Not all Mistakes are Equal

Extended Abstract

Murat Sensoy
Blue Prism AI Labs
London, UK
murat.sensoy@blueprism.com

Maryam Saleki
Ozyegin University
Istanbul, Turkey
maryam.saleki@ozu.edu.tr

Simon Julier
University College London
London, UK
s.julier@cs.ucl.ac.uk

Reyhan Aydoğan
Ozyegin University, Istanbul, Turkey
TU Delft, Delft, The Netherlands
reyhan.aydogan@ozyegin.edu.tr

John Reid
Blue Prism AI Labs
London, UK
john.reid@blueprism.com

ABSTRACT

In many tasks, classifiers play a fundamental role in the way an agent behaves. Most rational agents collect sensor data from the environment, classify it, and act based on that classification. Recently, deep neural networks (DNNs) have become the dominant approach to develop classifiers due to their excellent performance. When training and evaluating the performance of DNNs, it is normally assumed that the cost of all misclassification errors are equal. However, this is unlikely to be true in practice. Incorrect classification predictions can cause an agent to take inappropriate actions. The costs of these actions can be asymmetric, vary from agent-to-agent, and depend on context. In this paper, we discuss the importance of considering risk and uncertainty quantification together to reduce agents' cost of making misclassifications using deep classifiers.

KEYWORDS

Deep learning; Uncertainty; Risk; Cost-sensitive learning

ACM Reference Format:
Murat Sensoy, Maryam Saleki, Simon Julier, Reyhan Aydoğan, and John Reid. 2020. Not all Mistakes are Equal. In *Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), Auckland, New Zealand, May 9–13, 2020,* IFAAMAS, 3 pages.

1 INTRODUCTION

Recent developments in deep learning and neural networks mean that they have become one of the most prevalent approaches for machine learning. They are used in many applications from medical diagnosis to image recognition. In some classification problems, their performance has been shown to meet or even exceed human levels of performance. However, unlike humans, existing models do not reason about the consequences of their possible mistakes. This is evidenced by the choice of the cross entropy as the most common measure of loss in deep classifiers. The cross entropy loss is computed using the predicted probability of the correct category and completely ignores how the remaining probability mass is distributed over the wrong categories. Thus, the classifier is trained without regard to the risk of incorrect classification decisions.

Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, G. Sukthankar (eds.), May 9–13, 2020, Auckland, New Zealand.

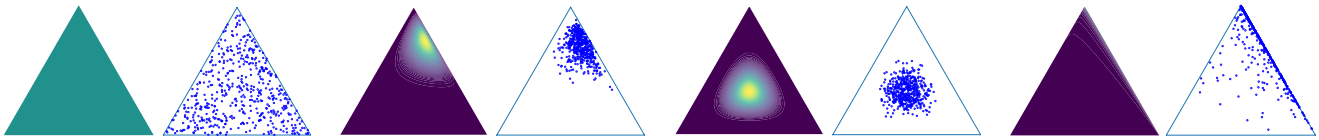
An agent solely depending on the predictions of such classifiers to make a decision or take an action may pay a high cost when these prediction are wrong. A striking example is an incident happened on 7th May 2016, near Williston, Florida, USA. A car operating with automated vehicle control systems crashed into a tractor semi-trailer truck. Unfortunately, the car driver died due to the sustained injuries. The car manufacturer stated that the accident originated from the vision system which incorrectly classified the white truck as a bright sky [7] and acted based on this erroneous prediction. As deep learning is used more widely and in more critical decision making operations, these difficulties will become more prevalent.

To incorporate the cost of misclassification into the training of deep classifiers, one approach is to use cost-sensitive learning [3]. This aims to minimize the expected cost of classification errors, e.g., by avoiding predictions placing high probabilities for high-risk categories, whilst maintaining classification accuracy. However, similarly to standard classifiers, cost-sensitive classifiers do not have any mechanism for quantifying the uncertainty of their predictions. Thus, an autonomous agent using these classifiers cannot know if it can rely on their predictions to make a decision or take an action. Hence, the main advantage of these classifiers for the agent is limited to decreasing the cost of classification errors due to their tendency to predict less risky categories. On the other hand, if the agent is equipped with tools to quantify the uncertainty of these predictions, it can avoid taking actions based on ungrounded and most likely wrong predictions. Recently, a number of methods have been proposed to quantify uncertainty of deep classifier predictions. Among those, evidential deep learning is the state-of-the-art and a practical approach for uncertainty quantification in deep classifiers [8]. However, these methods do not take into account the risk of classification errors and may still be overconfident for some high-risk categories.

In the rest of the paper, we discuss how uncertainty quantification in deep classifiers and the risk of making wrong classification decisions can be combined to reduce the risk for decision making for autonomous agents.

2 UNCERTAINTY AND DEEP CLASSIFIERS

In neural networks, the *softmax* function is frequently used to compute a predictive categorical distribution over possible categories for an input sample. Given a vector of logit values z, the i^{th} element for the output vector of the softmax function is defined as



(a) $\boldsymbol{\alpha} = [1, 1, 1]$ (b) $\boldsymbol{\alpha} = [2, 5, 15]$ (c) $\boldsymbol{\alpha} = [10, 10, 10]$ (d) $\boldsymbol{\alpha} = [0.1, 0.9, 2]$

Figure 1: Density plots (blue = low, red = high) for the Dirichlet distributions over the probability simplex in $\mathbb{R}^3$ for various values of the $\boldsymbol{\alpha}$ parameters and 500 categorical distributions sampled from each of these Dirichlet distributions.

$softmax_i(\boldsymbol{z}) = e^{z_i}/\sum_j e^{z_j}$. Since Dirichlet distribution is prior for categorical distribution, it can be used as a distribution over all possible softmax outputs for the classification of a given sample. This allows us to represent uncertainty of predictions for the classification of a sample through the variance of the corresponding Dirichlet distribution.

The Dirichlet distribution is the conjugate prior of the categorical and multinomial distributions. It is a probability density function (pdf) for possible values of the probability mass function (pmf) $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_K]$ over K categories. It is characterized by parameters $\boldsymbol{\alpha} = [\alpha_1, \cdots, \alpha_K]$ and is given by

$$\text{Dirichlet}(\boldsymbol{\pi}|\boldsymbol{\alpha}) = \begin{cases} \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{K} \pi_i^{\alpha_i - 1} & \text{for } \boldsymbol{\pi} \in \mathcal{S}_K, \\ 0 & \text{otherwise,} \end{cases} \quad (1)$$

where $\mathcal{S}_K$ is the K-dimensional unit simplex and $B(\boldsymbol{\alpha})$ is the K-dimensional multinomial beta function [5].

A Dirichlet distribution can be used to model the probability density of categorical distributions, each of which can be interpreted as a probability distribution for assigning a sample to one of K categories, e.g., as in the classification problems. Figure 1 demonstrates Dirichlet distributions over three categories. In this case, each Dirichlet distribution has three parameters ($K = 3$), i.e., one parameter for each category. When all parameters are one (i.e., $\boldsymbol{\alpha} = [1, 1, 1]$), the Dirichlet distribution is uniform, which means that all categorical distributions over these three categories are equally likely.

The parameters of a Dirichlet distribution are considered as real-valued pseudocounts [6]. The parameters of the uniform Dirichlet distribution is usually taken as the prior counts $\boldsymbol{\beta}$ to which observations or evidence for the training data is added. The resulting parameters (pseudocounts) define the updated (posterior) Dirichlet distribution. Let $[1, 4, 14]$ be the evidence (e.g., observations) to be added to the prior counts $\boldsymbol{\beta} = [1, 1, 1]$, then the posterior Dirichlet distribution will have the parameters $\boldsymbol{\alpha} = [2, 5, 15]$, which indicates that categorical distributions placing more mass on the third category are more likely than others, as shown in Figure 1(b). Similarly, if the evidence vector is $[9, 9, 9]$, the resulting Dirichlet distribution parameters become $\boldsymbol{\alpha} = [10, 10, 10]$, which indicates that the categorical distributions placing similar amount of mass on all categories become more likely, as shown in Figure 1(c).

3 REDUCING DECISION MAKING RISK

Agents usually use classifiers when they need to choose one option among several alternatives during decision making. Pignistic probabilities have been introduced in decision theory to represent the probability that a rational agent will choose a particular option when it is required to make a decision [1, 9]. The pignistic probabilities $\boldsymbol{p} = [p_1, \ldots, p_K]$ are mathematically equivalent to the Shapley value in game theory [2] and inherently incorporate the decision maker's uncertainty when choosing one of K options (i.e., the uncertainty related to $\boldsymbol{\pi}$) and the incurred risk of choosing each one. Hence, while calculating the expected risk of choosing one category as the label of a sample, the pignistic probabilities ($\boldsymbol{p}$) should replace the categorical probabilities ($\boldsymbol{\pi}$) to account for the risk of misclassification. In the settings where there is no risk for misclassification or each misclassification has the same risk, $\boldsymbol{p}$ should be equal to $\boldsymbol{\pi}$. However, in other settings, there may be some divergence between these two probabilities.

In this paper, we argue that the pignistic probabilities can be implemented by redistributing the prior counts in the uniform distribution to account for the associated risk of making wrong classification decisions when an evidential deep classifier is uncertain. Let $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_K]$ be the predicted parameters for the Dirichlet distribution for $\boldsymbol{\pi}$, for the classification of a sample $\boldsymbol{x}$. Then, the parameters of the Dirichlet distribution for the pignistic probabilities $\boldsymbol{p} = [p_1, \ldots, p_K]$, is calculated as $\boldsymbol{\alpha} - \mathbf{1} + K\boldsymbol{\gamma_\theta}(\boldsymbol{x})$, where $\mathbf{1} = [1, \ldots, 1]$ is the uniform prior counts and $\boldsymbol{\gamma_\theta}(\cdot)$ is a function (deeply parametrized by θ) calculating how to redistribute the prior counts for $\boldsymbol{x}$ to reduce the risk. For example, Figure 1(d) shows the resulting Dirichlet distribution after redistributing the counts in the uniform Dirichlet distribution in Figure 1(a) when $[0.033, 0.3, 0.667]$ is the output of $\boldsymbol{\gamma_\theta}(\boldsymbol{x})$. This Dirichlet distribution generates almost no probability mass for the first category while placing more probability mass on the third category. This is a desired prior for the pignistic probabilities of a sample if the misclassification risk for the sample is inversely proportional to the redistributed counts.

4 CONCLUSIONS

As a result of the success of deep learning in recent years, deep classifiers are now an indispensable part of autonomous systems. However, these black-box models may be very confident when their predictions are wrong and lead autonomous agents to make mistakes in their decisions [4]. Furthermore, standard training of deep models neglects that different mistakes involve in different level of risk for the agents depending them. In this paper, we discuss how one of the recent methods for uncertainty quantification for deep classifiers, i.e., evidential deep learning [8], can be extended to reduce misclassification risk for autonomous agents. In future, we will implement this approach by incorporating the notion of risk and pignistic probabilities into deep evidential classifiers, and evaluate how much it minimizes the cost of misclassification for autonomous agents.

REFERENCES

[1] A.P. Dempster. 2008. A generalization of Bayesian inference. In *Classic works of the Dempster-Shafer theory of belief functions*. Springer, 73–104.

[2] Didier Dubois, Henri Prade, and Philippe Smets. 2008. A definition of subjective possibility. *International Journal of Approximate Reasoning* 48, 2 (2008), 352–364.

[3] Charles Elkan. 2001. The Foundations of Cost-Sensitive Learning. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*. 973–978.

[4] Y. Gal and Z. Ghahramani. 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *ICML*.

[5] S. Kotz, N. Balakrishnan, and N.L. Johnson. 2000. *Continuous Multivariate Distributions*. Vol. 1. Wiley, New York.

[6] Kevin P Murphy. 2012. *Machine learning: a probabilistic perspective*. MIT press.

[7] NHTSA. 2016. PE 16-007. Technical report, U.S. Department of Transportation, National Highway Traffic Safety Administration, Jan 2017. Tesla Crash Preliminary Evaluation Report. (2016).

[8] Murat Sensoy, Lance Kaplan, and Melih Kandemir. 2018. Evidential Deep Learning to Quantify Classification Uncertainty. In *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.). 3179–3189.

[9] Philippe Smets and Robert Kennes. 1994. The transferable belief model. *Artificial intelligence* 66, 2 (1994), 191–234.