ to approximate it for regression problems w.r.t. the training dataset D:

$$\hat{l}(\mathbf{x}, \mathbf{s}) = -\sum_{(\mathbf{x},y)\in D} (||y - f^{\phi}(\mathbf{x}, \mathbf{s})||_2 - ||y - f^{\gamma}(\mathbf{x})||_2), \tag{3}$$

where f^{γ} and f^{ϕ} is the baseline network fed with the whole feature set $\mathbf{x}$ and the predictor relied on the selected feature subset $(\mathbf{x}, \mathbf{s})$ parameterized by γ and ϕ, respectively; $||\cdot||_2$ refers to the value of l_2 loss. Intuitively, the INVASE is intended to *choose a subset* $(\mathbf{x}, \mathbf{s})$*, upon which the performance surpasses that based on all features* $\mathbf{x}$ *as much as possible*. As for classification problems, (Yoon, Jordon, and van der Schaar 2019) point out that optimizing the l_2 loss is equivalent to minimizing the KL divergence for classification when the distribution of $Y|X$ is Gaussian. Therefore, we only discuss the model under the regression setting in our remaining paper for simplicity.

Extra Optimizing Objective beyond INVASE

In Equation (3), the term $||y - f^{\phi}(\mathbf{x}, \mathbf{s})||_2$ is designed to estimate the predictive loss of selected features $\mathbf{x}^{(\mathbf{s})}$. By doing this, it treats an observation $(\mathbf{x}, y)$ in the dataset D as a sample from a distribution of which the probability density function is a *Dirac delta function* δ *without capturing uncertainty*:

$$P_D(y|\mathbf{x}, \mathbf{s}) = P^{\delta}(y - f^{\phi}(\mathbf{x}, \mathbf{s})). \tag{4}$$

Unlike INVASE, our model regards $(\mathbf{x}, y)$ in the training dataset D as a sample from a *learnable uncertainty-aware distribution parameterized by* ψ:

$$P_D(y|\mathbf{x}, \mathbf{s}) = P^{\psi}(y - f^{\phi}(\mathbf{x}, \mathbf{s})). \tag{5}$$

In this context, our modeling parameters ψ can be trained by minimizing its KL Divergence from the dataset's distribution:

$$\psi^* = \arg\min_{\psi} E_{\mathbf{x}\sim P_D(\mathbf{x})}(d_{KL}(P_D(y|\mathbf{x},\mathbf{s})||P^{\psi}(y-f^{\phi}(\mathbf{x},\mathbf{s})))) \,, \tag{6}$$

where d_{KL} is the KL divergence and E is the mathematical expectation. Besides the original goals of INVASE, our *extra optimizing objective* is to obtain ψ^* with a suitable distribution type instead of simply using the Dirac delta function. Conceptually, *our framework is highly scalable since any distribution with differentiable parameters is an eligible tool*.

Uncertain-aware INVASE

Following some prior research (Kendall and Gal 2017; He et al. 2019) on modeling uncertainty, we specify a set of Gaussian distributions to analyze the uncertainty of instance-wise feature selection.

As illustrated in Figure 1, three neural networks constitute our whole model. Among them, the baseline network f^{γ} is identical with its counterpart in vanilla INVASE. As for the other two networks, we introduce two additional modules respectively, uncertainty quantification of the predictor and reward shaping of the selector networks.

Predictor with Uncertainty Quantification Fed into the selected feature, the predictor outputs the corresponding regression result to evaluate the performance of selection. We devise an uncertainty quantification module to capture the output uncertainty of our predictor. Specifically, we introduce a network branch parameterized by ψ to learn the mean value $\mu(\mathbf{x},\mathbf{s};\psi)$ and variance $\sigma^2(\mathbf{x},\mathbf{s};\psi)$ for a certain selected feature $\mathbf{x}^{(\mathbf{s})}$. To optimize ψ as described in Equation (6), we minimize the negative log-likelihood cost in the predictor ϕ:

$$\begin{aligned} l^{\phi}(\mathbf{x},\mathbf{s};\psi) &= -\log P_D^{\phi}(y|\mathbf{x},\mathbf{s};\psi) \\ &= \frac{\log\sigma^2(\mathbf{x},\mathbf{s};\psi)}{2} + \frac{||y-\mu(\mathbf{x},\mathbf{s};\psi)||_2}{2\sigma^2(\mathbf{x},\mathbf{s};\psi)} + constant \,. \end{aligned} \tag{7}$$

Intuitively, if $\mu(\mathbf{x},\mathbf{s};\psi)$ easily fits y (with low uncertainty), the predictive bias term $||y-\mu(\mathbf{x},\mathbf{s};\psi)||_2$ tends to be small so that the first term $\frac{\log\sigma^2(\mathbf{x},\mathbf{s};\psi)}{2}$ dominates our cost. By minimizing the cost in this case, we will obtain a smaller variance σ^2. Otherwise, σ^2 is inclined to be larger if the label y is difficult to approximate (with high uncertainty).

Thus, we *quantify uncertainty with σ^2: a larger variance means higher uncertainty*. Based on the sample-wise Gaussian distribution $N(\mu,\sigma^2)$, our loss estimator is computed as:

$$\begin{aligned} \hat{l}(\mathbf{x},\mathbf{s}) = -\sum_{(\mathbf{x},y)\in D} ((&\frac{\log\sigma^2(\mathbf{x},\mathbf{s};\psi)}{2} + \frac{||y-\mu(\mathbf{x},\mathbf{s};\psi)||_2}{2\sigma^2(\mathbf{x},\mathbf{s};\psi)}) \\ &-||y-f^{\gamma}(\mathbf{x})||_2) \,, \end{aligned} \tag{8}$$

where the meaning of all notations follows Equation (3) and (7). In practice, we append a fully-connected branch to the predictor for computation of ψ.

Selector with Reward Shaping The selector f^{θ} is trained to choose an appropriate subset of instance-wise variables. The reward of its original version is defined as:

$$R(\mathbf{x},\mathbf{s}) = -\hat{l}(\mathbf{x},\mathbf{s}) - \lambda||\mathbf{s}||_0 \,, \tag{9}$$

where the l_0-norm $||\mathbf{s}||_0$ constrains the dimension number of selected features and λ is a weighting hyper-parameter. In our model, a reward shaping module encourages the selector to explore more uncertain samples, which assists in estimating uncertainty more accurately. We shape the reward by adding an uncertainty preference term to optimize the policy of our selector θ:

$$R(\mathbf{x},\mathbf{s}) = \omega\sigma^2(\mathbf{x},\mathbf{s};\psi) - \hat{l}(\mathbf{x},\mathbf{s}) - \lambda||\mathbf{s}||_0 \,, \tag{10}$$

where ω is a hyper-parameter to control the balance between uncertainty preference and the other rewards. As shown in Section , the reward shaping module makes the uncertainty adequately explored. In the testing phase, the prediction $f^{\phi}(\mathbf{x},\mathbf{s};\psi)$ for an input feature $\mathbf{x}$ with selection vector $\mathbf{s}$ gives the regression result as $\mu(\mathbf{x},\mathbf{s};\psi)$ and the uncertainty as $\sigma^2(\mathbf{x},\mathbf{s};\psi)$. In **Appendix**, we will discuss the relationship between our model and the vanilla INVASE.

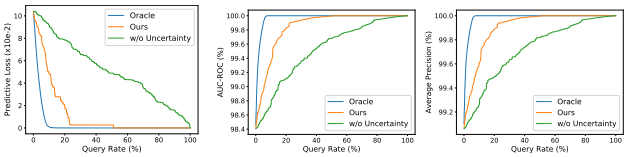
Experiments

Criteria for Evaluation Suppose that we are utilizing the uncertainty-aware INVASE to diagnose breast cancers with selected features. When we meet a highly uncertain prediction of our model, we will naturally query the exact answer from a skillful doctor. According to uncertainty scores, our goal is to achieve higher performance with fewer queries. Thus, we evaluate the model by observing the *performance gain on test data across different query rates*. For simplicity, we assume that the doctor does not make any error, *i.e.*, answers to all queries are always right. The queried samples by doctor consists of the data with uncertain of prediction thus need to be verified by doctor for correction.

Dataset and Evaluation Metric As the given implementation of INVASE, we carry out experiments on UCI-WDBC (Dua and Graff 2017) dataset, which has 569 records of a breast cancer diagnosis with 30-dimension features. Following the original setting, we hold out 80% data for training and randomly sample the test set 20 times. The weighting hyper-parameter of reward shaping ω is set as 0.1 empirically. In all the experiments, we keep the default settings identical to the vanilla INVASE if not specified particularly. We keep the same performance metrics as the vanilla INVASE, mainly area under the curve of receiver operating characteristic (AUC-ROC) and average precision (a.k.a. area under the curve of Precision-Recall, AUC-PR).

Benchmarks for Comparison Since no prior instance-wise feature selector explicitly quantifies the predictive uncertainty, there does not exist prior work for comparison. Hereby, we introduce two benchmarks, *i.e.*, "Oracle" and "w/o Uncertainty". *Oracle* is an ideal selection strategy: a sample with the largest predictive bias from the ground truth takes precedence. That can be deemed as the upper bound

of an uncertainty-aware model since every query is capable of maximizing the performance gain. Here the query means the data that are in need of further doctor verification. Another selection method denoted as *w/o Uncertainty* is that we know nothing about uncertainty and randomly choose our queries, corresponding to uncertainty-agnostic models (*e.g.*, the vanilla INVASE). The uncertainty here refers to the probability of a sample needs to be sent to doctor for further verification. As for the *our model*, we just prioritize the query about an uncertain prediction: queries are submitted in descending order of uncertainty scores (*i.e.*, the variance σ^2).



(a) Predictive Bias from Ground Truths (b) AUC-ROC (c) Average Precision (AUC-PR)

Figure 2: *Performance change tendency at different query rates.* Note: around 20 % query rate, the model accuracy has become significantly better - nearly perfect. Query rate: it is defined as the lower confidence prediction of sample versus the total of test sample.

Results As depicted in Figure 2, we make comparisons with three metrics at various query rates. It is observed that our model quickly reduces the predictive bias (l_2 testing loss) on test data from Figure 2(a). Its predictive bias decreases to nearly 0 with only about 20% query samples, whereas the uncertainty-agnostic model requires almost 100% query data to achieve similar performance. In terms of the remaining three measurements, our approach also outperforms the uncertainty-agnostic one by a large margin. For quantitative evaluation, we report the performance gain w.r.t AUC-ROC and AUC-PR in Table 1 and Table 2 at various query rates. The purpose of this table is to show with such queried data, the labels of the data will be corrected correspondingly thus leading to the increase of performance gain. The larger increase of the performance gain indicates the increased likelihood of the model to predicate the uncertain sample. Along with the growth of query rates, the performance gain of our model rises much faster than uncertainty-agnostic predictions.

Table 1: *Performance gain (%) of AUC-ROC at various query rates.* The value of AUC-ROC at 0% query rate is 98.40%.

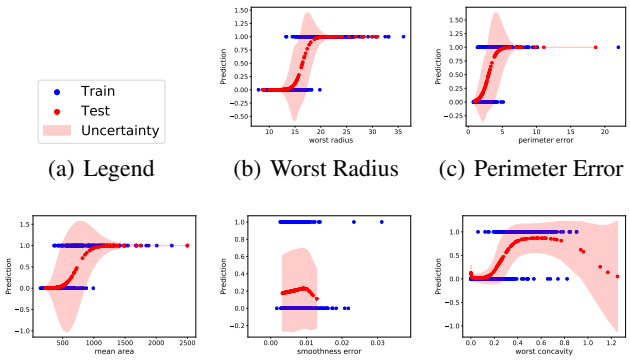
Methods	**0.1%**	**0.5%**	**1%**	**5%**	**10%**	**50%**
Oracle	0.18	0.47	0.73	1.50	1.60	1.60
w/o Uncertainty	0.00	0.01	0.02	0.15	0.33	1.28
Ours	0.03	0.11	0.18	0.54	0.95	1.60

Table 2: *Performance gain (%) of average precision (AUC-PR) at various query rates.* The value of average precision at 0% query rate is 99.06%.

Methods	**0.1%**	**0.5%**	**1%**	**5%**	**10%**	**50%**
Oracle	0.13	0.31	0.47	0.89	0.94	0.94
w/o Uncertainty	0.00	0.01	0.01	0.09	0.20	0.76
Ours	0.02	0.06	0.11	0.31	0.54	0.94

Exploration and Ablation Studies

In this paper, two additional modules is introduced to enhance the vanilla INVASE, *i.e.*, uncertainty quantification of the predictor and reward shaping of the selector. Through exploration and ablation studies, we attempt to verify their efficacy individually. It is important to note that the shaded area in Figure 3 represents the uncertainty of the model to predicate the correct positive or negative sample.



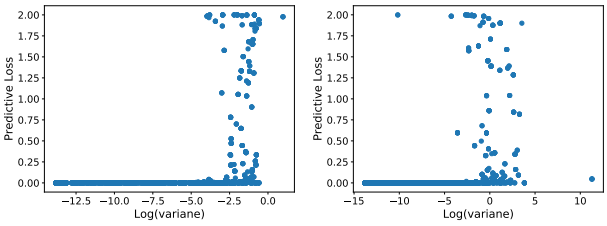
(a) Legend (b) Worst Radius (c) Perimeter Error

(d) Mean Area (e) Smoothness Error (f) Worst Concavity

Figure 3: *Exploration of uncertainty quantification.* The x-axis is the value of a certain feature and the y-axis represents the testing prediction (the training label). The red shade is the area with $x =$*feature value* and $\mu(\mathbf{x}, \mathbf{s}; \psi) - \sigma(\mathbf{x}, \mathbf{s}; \psi) \leq y \leq \mu(\mathbf{x}, \mathbf{s}; \psi) + \sigma(\mathbf{x}, \mathbf{s}; \psi)$.

Can the uncertainty quantification module capture confidence? As shown in Figure 3(b) w.r.t. the feature "Worst Radius", the test of uncertainty within the range from 12 to 20 is large since the training labels are ambiguous (mix of 0 and 1), whereas the uncertainty beyond that range is approximately 0 because the training annotations are exclusively 0 or 1. Similar correspondences also occur in the remaining features of Figure 3. The aforementioned results show that our uncertainty quantification module is helpful to uncertainty modeling.

Can the reward shaping module benefit uncertainty estimation? We adopt the predictive bias to investigate whether uncertainty is correctly estimated, As shown in Figure 4(a), variances learned with reward shaping are indeed effective indicators for mis-classified results. However, variances of the model without reward shaping scatter across a



(a) w/ Reward Shaping (b) w/o Reward Shaping

Figure 4: *Ablation of reward shaping.* (Left: with reward shaping, Right: without reward shaping) The x-axis is the value of $\log \sigma^2$, the log of predicative variance (Equation (7)) and the y-axis represents the predictive bias for the corresponding test data points.

wider range of the x-axis and have weaker relevance with the predictive bias.

Conclusion

In this paper, we present an uncertainty-aware INVASE to quantify predictive confidence. The model is able to quantify the potential errors of our instance-wise feature selection, which may be beneficial to some healthcare problems. In theory, the proposed approach extends the modeling perspective from a Dirac delta function to a learnable uncertainty-aware distribution. Conceptually, it is a highly scalable framework, of which any distribution with differentiable parameters is an eligible tool. To be specific, we apply Gaussian distributions to capture uncertainty with their variances. Accordingly, we implement the uncertainty-aware model with two extra modules over the raw INVASE, *i.e.*, uncertainty quantification of the predictor and reward shaping of the selector. To evaluate our method, we carry out experiments on UCI-WDBC w.r.t. the whole model and each new component. Experimental results show that our approach discovers overwhelming majority testing errors with only about 20% queries, whereas the uncertainty-agnostic counterparts need nearly 100% query samples for the same performance gain.

References

Candes, E.; Fan, Y.; Janson, L.; and Lv, J. 2018. Panning for gold:'model-X'knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80(3): 551–577.

Chen, J.; Song, L.; Wainwright, M.; and Jordan, M. 2018. Learning to Explain: An Information-Theoretic Perspective on Model Interpretation. In *International Conference on Machine Learning*, 883–892.

Dua, D.; and Graff, C. 2017. UCI Machine Learning Repository. URL http://archive.ics.uci.edu/ml.

He, Y.; Zhu, C.; Wang, J.; Savvides, M.; and Zhang, X. 2019. Bounding box regression with uncertainty for accurate object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2888–2897.

Howell, A.; Anderson, A. S.; Clarke, R. B.; Duffy, S. W.; Evans, D. G.; Garcia-Closas, M.; Gescher, A. J.; Key, T. J.; Saxton, J. M.; and Harvie, M. N. 2014. Risk determination and prevention of breast cancer. *Breast Cancer Research* 16(5): 446.

Kendall, A.; and Gal, Y. 2017. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, 5574–5584.

Lin, C.; Miller, T.; Dligach, D.; Plenge, R.; Karlson, E.; and Savova, G. 2012. Maximal information coefficient for feature selection for clinical document classification. In *ICML Workshop on Machine Learning for Clinical Data. Edingburgh, UK*.

Lin, Y.; Hu, Q.; Liu, J.; and Duan, J. 2015. Multi-label feature selection based on max-dependency and min-redundancy. *Neurocomputing* 168: 92 – 103. ISSN 0925-2312. doi:https://doi.org/10.1016/j.neucom.2015.06.010. URL http://www.sciencedirect.com/science/article/pii/S0925231215008309.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, 4765–4774.

McDermott, P. L.; and Wikle, C. K. 2019. Bayesian recurrent neural network models for forecasting and quantifying uncertainty in spatial-temporal data. *Entropy* 21(2): 184.

Postels, J.; Ferroni, F.; Coskun, H.; Navab, N.; and Tombari, F. 2019. Sampling-free Epistemic Uncertainty Estimation Using Approximated Variance Propagation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2931–2940.

Shrikumar, A.; Greenside, P.; and Kundaje, A. 2017. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 3145–3153. JMLR. org.

Tagasovska, N.; and Lopez-Paz, D. 2019. Single-Model Uncertainties for Deep Learning. In *Advances in Neural Information Processing Systems*, 6414–6425.

Tonekaboni, S.; Joshi, S.; McCradden, M. D.; and Goldenberg, A. 2019. What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use. In *Machine Learning for Healthcare Conference*, 359–380.

Wang, D.; Khosla, A.; Gargeya, R.; Irshad, H.; and Beck, A. H. 2016. Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718* .

Yoon, J.; Jordon, J.; and van der Schaar, M. 2019. INVASE: Instance-wise Variable Selection using Neural Networks. In *International Conference on Learning Representations*. URL https://openreview.net/forum?id=BJg_roAcK7.

Appendix
Uncertainty-aware INVASE: Enhanced Breast Cancer Diagnosis Feature Selection

Jia-Xing Zhong,[1] **Hongbo Zhang** [2]
[1] School of Electronic and Computer Engineering, Peking University
[2] Department of Computer and Information Sciences, Virginia Military Institute
jxzhong@pku.edu.cn, hbzhang@vt.edu

1 Detailed Analysis on the Equivalency between our Model with $\sigma^2 \to 0$ and the vanilla INVASE

The differences between the vanilla INVASE and ours are based on two new modules: uncertainty quantification and reward shaping. To analyze the equivalency, *all we need to do is to prove the equivalency w.r.t. those two components*.

Uncertainty Quantification In our model, we treat data as samples from learnable uncertainty-aware distributions as shown in Equation (5):

$$P_D(y|\mathbf{x},\mathbf{s}) = P^{\psi}(y - f^{\phi}(\mathbf{x},\mathbf{s}))\,.$$

where $P_D(y|\mathbf{x},\mathbf{s}) \sim N(\mu,\sigma^2)$ in our model. We should prove the Gaussian distribution approaches to a Dirac delta function when $\sigma^2 \to 0$. Given a Gaussian distribution $N(\mu,\sigma^2)$, the probability density function of a variable t is:

$$f(t;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\exp(-\frac{(t-\mu)^2}{2\sigma^2})\,.$$

Hence,

$$\lim_{\sigma^2\to 0} f(t;\mu,\sigma) = \begin{cases} \infty & (t=\mu) \\ 0 & (t\neq\mu) \end{cases},$$

where $\int f(t;\mu,\sigma)dt = 1$ according to the characteristics of a probability density function. By definition, $\delta(t-\mu) = \lim_{\sigma^2\to 0} f(t;\mu,\sigma)$.

Reward Shaping In Equation (10), the total reward is defined as:

$$R(\mathbf{x},\mathbf{s}) = \omega\sigma^2(\mathbf{x},\mathbf{s};\psi) - \hat{l}(\mathbf{x},\mathbf{s}) - \lambda||\mathbf{s}||_0\,,$$

if $\sigma^2(\mathbf{x},\mathbf{s};\psi)\to 0$, then $R(\mathbf{x},\mathbf{s}) \to -\hat{l}(\mathbf{x},\mathbf{s}) - \lambda||\mathbf{s}||_0$. That is just the form of Equation (9), so they are equivalent.

Therefore, the vanilla INVASE is a *particular case* of our uncertainty-aware INVASE. When the variance $\sigma^2 \to 0$, the Gaussian distribution $N(\mu,\sigma^2)$ of the predictor approaches to a Dirac delta function δ, in which case $\mu(\mathbf{x},\mathbf{s};\psi) \to f^{\phi}(\mathbf{x},\mathbf{s})$. Meanwhile, the additional shaping reward of the selector $\omega\sigma^2(\mathbf{x},\mathbf{s};\psi) \to 0$. Hence, our uncertainty-aware model degrades into the raw INVASE if the variance of all data points is infinitesimal. That corresponds to the condition in which every prediction is considered to be absolutely sure.

2 Proof of the Formula in Equation (7)

Maximum Likelihood Estimation as minimizing KL Divergence The Equation (6) is our optimization objective:

$$\psi^* = \arg\min_{\psi} E_{\mathbf{x}\sim P_D(\mathbf{x})}(d_{KL}(P_D(y|\mathbf{x},\mathbf{s})||P^{\psi}(y - f^{\phi}(\mathbf{x},\mathbf{s})))\,.$$

The equivalence to maximum likelihood can be found in (Bishop 2006) as a ready-made theorem, which is omitted here for space limitations.

Loss of Gaussian Maximum Likelihood Estimation (Nix and Weigend 1994) provide a similar conclusion for common discriminative problems. In terms of our instance-wise feature selection problem:

$$\begin{aligned} l^{\phi}(\mathbf{x},\mathbf{s};\psi) &= -\log P_D^{\phi}(y|\mathbf{x},\mathbf{s};\psi) \\ &= -\log(\frac{1}{\sqrt{2\pi\sigma^2(\mathbf{x},\mathbf{s};\psi)}}\exp\,(-\frac{||y-\mu(\mathbf{x},\mathbf{s};\psi)||_2}{2\sigma^2(\mathbf{x},\mathbf{s};\psi)})) \\ &= \frac{\log\sigma^2(\mathbf{x},\mathbf{s};\psi)}{2} + \frac{||y-\mu(\mathbf{x},\mathbf{s};\psi)||_2}{2\sigma^2(\mathbf{x},\mathbf{s};\psi)} + \frac{\log 2\pi}{2} \\ &= \frac{\log\sigma^2(\mathbf{x},\mathbf{s};\psi)}{2} + \frac{||y-\mu(\mathbf{x},\mathbf{s};\psi)||_2}{2\sigma^2(\mathbf{x},\mathbf{s};\psi)} + constant\,. \end{aligned}$$

That is identical to Equation (7).

3 Pseudo-codes

Please refer to Algorithm 1.

4 Implementation Details

In practice, we append a 2-layer fully-connected 100-dimension branch to the predictor for computation of ψ, of which the shape and BatchNorm settings are consistent with the raw predictor.

In terms of Equation (7), we have a term to optimize:

$$\frac{\log\sigma^2(\mathbf{x},\mathbf{s};\psi)}{2} + \frac{||y-\mu(\mathbf{x},\mathbf{s};\psi)||_2}{2\sigma^2(\mathbf{x},\mathbf{s};\psi)}\,.$$

Algorithm 1 Training Process of Uncertainty-aware INVASE.

Input:

α: learning rate of selector
β: learning rate of baseline network and predictor
n: batch size
D: dataset
ω: a hyper-parametric weight for reward shaping
λ: a hyper-parametric weight for l_1-norm of feature dimensions

Output:

θ: learned parameters of selector
ϕ: learned parameters of predictor
γ: learned parameters of baseline network
ψ: learned parameters of uncertainty quantification

repeat
  Sample a mini-batch $(\mathbf{x}_j, y_j)_{j=1}^n$ from D
  for j=1,...,n **do**
    Compute selection probabilities: $\mathbf{p}_j = S^\theta(\mathbf{x}_j)$
    Obtain selection vector: $\mathbf{s}_j \sim Ber(\mathbf{x}_j)$
    Estimate loss difference:
    $\hat{l}(\mathbf{x}_j, \mathbf{s}_j) = -(\frac{\log \sigma^2(\mathbf{x}_j, \mathbf{s}_j; \psi)}{2} + \frac{||y_j - \mu(\mathbf{x}_j, \mathbf{s}_j; \psi)||_2}{2\sigma^2(\mathbf{x}_j, \mathbf{s}_j; \psi)}) - ||y - f^\gamma(\mathbf{x}_j)||_2)$
  Update the selector:
  $\theta = \theta - \alpha \frac{1}{n} \sum_{(\mathbf{x},y)\in batch} (\omega\sigma^2(\mathbf{x}, \mathbf{s}; \psi) - \hat{l}(\mathbf{x}, \mathbf{s}) - \lambda||\mathbf{s}||_0) \nabla_\theta \log \pi_\theta(\mathbf{x}, \mathbf{s})$
  Update predictor with uncertainty quantification:
  $\phi = \phi - \beta \frac{1}{n} \sum_{(\mathbf{x},y)\in batch} \nabla_\phi l^\phi(\mathbf{x}, \mathbf{s}; \psi)$
  $\psi = \psi - \beta \frac{1}{n} \sum_{(\mathbf{x},y)\in batch} \nabla_\psi l^\phi(\mathbf{x}, \mathbf{s}; \psi)$
  Update baseline network:
  $\gamma = \gamma - \beta \frac{2}{n} \sum_{(\mathbf{x},y)\in batch} \mathbf{x}(f^\gamma(\mathbf{x}) - y)$
until Convergence

However, that is numerically unstable: if $\sigma^2(\mathbf{x}, \mathbf{s}; \psi) = 0$, the second component $\frac{||y-\mu(\mathbf{x},\mathbf{s};\psi)||_2}{2\sigma^2(\mathbf{x},\mathbf{s};\psi)}$ will become infinite. Following (Kendall and Gal 2017), we actually utilize $\log \sigma^2$ as the computing unit. Therefore, the term is rewritten as:

$$\frac{\log \sigma^2(\mathbf{x}, \mathbf{s}; \psi)}{2} + \frac{1}{2} \exp(-\log \sigma^2(\mathbf{x}, \mathbf{s}; \psi)) \, ||y - \mu(\mathbf{x}, \mathbf{s}; \psi)||_2 \, .$$

References

Bishop, C. M. 2006. *Pattern Recognition and Machine Learning*. Springer.

Kendall, A.; and Gal, Y. 2017. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, 5574–5584.

Nix, D. A.; and Weigend, A. S. 1994. Estimating the mean and variance of the target probability distribution. In *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, volume 1, 55–60 vol.1. ISSN null. doi:10.1109/ICNN.1994.374138.