EMPIRICAL CONFIDENCE ESTIMATES FOR CLASSIFICATION BY DEEP NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

How well can we estimate the probability that the classification predicted by a deep neural network is correct (or in the Top 5)? It is well-known that the softmax values of the network are not estimates of the probabilities of class labels. However, there is a misconception that these values are not informative. We define the notion of *implied loss* and prove that if an uncertainty measure is an implied loss, then low uncertainty means high probability of correct (or top k) classification on the test set. We demonstrate empirically that these values can be used to measure the confidence that the classification is correct. Our method is simple to use on existing networks: we proposed confidence measures for Top k which can be evaluated by binning values on the test set.

1 INTRODUCTION

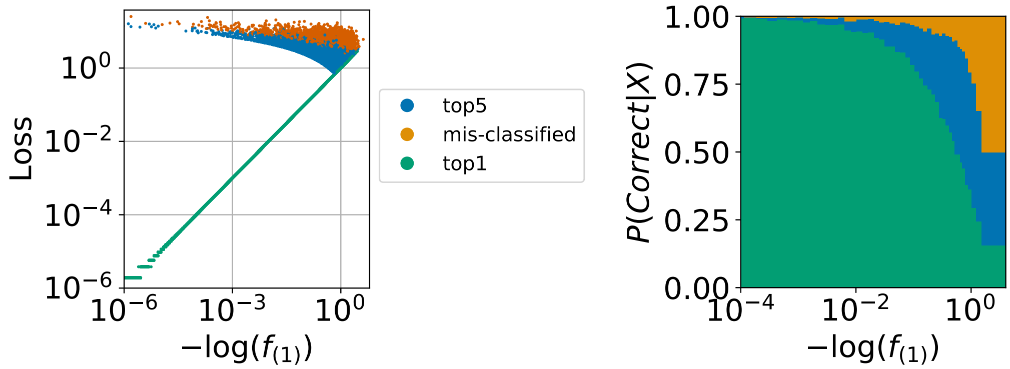
Despite lots of effort to build confidence measures for classification by deep neural networks, there is still a lot of confusion about the value and applicability of these measures. In this article we present a simple method for estimating confidence based on *implied loss* values which leads to results which are empirically more accurate than benchmarks on test sets. We prove that high confidence values imply a high probability of correct classification on test sets.

Many have observed that used blindly, the maximum softmax probability of a network does a poor job of predicting uncertainty (Nguyen & O'Connor, 2015; Provost et al., 1998; Nguyen et al., 2015; Yu et al., 2011; Lakshminarayanan et al., 2017). However, Zaragoza & d'Alché Buc (1998) showed in the 1990s that on on shallow networks the maximum softmax probability and the (negative) entropy of the probabilities strongly correlate with model confidence on in-distribution images. More recently, in the deep setting, Hendrycks & Gimpel (2017) showed empirically that the maximum softmax probability can be used to predict network confidence. Our implied loss interpretation justifies both methods, since we demonstrate that both these quantities are uncertainly measures. Moreover, we extend the uncertainty metric to Top k predictions. We show that, in conjunction with binning, simple uncertainty statistics outperform common Bayesian approaches like MC-dropout as a measure of confidence, at a fraction of the computational cost.

Using this simple idea, we make the following contributions.

1. We make accurate estimates of the probability that the classification of the model on a test set is correct. This works for existing models (no need to retrain), using a simple tabular form (see Table 2 for Imagenet).
2. We can discover mislabelled data in a consistent manner, see Figure 2 and we can detect off manifold data and adversarial examples.
3. We give a simple definition of uncertainty, which applies to previously proposed methods, and leads to a proof that low uncertainty (high confidence) implies high probability of correct classification. It applies to both Top 1 and Top k uncertainty.

We advocate evaluating model uncertainty via expected *Bayes factors* (Kass & Raftery, 1995), which provide a rigorous probabilistic approach to evaluating uncertainty, and are widely used for hypothesis testing in other scientific fields, see for example (Good, 1979) and (Jeffreys, 2003). Bayes factors are more informative than Brier scores in the current setting, where the probability of correct classification is high.



(a) Loss vs U_1

(b) Histogram

Figure 1: Figure 1(a) Scatter plot to indicate how predictive U_1 is compared to the loss. For small values of U_1, the loss is small with high probability. Figure 1(b): the probability correct (green) or Top5 (blue) given the value of U_1.

2 Prior work

As neural networks are adopted into safety critical systems, the need for neural network uncertainty estimates has become abundantly clear. Indeed, any accident adverse system must by design incorporate notions of uncertainty (Amodei et al., 2016). Real-world examples abound: uncertainty measures are needed in autonomous vehicles (Feng et al., 2018), robotics (Richter & Roy, 2017), medical imaging (Ching et al., 2018; DeVries & Taylor, 2018b) and medical decision making (Begoli et al., 2019), and semantic understanding (Kendall et al., 2017).

Much effort has been dedicated to addressing this deficiency. Many works have placed neural networks within a Bayesian probabilistic framework. Initial work placed Bayesian priors on model weights (MacKay, 1992a; Neal, 1996), leading to Bayesian neural networks, however this has proven difficult to implement in practice. Many techniques have been developed to overcome this difficulty (MacKay, 1992b; Neal, 1996; Graves, 2011; Hasenclever et al., 2017; Li et al., 2015; Balan et al., 2015; Welling & Teh, 2011; Springenberg et al., 2016). One promising approach in the deep learning setting is to perform *approximate* posterior inference (Louizos & Welling, 2016; Hernández-Lobato & Adams, 2015; Blundell et al., 2015; Sun et al., 2017).

Due to its simplicity, dropout is widely used as a surrogate for uncertainty. Dropout (Srivastava et al., 2014) was interpreted in a Bayesian setting by Gal & Ghahramani (2016) and Kingma et al. (2015), however, there are problems with this interpretation, see (Hron et al., 2018) for a recent discussion. Dropout involves evaluating an ensemble of models at test time, which can be both memory and computationally intensive for very large networks.

Non-Bayesian model ensembles have also been developed (Dietterich, 2000), for a recent survey see (Li et al., 2018). Lakshminarayanan et al. (2017) train an ensemble of adversarially robust models and empirically showed an improvement in uncertainty estimates over dropout based methods. Geifman et al. (2018) proposed using an early stopping criteria to collate an ensemble of models. Kristiadi & Fischer (2019) use mixture modeling to chose ensemble weights.

Several deep learning specific approaches have been proposed in recent years, especially in the context of detecting out-of-distribution samples. Oliveira et al. (2016) suggest detecting outliers via an anomaly detector. Lee et al. (2018) generator out-of-distribution images through a GAN; the classifier is trained to assign the equal weight probability vector to these images. Hendrycks et al. (2018) train networks on two distributions: the in-distribution samples, and out-of-distribution samples. Liu et al. (2018) develop PAC-style guarantees on detection of out-of-distribution samples. Several recent works (Jiang et al., 2018; Papernot & McDaniel, 2018; Mandelbaum & Weinshall, 2017) have suggested using nearest neighbour distances, in feature space, for outlier detection and confidence measures. DeVries & Taylor (2018a) suggest training an additional network to predict

Figure 2: Visualization of the images in the upper left of Figure 1(a). The confident images which were labelled incorrectly turned out to be mislabelled or ambiguous. For example, in the second image, the animal is a wallaby, not a wombat. In the fourth image, a paintbrush is a kind of plant, but there is also a pot in the image.

uncertainty; Malinin & Gales (2018) specifically model prediction probabilities with a Dirichlet distribution, which implicitly describes model uncertainty.

Platt (2000) proposed scaling SVM predictions to better match the validation set; this has been generalized to neural networks and multiclass classification (Niculescu-Mizil & Caruana, 2005; Guo et al., 2017). Other scaling approaches, such as changing the softmax temperature, have shown promise (Guo et al., 2017; Liang et al., 2018). Another popular approach to calibration is based on *binning* model probabilities, developed by Zadrozny & Elkan (2001). Each bin is assigned a probability of being correct, which is obtained by minimizing the Brier score of the bins (Brier, 1950). Bins edges may be optimized as well (Zadrozny & Elkan, 2002); and can be extended to the Bayesian setting by assigning a prior on binning schemes (Naeini et al., 2015).

3 Confidence measures based on implied loss

Suppose a model $f(x)$, generalizes well, so that it has a high probability, p, of a correct prediction on an image x sampled from the same underlying distribution. Write

$$I_k(f) = \{\text{indices of the } k \text{ largest components of } f\} \tag{1}$$

for the top k indices. The classification of the vector f is given by the largest component, $C(f) = I_1(f)$. Define the random variables

$$X_k = \begin{cases} 1 & \text{if } y(x) \in I_k(f(x)) \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

which are Bernoulli random variables with expected values

$$p_k = \mathbb{E}\left[X_k\right] \tag{3}$$

of the probability that the correct label is in the Top k.

We want to estimate p_k. Define random variables U_k, which we call *uncertainties*, whose statistics allow us to better estimate p_k. We define $U_1(x)$, the implied loss, to be the *loss, given that the classification was correct.*

$$U_1(x) = \{\mathcal{L}(f(x), y) \mid y = C(f(x))\} \tag{4}$$

where $\mathcal{L}$ is the loss used to train the network. In the case of interest corresponding to the Kullback-Leibler losses, define

$$U_k(x) = -\log\left(\sum_{i=1}^{k} f_i^{sort}\right), \tag{5}$$

where f^{sort} corresponds to the indices of f sorted in decreasing order.

The histograms of the uncertainty variables will result in an estimate of the conditional probability that the classification is correct, given the uncertainty value,

$$\text{Prob}\left(X_k(x) = 1 \mid U_k(x) = t\right).$$

The histogram of U_1 is plotted on the test set in Figure 1(a). Note that for small values of U_1, the images have a very high probability of being correct. In fact, we can use U_1 to detect incorrectly classified images: we visualized the images which smallest value of U_1 (i.e. highest confidence), which correspond to the few isolated points in the upper left of the figure. It turned out that all of these were either incorrectly labelled, or were ambiguous images, see illustrations in Figure 2. In the second part of Figure 1(b) we illustrate more quantitatively the Top 1 (green) and Top 5 (green or blue) probabilities conditioned on the 100 histograms bins of $-\log(p_{\max})$ on test set for ResNet152 on ImageNet. The Top 1 probability conditioned on the lower bins is very close to 100%. The Top 5 probability is no better than 50% on the last few bins. The intermediate bins are less informative.

4 Uncertainty estimates

We establish asymptotic confidence estimates for the uncertainty measures are given in (5).

We give a definition of uncertainty measures for general losses.

Definition 4.1. Given $\epsilon > 0$, and the uncertainty measure $U(x)$, define the set

$$S_k^\epsilon = \{U_k(x) \leq \epsilon \text{ and } y \notin I_k(x)\} \tag{6}$$

The uncertainty measure $U_k(x)$ is an implied loss if the event S_k^ϵ has high expected loss. For general losses, an implied loss is given by

$$U_k(x) = \mathcal{L}(f, y_w)$$

where y_w is the $(k+1)$-th ranked label. With the Kullback-Leibler loss, the (negative) entropy of the probabilities is also an uncertainty measure.

4.1 Top 1 uncertainty

The next theorem shows that if the uncertainty is small, then the probability of correct classification must be high.

Theorem 4.2 (Confidence estimate). *Define $U_1(x)$ by (5) and define S^ϵ by (6), and let $\mathcal{L}_{KL}$ be the Kullback-Leibler loss. Then*

$$\text{Prob}\left(S^\epsilon\right) \leq \frac{\mathbb{E}\left[\mathcal{L}_{KL}(f(x), y)\right]}{\log\left(\frac{1}{\epsilon}\right)} \tag{7}$$

Proof. Claim: Let $\epsilon > 0$ be small. By assumption, $-\log f_1^{sort} \leq \epsilon$. Thus $f_1^{sort} \geq \exp(-\epsilon)$. Let e_k be the correct label. Then $f_k \leq f_1^{sort}$, so

$$f_k \leq 1 - \exp(-\epsilon)$$

and

$$-\log(f_k) \geq -\log(1-\exp(-\epsilon)) \geq \log(1/\epsilon).$$

Thus for $x \in S^\epsilon$, $\mathcal{L}_{KL}(f(x), y(x)) \geq \log(1/\epsilon)$. Apply Markov's inequality (11) to the random variable $L(x) = \mathcal{L}(f(x), y(x))$ to obtain the result. □

Remark 4.3 (Neural Networks are always overconfident). Note that the uncertainly is always less than the loss,

$$U_1(f) \leq \mathcal{L}_{KL}(f, e_k) \tag{8}$$

with equality when $C(f(x)) = y(x)$.

4.2 Top k uncertainty

In the next result we show that if the top k uncertainty is small, then the probability that the correct labels is in the top k must be high. The result can also be proven in the case of general losses, and uncertainly measures satisfying (6).

Consider the event S_k^ϵ (6) for a given $k \geq 1$. If the correct label is not in the top k, then the probability of the correct label, f_c, must satisfy

$$f_c \leq f_{k+1}^{sort}$$

Table 1: Bayes ratio $\mathbb{E}[BR]$ against various measures of confidence. For CIFAR-10 we used X_1, the probability of the correct label; for CIFAR-100 and ImageNet-1K we used X_5 the probability that the correct label is in the $\mathrm{Top}5$. Data is binned into 100 bins, chosen to have equal weight.

Confidence measure	CIFAR-10	CIFAR-100	ImageNet-1K
Model Entropy	4.29	3.64	8.18
$-\log p_{\max}$	4.22	3.77	**8.87**
$-\log \sum p_{1:5}$	-	**4.25**	8.45
$\|\nabla_x \|p\|\|$	8.32	3.47	7.17
Dropout variance ($p = 0.002$)	10.39	3.11	6.84
Dropout variance ($p = 0.01$)	4.67	2.38	7.81
Dropout variance ($p = 0.05$)	1.69	1.35	1.60
Ensemble variance	**16.66**	4.03	6.13
Loss	∞	228.94	1242.55

with

$$f_{k+1}^{sort} \leq 1 - (f_1^{sort} + \cdots + f_k^{sort})$$

Thus

$$\mathcal{L}_{KL}(f, e_c) \geq -\log(1 - (f_1^{sort} + \cdots + f_k^{sort}))$$

Then, by an argument similar to the one for Top 1 error, we see that

$$\text{Prob}\left(S_k^\epsilon\right) \leq \frac{\mathbb{E}\left[X_k\right]}{\log\left(\frac{1}{\epsilon}\right)} \tag{9}$$

5 EMPIRICAL RESULTS

The previous section proved that, under fairly general conditions, we can define uncertainty measure which ensure that the top k classification is correct with high probability. The theory applies to uncertainties used in the literature, such as the negative entropy of the probabilities, and negative log softmax.

In practice, once we have an uncertainty measure, the method is simple

1. Compute the statistics on the test set of the uncertainty estimates.
2. Divide the test set into bins, based on uncertainty values.
3. Estimate the conditional probabilities based on the bin populations.

5.1 VALUE OF THE CONFIDENCE MEASURE USING THE BAYES FACTOR

The Bayes factor is a way to measure the value of new information, in terms of how much the expected winnings of a fair bet increase, when the information is available. The Bayes factor is explained in Appendix B.

In Figure 6 we plot the regularized Bayes factor for our two main measure of confidence, U_1 and U_5 along with the loss and the model Entropy. The entropy and U_1, U_5 have very large Bayes factor in the first 10 and last 3 bins, meaning that for these bins, the prediction is 10X (or more) likely to be correct (for the first 10) or wrong (for the last 3 bins) than average.

In Table 1 we show the expected Bayes factor for various confidence measures, on CIFAR-10, CIFAR-100, and ImageNet-1K. In addition to the confidence measures already discussed, we considered Bayesian dropout, and the norm of the gradient of the model. Larger expected Bayes factors means the information is more valuable.

5.2 CONFIDENCE BINS

In this section we present confidence bins for ImageNet-1K. These bins are concise summaries of the information presented in the larger bins. Table 2 presents short bins for ImageNet. Using these

Table 2: Confidence bins for ImageNet-1K. The values of a and b are chosen such that $P(\text{top5} \mid Y < a) = 0.99$ and $P(a \leq \text{top5} \mid Y < b) = 0.95$. For the model used here, $P(\text{top5}) = 0.9406$.

Confidence measure Y	(a, b)	$P(Y < a)$	$P(a \leq Y < b)$	$P(Y \geq b)$
Model Entropy	(0.31, 1.40)	0.55	0.31	0.14
$-\log p_{\text{max}}$	(0.047, 0.41)	0.52	0.26	0.22
$-\log \sum p_{1:5}$	(6.2e−3, 0.03)	0.66	0.13	0.21
$\|\nabla_x \|p\|\|$	(0.19, 0.30)	0.52	0.08	0.40
Dropout variance ($p = 0.002$)	(8.5e−4, 4.7e−3)	0.50	0.15	0.35
Ensemble variance	(0.014, 0.023)	0.54	0.05	0.41

Table 3: Discarding out-of-distribution images from ImageNet-1K. For each confidence measure Y, the value of a is chosen such that $P(Y \leq a \mid \text{image is from ImageNet-1k}) = 0.9$.

Image source	Confidence measure	a	$P(\text{image discarded})$
COCO	Model Entropy	1.75	**0.38**
	$-\log p_{\text{max}}$	0.77	0.34
	$-\log \sum p_{1:5}$	0.13	0.37
	$\|\nabla_x \|p\|\|$	1.06	0.23
	Dropout variance ($p = 0.002$)	0.024	0.
adversarially perturbed (L_2)	Model Entropy	1.75	0.28
	$-\log p_{\text{max}}$	0.77	0.25
	$-\log \sum p_{1:5}$	0.13	0.28
	$\|\nabla_x \|p\|\|$	1.06	**0.58**
	Dropout variance ($p = 0.002$)	0.024	0.39

bins, we can simply read of from the Uncertainty values, the probability that the model is correct. For example, on the model, $P(top5) = 0.9406$, however, using entropy, 55% of the images had entropy low enough to be confidently classified with probability .99. Using U_5, 66% of images could be binned to have probability .99.

Bins for CIFAR-10 and CIFAR-100 are given in Tables 7 and 6, respectively.

6 Extensions

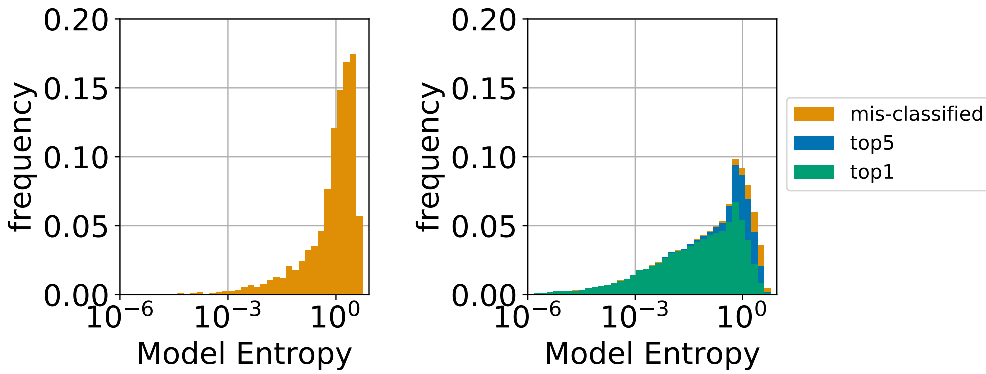
In this section we discuss some extensions of the confidence results. We show that we can detect mislabeled images in the test set. We also show that we can obtain some confidence results for off manifold images, as well as adversarial images.

6.1 Detection of mislabeled images

We are able to detect test images which are mis-labeled: images which the network correctly classified, but who's label is incorrect, or for which multiple labels could apply. These are images with high loss but low model entropy. For example in Figure 2 we show six images from the ImageNet-1k test set who's predictions where not in the top5, but had low model entropy. All six of these images either have an incorrect dataset label, or could be described by multiple labels.

6.2 Confidence on out-of-distribution and adversarial images

Next we studied whether we could detect out-of-distribution images generated by COCO. In Figure 3 we show how the histogram of the model entropy is shifted to the right compared to the on-distribution images. Table 3 give the results of our test: choosing a confidence measure which rejects 10% of the on-distribution images, our confidence measures rejected as much as 38% of COCO images (for Entropy) with similar values for U_1, U_5. On the other hand Dropout was completely ineffective.



(a) COCO images (b) ImageNet images

Figure 3: Figure 3(a): Confidence of a model trained on ImageNet-1k, evaluated on the COCO dataset. Figure 3(b): ImageNet images.

Table 4: Adversarial detection with ResNeXt-34 (2x32) on CIFAR-10. Clean images which the model correctly labels are perturbed until they are misclassified with four attack methods (PGD, Boundary attack, Carlini-Wagner, and an evasive Carlini-Wagner designed to avoid detection). Images are rejected if $|\nabla f(x)|_{2,\infty} > 2.45$.

	clean	PGD	Boundary	CW	evasive CW
percent detected	6%	96%	100%	100%	22%
median ℓ_2	-	0.31	0.36	0.34	0.81

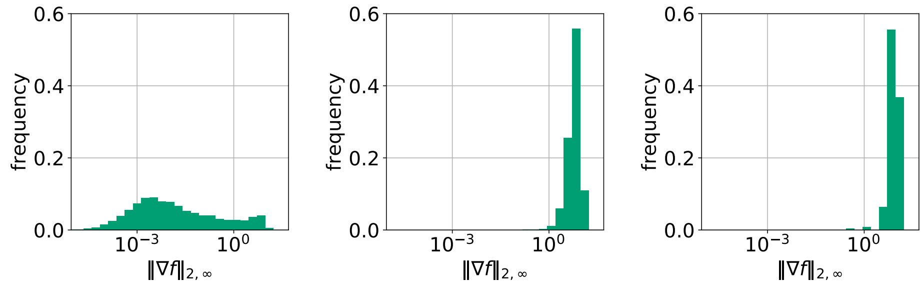
6.3 Adversarial attack detection

In this section we empirically demonstrate that image vulnerability may also be used to *detect* adversarial examples. We hypothesize that unless otherwise penalized, gradient based attacks will tend to move images to regions where the gradient of the loss is large. Based on this heuristic, we propose the norm of the loss gradient norm as criterion for detecting adversarial perturbations. Because the loss is not available during inference, we propose using the norm of the model gradient as a rejection criteria: an image has been adversarially perturbed if

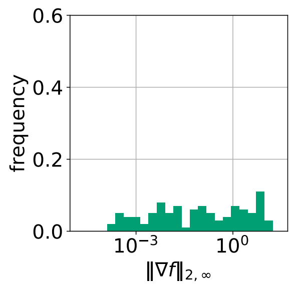
$$\|\nabla f(x)\|_{2,\infty} \geq c, \tag{10}$$

for some threshold value, c. The threshold is determined by setting the significance level (the rate of false positives) to 5%. For example on CIFAR-10 we obtained $c = 2.45$ for our model. The results are reported in Table 4 and in Figure 4. Only 6% of clean test images were rejected. However, 100% of Boundary attacks and Carlini-Wagner attacks were detected, as well as 96% of PGD attacked images.

This leads to the question, is it possible to successfully perturb all images in the test set, and avoid detection? We built a targeted attack, designed to avoid detection. We use a Carlini-Wagner style attack, modified with a penalty to avoid detection. We augmented the attack loss function with a penalty for $\|\nabla \ell(x)\|_*^2$, which penalizes attacks for being detectable. We call this attack an evasive Carlini-Wagner attack. The evasive CW attack was successful at avoiding detection 78% of the time, but in order to do so, it increased the median adversarial distance significantly, from 0.31 to 0.81, see Table 4.

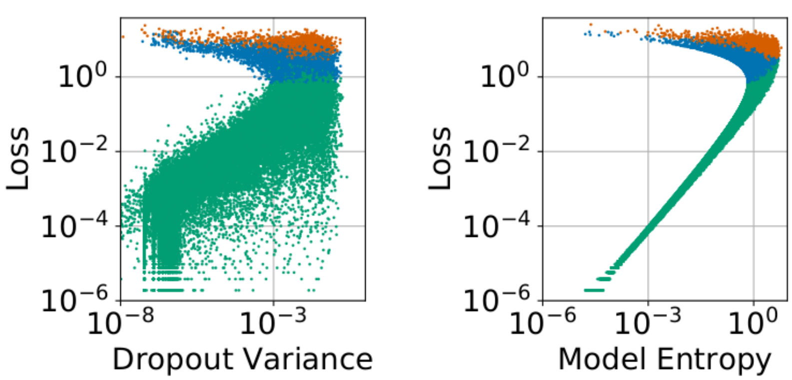


(a) Clean (b) PGD attacked (c) Boundary attack



(d) evasive CW attack

Figure 4: Frequency distribution of the norm of the model Jacobian $|\nabla f(x)|_{2,\infty}$ on ResNeXt-34 (2x32) on CIFAR-10, using 4(a): Clean, 4(b): PGD attacked 4(c): Boundary attacked, 4(d): evasive-CW attacked test images.



(a) Dropout variance (b) Model entropy

Figure 5: Illustration of uncertainty measures on ImageNet. Dropout $p = 0.002$.

REFERENCES

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul F. Christiano, John Schulman, and Dan Mané. Concrete Problems in AI Safety. *CoRR*, abs/1606.06565, 2016. URL http://arxiv.org/abs/1606.06565.

Anoop Korattikara Balan, Vivek Rathod, Kevin P. Murphy, and Max Welling. Bayesian dark knowledge. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp. 3438–3446, 2015. URL http://papers.nips.cc/paper/5965-bayesian-dark-knowledge.

Edmon Begoli, Tanmoy Bhattacharya, and Dimitri Kusnezov. The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence*, 1(1):20, January 2019. ISSN 2522-5839. doi: 10.1038/s42256-018-0004-1. URL https://www.nature.com/articles/s42256-018-0004-1.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. *CoRR*, abs/1505.05424, 2015. URL http://arxiv.org/abs/1505.05424.

Glenn W Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, January 1950. doi: 10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2.

Travers Ching, Daniel S Himmelstein, Brett K Beaulieu-Jones, Alexandr A Kalinin, Brian T Do, Gregory P Way, Enrico Ferrero, Paul-Michael Agapow, Michael Zietz, Michael M Hoffman, and others. Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface*, 15(141):20170387, 2018.

Terrance DeVries and Graham W. Taylor. Learning Confidence for Out-of-Distribution Detection in Neural Networks. *CoRR*, abs/1802.04865, 2018a. URL http://arxiv.org/abs/1802.04865.

Terrance DeVries and Graham W. Taylor. Leveraging Uncertainty Estimates for Predicting Segmentation Quality. *CoRR*, abs/1807.00502, 2018b. URL http://arxiv.org/abs/1807.00502.

Thomas G. Dietterich. Ensemble methods in machine learning. In *Multiple Classifier Systems, First International Workshop, MCS 2000, Cagliari, Italy, June 21-23, 2000, Proceedings*, pp. 1–15, 2000. doi: 10.1007/3-540-45014-9_1. URL https://doi.org/10.1007/3-540-45014-9_1.

Di Feng, Lars Rosenbaum, and Klaus Dietmayer. Towards Safe Autonomous Driving: Capture Uncertainty in the Deep Neural Network For Lidar 3d Vehicle Detection. In *21st International Conference on Intelligent Transportation Systems, ITSC 2018, Maui, HI, USA, November 4-7, 2018*, pp. 3266–3273, 2018. doi: 10.1109/ITSC.2018.8569814. URL https://doi.org/10.1109/ITSC.2018.8569814.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pp. 1050–1059, 2016. URL http://proceedings.mlr.press/v48/gal16.html.

Yonatan Geifman, Guy Uziel, and Ran El-Yaniv. Bias-Reduced Uncertainty Estimation for Deep Neural Classifiers. September 2018. URL https://openreview.net/forum?id=SJfb5jCqKm.

Irving J Good. Studies in the history of probability and statistics. XXXVII AM Turing's statistical work in world war ii. *Biometrika*, pp. 393–396, 1979.

Alex Graves. Practical variational inference for neural networks. In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain.*, pp. 2348–2356, 2011. URL http://papers.nips.cc/paper/4329-practical-variational-inference-for-neural-networks.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pp. 1321–1330, 2017. URL http://proceedings.mlr.press/v70/guo17a.html.

Leonard Hasenclever, Stefan Webb, Thibaut Lienart, Sebastian Vollmer, Balaji Lakshminarayanan, Charles Blundell, and Yee Whye Teh. Distributed bayesian learning with stochastic natural gradient expectation propagation and the posterior server. *Journal of Machine Learning Research*, 18: 106:1–106:37, 2017. URL http://jmlr.org/papers/v18/16-478.html.

Dan Hendrycks and Kevin Gimpel. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017. URL https://openreview.net/forum?id=Hkg4TI9xl.

Dan Hendrycks, Mantas Mazeika, and Thomas G. Dietterich. Deep Anomaly Detection with Outlier Exposure. *CoRR*, abs/1812.04606, 2018. URL http://arxiv.org/abs/1812.04606.

José Miguel Hernández-Lobato and Ryan P. Adams. Probabilistic backpropagation for scalable learning of bayesian neural networks. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pp. 1861–1869, 2015. URL http://proceedings.mlr.press/v37/hernandez-lobatoc15.html.

Jiri Hron, Alexander G. de G. Matthews, and Zoubin Ghahramani. Variational Bayesian dropout: pitfalls and fixes. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pp. 2024–2033, 2018. URL http://proceedings.mlr.press/v80/hron18a.html.

Harold Jeffreys. *Theory of probability, 3rd Edition*. Oxford Classic Texts in the Physical Sciences. Clarendon Press, 3 edition, 2003. ISBN 0198503687,9780198503682. URL http://gen.lib.rus.ec/book/index.php?md5=466fd89ad88ccb6e1aa6f53d937cf93e.

Heinrich Jiang, Been Kim, Melody Y. Guan, and Maya R. Gupta. To Trust Or Not To Trust A Classifier. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, pp. 5546–5557, 2018. URL http://papers.nips.cc/paper/7798-to-trust-or-not-to-trust-a-classifier.

Robert E Kass and Adrian E Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding. In *British Machine Vision Conference 2017, BMVC 2017, London, UK, September 4-7, 2017*, 2017. URL https://www.dropbox.com/s/jgozsaobbk98azy/0205.pdf?dl=1.

Durk P Kingma, Tim Salimans, and Max Welling. Variational Dropout and the Local Reparameterization Trick. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 28*, pp. 2575–2583. Curran Associates, Inc., 2015. URL http://papers.nips.cc/paper/5666-variational-dropout-and-the-local-reparameterization-trick.pdf.

Agustinus Kristiadi and Asja Fischer. Predictive uncertainty quantification with compound density networks. *CoRR*, abs/1902.01080, 2019. URL http://arxiv.org/abs/1902.01080.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pp. 6405–6416, 2017. URL http://papers.nips.cc/paper/7219-simple-and-scalable-predictive-uncertainty-estimation-using-deep-ensembles.

Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training Confidence-calibrated Classifiers for Detecting Out-of-Distribution Samples. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018. URL https://openreview.net/forum?id=ryiAv2xAZ.

Hui Li, Xuesong Wang, and Shifei Ding. Research and development of neural network ensembles: a survey. *Artif. Intell. Rev.*, 49(4):455–479, 2018. doi: 10.1007/s10462-016-9535-1. URL https://doi.org/10.1007/s10462-016-9535-1.

Yingzhen Li, José Miguel Hernández-Lobato, and Richard E. Turner. Stochastic expectation propagation. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp. 2323–2331, 2015. URL http://papers.nips.cc/paper/5760-stochastic-expectation-propagation.

Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018. URL https://openreview.net/forum?id=H1VGkIxRZ.

Si Liu, Risheek Garrepalli, Thomas G. Dietterich, Alan Fern, and Dan Hendrycks. Open Category Detection with PAC Guarantees. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pp. 3175–3184, 2018. URL http://proceedings.mlr.press/v80/liu18e.html.

Christos Louizos and Max Welling. Structured and efficient variational deep learning with matrix gaussian posteriors. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pp. 1708–1716, 2016. URL http://proceedings.mlr.press/v48/louizos16.html.

David J. C. MacKay. A practical bayesian framework for backpropagation networks. *Neural Computation*, 4(3):448–472, 1992a. doi: 10.1162/neco.1992.4.3.448. URL https://doi.org/10.1162/neco.1992.4.3.448.

David JC MacKay. *Bayesian methods for adaptive models*. PhD thesis, California Institute of Technology, 1992b.

Andrey Malinin and Mark J. F. Gales. Predictive Uncertainty Estimation via Prior Networks. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, pp. 7047–7058, 2018. URL http://papers.nips.cc/paper/7936-predictive-uncertainty-estimation-via-prior-networks.

Amit Mandelbaum and Daphna Weinshall. Distance-based confidence score for neural network classifiers. *CoRR*, abs/1709.09844, 2017. URL http://arxiv.org/abs/1709.09844.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA.*, pp. 2901–2907, 2015. URL http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9667.

Radford M Neal. Bayesian learning for neural networks. 1996.

Anh Mai Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 427–436, 2015. doi: 10.1109/CVPR.2015.7298640. URL https://doi.org/10.1109/CVPR.2015.7298640.

Khanh Nguyen and Brendan O'Connor. Posterior calibration and exploratory analysis for natural language processing models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pp. 1587–1598, 2015. URL http://aclweb.org/anthology/D/D15/D15-1182.pdf.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Machine Learning, Proceedings of the Twenty-Second International Conference (ICML 2005), Bonn, Germany, August 7-11, 2005*, pp. 625–632, 2005. doi: 10.1145/1102351.1102430. URL https://doi.org/10.1145/1102351.1102430.

Ramon Oliveira, Pedro Tabacof, and Eduardo Valle. Known Unknowns: Uncertainty Quality in Bayesian Neural Networks. *CoRR*, abs/1612.01251, 2016. URL http://arxiv.org/abs/1612.01251.

Nicolas Papernot and Patrick McDaniel. Deep k-Nearest Neighbors: Towards Confident, Interpretable and Robust Deep Learning. March 2018. URL https://arxiv.org/abs/1803.04765v1.

John Platt. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. *Adv. Large Margin Classif.*, 10, 2000.

Foster J. Provost, Tom Fawcett, and Ron Kohavi. The case against accuracy estimation for comparing induction algorithms. In *Proceedings of the Fifteenth International Conference on Machine Learning (ICML 1998), Madison, Wisconsin, USA, July 24-27, 1998*, pp. 445–453, 1998.

Charles Richter and Nicholas Roy. Safe Visual Navigation via Deep Learning and Novelty Detection. In *Robotics: Science and Systems XIII, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA, July 12-16, 2017*, 2017. doi: 10.15607/RSS.2017.XIII.064. URL http://www.roboticsproceedings.org/rss13/p64.html.

Jost Tobias Springenberg, Aaron Klein, Stefan Falkner, and Frank Hutter. Bayesian optimization with robust bayesian neural networks. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 4134–4142, 2016. URL http://papers.nips.cc/paper/6117-bayesian-optimization-with-robust-bayesian-neural-networks.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

Shengyang Sun, Changyou Chen, and Lawrence Carin. Learning structured weight uncertainty in bayesian neural networks. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*, pp. 1283–1292, 2017. URL http://proceedings.mlr.press/v54/sun17b.html.

Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pp. 681–688, 2011. URL https://icml.cc/2011/papers/398_icmlpaper.pdf.

Dong Yu, Jinyu Li, and Li Deng. Calibration of confidence measures in speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(8):2461–2473, 2011.

Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*, pp. 609–616, 2001.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July 23-26, 2002, Edmonton, Alberta, Canada*, pp. 694–699, 2002. doi: 10.1145/775047.775151. URL https://doi.org/10.1145/775047.775151.

Hugo Zaragoza and Florence d'Alché Buc. Confidence measures for neural network classifiers. In *Proceedings of the Seventh Int. Conf. Information Processing and Management of Uncertainty in Knowlegde Based Systems*, 1998.

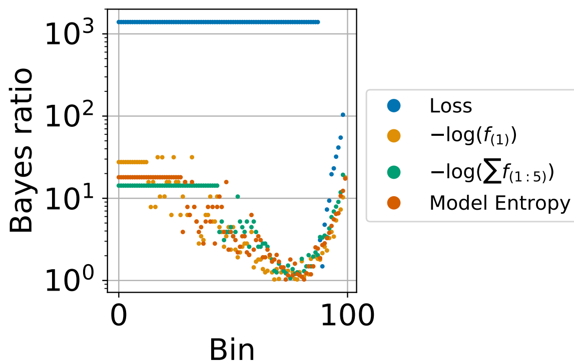


Figure 6: Bayes ratio over equal 100 quantile bins on test set for ImageNet: loss, entropy, U_1, U_5. The entropy and U_1, U_5 have very large Bayes ratio in the first 10 and last 3 bins.

A Markov's inequality

Lemma A.1 (Markov's Inequality)**.** *For a random variable Z with finite expectation, let $S \subset \{Z \geq a\}$ then*

$$\text{Prob}\,(S) \leq \frac{\mathbb{E}\,[Z]}{a} \tag{11}$$

B Measuring confidence using the expected Bayes factor

In this section we define a metric for measuring the quality of an uncertainty random variable. Suppose the random variable $U(x)$ takes values $U(x) \in [0, \infty)$. We can use the histogram of $U(x)$ to define bins where we measure the conditional probabilities.

B.1 The Bayes factor

Consider a Bernoulli random variable $X = B(p_X)$. The odds for X are given by $O(p) = \frac{p}{1-p}$. Now consider a test, $Y = B(p_Y)$, for which

$$p_{X,Y} = \text{Prob}\,(X = 1 \mid Y = 1)$$

Then the odds, given the test succeeds, are $O(p_{X,Y})$. In the odds have increased, we define the Bayes Factor to be

$$BF(X \mid Y) = \frac{O(p_{X,Y})}{O(p_X)},$$

On the other hand, if the odds have decreased, then the value of the information provided by Y is to bet against, so we define the Bayes factor to be

$$BF(X \mid Y) = \frac{O(p_X)}{O(p_{X,Y})},$$

Note that the Bayes factor of a test does not depend on the probability of success for the test. Generally, we define the Bayes factor as follows. Given $X = B(p_X)$ and the test $Y = B(p_y)$, the Bayes factor for Y is given by

$$BF(X \mid Y) = \max\left(\frac{O(p_{X,Y})}{O(p_X)}, \frac{O(p_X)}{O(p_{X,Y})}\right) \tag{12}$$

In the case where the test is certain, the Bayes factor is infinite, so we cap the odds at $\overline{T}$ for a large number $\overline{T}$

Definition B.1. Define the regularized Bayes Factor by

$$\overline{BF}(X \mid Y) = \min(\overline{T}, BF(X \mid Y)) \tag{13}$$

Example B.2. For example, if $p_X = .95$ then $O(p_X) = 19$. If $p_{X,Y} = .99$ then $O(p_{X,Y}) = 99$, and $BF(X \mid Y) = 5.25$. On the other hand, if $p_{X,Y} = 2/3$ then $O(p_{X,Y}) = 2$ and $BF(X \mid Y) = 9.5$

Table 5: Brier score of various measures of confidence. For CIFAR-10 we used X_1, the probability of the correct label; for CIFAR-100 and ImageNet-1K we used X_5 the probability that the correct label is in the Top5. Data is binned into 100 bins, chosen to have equal weight.

Confidence measure	CIFAR-10	CIFAR-100	ImageNet-1K
Model Entropy	**0.033**	0.067	0.041
$-\log p_{\max}$	**0.033**	0.067	0.042
$-\log \sum p_{1:5}$	-	0.067	**0.040**
$\|\nabla_x \|p\|\|$	0.034	0.073	0.046
Dropout variance ($p = 0.002$)	0.036	0.074	0.047
Dropout variance ($p = 0.01$)	0.04	0.075	0.048
Dropout variance ($p = 0.05$)	0.043	0.076	0.049
Ensemble variance	0.040	**0.050**	0.047
Loss	0	0.029	0.019

Table 6: Confidence bins for CIFAR-100. The values of a and b are chosen such that $P(\text{top5} \mid Y < a) = 0.99$ and $P(a \leq \text{top5} \mid Y < b) = 0.95$. For the model used here, $P(\text{top5}) = 0.916$.

Confidence measure Y	(a, b)	$P(Y < a)$	$P(a \leq Y < b)$	$P(Y \geq b)$
Model Entropy	(0.082, 2.1)	0.24	0.50	0.26
$-\log p_{\max}$	(7.9e−3, 0.42)	0.24	0.49	0.27
$-\log \sum p_{1:5}$	(4.8e−3, 0.34)	0.19	0.57	0.24
$\|\nabla_x \|p\|\|$	(0.46, 1.70)	0.27	0.17	0.56
Dropout variance ($p = 0.002$)	(6.4e−4, 2.2e−3)	0.27	0.06	0.67
Ensemble variance	(4.2e−4, 0.052)	0.42	0.18	0.40

B.2 Expected Bayes factor

Definition B.3 (Histogram random variables)**.** Next, given a random variable $U(x) \in [a, b)$ and a partition of $[a, b]$ into bins

$$a = t_0 < t_1 \cdots < t_Q = b, \tag{14}$$

Define the (histogram) random variables Y_i corresponding to each interval

$$Y_i(x) = \begin{cases} 1 & t_{i-1} \leq U(x) < t_i \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

so that

$$\text{Prob}\left(t_{i-1} \leq U < t_i\right) = \mathbb{E}\left[Y_i\right] \tag{16}$$

Each Bayes factor measures the value of information that x lies in each quantile. The value of the test itself is defined to be the expected value of the Bayes factors.

Definition B.4 (Histogram Bayes Factors)**.** Given X, U and the histogram random variables Y_i, define the conditional probabilities

$$p_{X,i} = \text{Prob}\left(X = 1 \mid Y_i = 1\right), \quad i = 1, \ldots, Q \tag{17}$$

Write $\overline{BF}(X \mid Y_i)$ for the regularized Bayes Factor of each Y_i, given by (13). The predictive value for X of the random variable U with respect to the histogram, is given by

$$\mathbb{E}\left[\overline{BF}(X \mid Y_i)\right] = \sum_{i=1}^{Q} \overline{BF}(X \mid Y_i)\, \mathbb{E}\left[Y_i\right] \tag{18}$$

B.3 Worked example of Bayes Factors

Consider the situation where you have exchanged phone numbers with someone, and you wish to contact them. The question is whether to send a text message or phone their number. Approximately

Table 7: Confidence bins for CIFAR-10. The value of a is chosen such that $P(\text{top1} \mid Y < a) = 0.975$.

Confidence measure Y	a	$P(Y < a)$	$P(Y \geq a)$
Model Entropy	1.6	0.95	0.05
$-\log p_{\max}$	0.57	0.95	0.05
$\lVert\nabla_x \lVert p\rVert\rVert$	8.16	0.93	0.07
Dropout variance ($p = 0.002$)	0.045	0.92	0.08
Ensemble variance	0.019	0.88	0.12

95% of people prefer to message. Let X be the probability that a person prefers to message. The expected value and odds for X is given by

$$p_X = 0.95, \qquad O(p_X) = 19$$

Now suppose we have additional information, which gives these statistics based on age. Suppose we wish to predict X. Knowing the age U has a value. Let $U(x)$ be the age, and consider three bins for U given by the values $20, 65$ and let Y_1, Y_2, Y_3 be the corresponding histogram random variables.

$$\begin{cases} Y_1 = 1_{\{U<20\}}, & \mathbb{E}[Y_1] = .4 \\ Y_2 = 1_{\{20 \leq U \leq 65\}}, & \mathbb{E}[Y_2] = .5 \\ Y_3 = 1_{\{65<U\}}, & \mathbb{E}[Y_2] = .1 \end{cases} \tag{19}$$

Since older people are more likely to prefer to use a phone, the conditional probabilities and corresponding odds are given by

$$\begin{cases} p(X \mid Y_1) = .999, & O(p_{X,Y_1}) = 999 \\ p(X \mid Y_2) = .94, & O(p_{X,Y_2}) = 15.7 \\ p(X \mid Y_3) = .9, & O(p_{X,Y_2}) = 9 \end{cases} \tag{20}$$

In particular, knowing if they are younger or older is more valuable than the middle range. The Bayes ratio (relative odds) expresses the value of knowing the age if someone is willing to bet with the odds $O(p_X)$. So this information allows an expected profit on the bet given by the ratio.

$$\begin{cases} BF(X \mid Y_1) = 999/19 = 53 \\ BF(X \mid Y_1) = 19/15.7 = 1.2 \\ BF(X \mid Y_1) = 19/9 = 2.1 \end{cases} \tag{21}$$

So the value of the information depends on the cases. Finally, if we wish to find the expected value of the information, we take an expectation with respect to the probabilities of the events.

$$\mathbb{E}[BR(X|Y_i)] = 53 \times .4 + 1.2 \times .5 + 2.1 \times .1 = 22 \tag{22}$$

Some other information about the person may be much less useful in prediction their preference. For example, suppose you know the region where they live and let Y_1, Y_2, Y_3 be the histogram random variables. Suppose

$$\begin{cases} p(X \mid Y_1) = .03 \\ p(X \mid Y_2) = .05 \\ p(X \mid Y_3) = .07 \end{cases} \qquad \begin{cases} \mathbb{E}[Y_1] = .3 \\ \mathbb{E}[Y_2] = .5 \\ \mathbb{E}[Y_3] = .3 \end{cases} \tag{23}$$

Since $\mathbb{E}[X] = .95$,

$$\begin{cases} BF(X \mid Y_1) = 1.9 \\ BF(X \mid Y_1) = 1.1 \\ BF(X \mid Y_1) = 1.3 \end{cases} \qquad \mathbb{E}[BF(X|Y_i)] = 1.5 \tag{24}$$

So with an expected value of 1.5, compared to age, with an expected value of 22, the location information is much less valuable.