Model averages sharpened into Occam's razors: Deep learning enhanced by Rényi entropy

David R. Bickel

Model averages sharpened into Occam's razors: Deep learning enhanced by Rényi entropy

December 6, 2019

David R. Bickel

Ottawa Institute of Systems Biology

Department of Biochemistry, Microbiology and Immunology

Department of Mathematics and Statistics

University of Ottawa

451 Smyth Road

Ottawa, Ontario, K1H 8M5

+01 (613) 562-5800, ext. 8670

dbickel@uottawa.ca

Abstract

Ensemble methods of machine learning combine neural networks or other machine learning models in order to improve predictive performance. The proposed ensemble method is based on Occam's razor idealized as adjusting hyperprior distributions over models according to a Rényi entropy of the data distribution that corresponds to each model.

The entropy-based method is used to average a logistic regression model, a random forest, and a deep neural network. As expected, the deep leaning machine more accurately recognizes handwritten digits than the other two models. The combination of the three models performs even better than the neural network when they are combined according to the entropy-based method or according to methods that average the log odds of the classification probabilities reported by the models.

Which of the best ensemble methods to choose for other applications may depend on the loss function that quantifies prediction performance and on a robustness consideration.

Keywords: big data; data science; deep learning; deep neural network; model averaging; model predictive distribution; predictivism; sharpened prior distribution

1 Introduction

In traditional statistics, there are often multiple methods that can legitimately be used to analyze the same data. Common choices available between very different methods include conditioning on one of multiple ancillary statistics for conditional inference (Fraser, 2004); selecting one of many appropriate priors for Bayesian statistics (Kass and Wasserman, 1996); choosing between frequentist and Bayesian methods; deciding whether to use a parametric test to analyze a sample of data too small to check the model assumptions, whether and how to adjust for testing multiple hypotheses, and whether to use a model averaging procedure. While each method may be considered appropriate, the results can differ markedly.

The same problems occur in machine learning given uncertainty about the neural network or other algorithmic model used to make predictions such as classifications. Decisions often require timely advice even when different models would give different advice and when the data do not decisively favor any model above the others. That type of uncertainty occurs with big data sets that lack the type of information needed to effectively discriminate between models.

In that setting, some form of model averaging would be ideal since that would strike a balance between two extremes. At one extreme, the user would be presented with the contradictory advice of many models, creating confusion and either paralysis or making intuitive decisions not fully based on the data. At the other extreme, the uncertainty in the models would be suppressed, and only the advice of what appears to be the best model would be reported to the user. The danger is that insights from the other models would be neglected. Model averaging takes into account the evidence-based insights of all available models without overloading the user with information.

Bayesian model averaging has often performed well when a suitable prior distribution over the models is available. It is applicable to deep neural networks and other non-Bayesian models that yield predictive distributions, including those reporting classification probabilities (Eklund and Karlsson, 2007; Bickel, 2019e). The risk is that Bayesian model averaging is no more reliable than the reliability of the prior distribution over the models.

Occam's razor addresses that problem by adjusting the prior distribution over models for the complexity of their predictive distributions, as described in Section 2. Alternative solutions featuring previous methods of model averaging, also called *ensemble methods* (Zhou, 2012), appear in Section 3. The complexity-based methods are then compared to many of the other ensemble methods by using them to average a deep neural network, a random forest, and a regularized logistic regression model for the analysis of a large handwriting recognition data set in Section 4.

Preliminary conclusions appear in Section 5.

2 Ensemble learning given the complexity of predictive distributions

To lay the foundations for combining ensembles of non-Bayesian classifiers and other prediction algorithms, Section 2.1 applies priors adjusted for the complexity of sampling distributions to Bayesian model averaging. Section 2.2 then handles the case in which predictive distributions are available instead of sampling distributions. The results are extended to non-Bayesian classifiers and other prediction algorithms in Section 2.3.

2.1 Bayesian model averaging given the complexity of sampling distributions

Let n denote the number of observations of the following structure. For observation index $i = 1$ to n, the observation is (x_i, y_i), where x_i is a vector of m independent variables and y_i the corresponding dependent variable. Each *Bayesian model* mdl is defined in terms of the following probability distributions. The dependent variable is assumed to be a random variate generated by a sampling distribution $f_{\mathrm{mdl}}(y_i \,|\, x_i, \theta_{\mathrm{mdl}})$ given a parameter a parameter θ_{mdl} in some set Θ_{mdl}. Each Bayesian model also has a prior distribution π_{mdl}.

The sample is (x, y), where $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$. It determines the posterior distribution $\pi_{\mathrm{mdl}}(\theta_{\mathrm{mdl}} | (x, y))$ according to Bayes's theorem. For $t = 1, 2, \ldots$, each predicted value $\widehat{y}_{n+t}$ of y_{n+t} given x_{n+t} has a posterior predictive distribution $f^{\mathrm{prdct}}_{\mathrm{mdl}}\left(\widehat{y}_{n+t} | (x, y), x_{n+t}\right)$ defined as the sampling distribution $f_{\mathrm{mdl}}\left(\widehat{y}_{n+t} | x_{n+t}, \theta_{\mathrm{mdl}}\right)$ with θ_{mdl} eliminated by integration with respect to $\pi_{\mathrm{mdl}}(\theta_{\mathrm{mdl}} | (x, y))$. For example, if θ_{mdl} is continuous and $\pi_{\mathrm{mdl}}(\theta_{\mathrm{mdl}})$ is a probability density, then

$$\pi_{\mathrm{mdl}}(\theta_{\mathrm{mdl}} | (x, y)) \propto \pi_{\mathrm{mdl}}(\theta_{\mathrm{mdl}}) f_{\mathrm{mdl}}(y \,|\, x, \theta_{\mathrm{mdl}});$$

$$f^{\mathrm{prdct}}_{\mathrm{mdl}}\left(\widehat{y}_{n+t} | (x, y), x_{n+t}\right) = \int f_{\mathrm{mdl}}\left(\widehat{y}_{n+t} | x_{n+t}, \theta_{\mathrm{mdl}}\right) \pi_{\mathrm{mdl}}(\theta_{\mathrm{mdl}} | (x, y)) \, d\theta_{\mathrm{mdl}}.$$

Supposing the number of Bayesian models under consideration to be finite, let $P(\mathrm{mdl})$ denote the prior probability and $P(\mathrm{mdl} | (x, y))$ the posterior probability of model mdl. *Bayesian model*

averaging means using

$$f^{\text{prdct}}\left(\widehat{y}_{n+t}|\,(x,y)\,,x_{n+t}\right)\propto\sum_{\text{mdl}}P\left(\text{mdl}\,|\,(x,y)\right)f_{\text{mdl}}^{\text{prdct}}\left(\widehat{y}_{n+t}|\,(x,y)\,,x_{n+t}\right)$$

as the posterior predictive distribution distribution of $\widehat{y}_{n+t}$.

Consider a family of Bayesian models that are degenerate in the sense that each π_{mdl} assigns 100% probability to a different parameter value θ_{mdl} in a common set Θ with a sampling distribution $f\left(y_i\,|\,x_i,\theta_{\text{mdl}}\right)$ such that $\Theta_{\text{mdl}}=\Theta$ and $f_{\text{mdl}}=f$ for every mdl. In that case, each model's predictive distribution $f_{\text{mdl}}^{\text{prdct}}\left(\widehat{y}_{n+t}|\,(x,y)\,,x_{n+t}\right)$ is the sampling distribution $f\left(\widehat{y}_{n+t}|\,x_{n+t},\theta_{\text{mdl}}\right)$ with probability 1, and the overall posterior predictive distribution is

$$f^{\text{prdct}}\left(\widehat{y}_{n+t}|\,(x,y)\,,x_{n+t}\right)\propto\sum_{\text{mdl}}P\left(\text{mdl}\,|\,(x,y)\right)f\left(\widehat{y}_{n+t}|\,x_{n+t},\theta_{\text{mdl}}\right).$$

That one-to-one correspondence between Bayesian models and parameter values presents the opportunity to modify the prior and posterior distributions according to the formalization of Occam's razor found in Bickel (2019b) and Bickel (2019a). The idea is to incorporate the complexity of the sampling distribution into the prior distribution, thereby sharpening the prior and the resulting posterior into "razors." For any *order* $\alpha\geq0$ and *sharpness* $\kappa\in[-\infty,\infty]$, the (α,κ)-*sharpened prior distribution* and (α,κ)-*sharpened posterior distribution* are given by

$$P^{(\alpha,\pm\infty)}\left(\text{mdl}\,|\,x_{n+t}\right)=\lim_{\kappa\to\pm\infty}P^{(\alpha,\kappa)}\left(\text{mdl}\,|\,x_{n+t}\right),$$

$$P^{(\alpha,\pm\infty)}\left(\text{mdl}\,|\,(x,y)\,,x_{n+t}\right)=\lim_{\kappa\to\pm\infty}P^{(\alpha,\kappa)}\left(\text{mdl}\,|\,(x,y)\,,x_{n+t}\right),\tag{1}$$

and, for any real κ, by

$$P^{(\alpha,\kappa)}\left(\text{mdl}\,|\,x_{n+t}\right)\propto P\left(\text{mdl}\right)e^{-\kappa\,S_\alpha\left(\theta_{\text{mdl}}\right)};$$

$$P^{(\alpha,\kappa)}\left(\text{mdl}\,|\,(x,y)\,,x_{n+t}\right)\propto P\left(\text{mdl}\,|\,(x,y)\right)e^{-\kappa\,S_\alpha\left(\theta_{\text{mdl}}\right)},\tag{2}$$

where $S_\alpha\left(\theta_{\text{mdl}}\right)$ is the order-α Rényi entropy of $f\left(\widehat{y}_{n+t}|\,x_{n+t},\theta_{\text{mdl}}\right)$ and, as such, depends on the suppressed independent variable x_{n+t}. (The definition of Rényi entropy (Rényi, 1965) is postponed until Section 2.2.) The corresponding (α,κ)-*sharpened posterior predictive distribution* is

$$f^{(\alpha,\kappa)}\left(\widehat{y}_{n+t}|\,(x,y)\,,x_{n+t}\right)=\sum_{\text{mdl}}P^{(\alpha,\kappa)}\left(\text{mdl}\,|\,(x,y)\,,x_{n+t}\right)f\left(\widehat{y}_{n+t}|\,x_{n+t},\theta_{\text{mdl}}\right).\tag{3}$$

2.2 Bayesian model averaging given the complexity of predictive distributions

Equation (3) holds exactly only in the degenerate case that each model corresponds to a single sampling distribution. An approximation to it also holds more generally when each model's predictive distribution approaches its sampling distribution.

Consider the predictive distribution $f_{\text{mdl}}^{\text{prdct}}\left(\widehat{y}_{n+t}|\,(x,y)\,,x_{n+t}\right)$ as an estimate of the sampling distribution $f_{\text{mdl}}\left(\widehat{y}_{n+t}|\,x_{n+t},\theta_{\text{mdl}}\right)$ for each mdl and the order-α Rényi entropy of $f_{\text{mdl}}^{\text{prdct}}\left(\widehat{y}_{n+t}|\,(x,y)\,,x_{n+t}\right)$ as an estimate of the order-α Rényi entropy of $f_{\text{mdl}}\left(\widehat{y}_{n+t}|\,x_{n+t},\theta_{\text{mdl}}\right)$, where θ_{mdl} is the true value of the parameter under model mdl. The latter estimate is denoted by $\widehat{S}_{\alpha}\left(\text{mdl}\right)$. (That follows Geisser (1971)'s *predictivism*, using predictive distributions and their functionals as estimates of sampling distributions and their functionals.) For example, if $f_{\text{mdl}}^{\text{prdct}}\left(\widehat{y}_{n+t}|\,(x,y)\,,x_{n+t}\right)$ specifies the classification probabilities $p_{\text{mdl}}\left(1\right),\ldots,p_{\text{mdl}}\left(K\right)$ as the predictive probabilities of each of K possible categories according to mdl, then, by the definition of Rényi entropy (Rényi, 1965),

$$\widehat{S}_{\alpha}\left(\text{mdl}\right)=\begin{cases}-\sum_{k=1}^{K}p_{\text{mdl}}\left(k\right)\log p_{\text{mdl}}\left(k\right) & \text{if }\alpha=1\\ -\log\left(\sum_{k=1}^{K}p_{\text{mdl}}\left(k\right)\left(p_{\text{mdl}}\left(k\right)\right)^{\alpha-1}\right)^{\frac{1}{\alpha-1}} & \text{if }\alpha\neq1\end{cases},$$

where each $p_{\text{mdl}}\left(k\right)$ depends not only on k and mdl but also on x, y, and x_{n+t}. Plugging the estimates into equations (1), (2), and (3) yields the *estimated* (α,κ)*-sharpened posterior distribution*

$$\widehat{P}^{(\alpha,\kappa)}\left(\text{mdl}\,|\,(x,y)\,,x_{n+t}\right)\propto\begin{cases}\lim_{\kappa\to\pm\infty}\widehat{P}^{(\alpha,\kappa)}\left(\text{mdl}\,|\,(x,y)\,,x_{n+t}\right) & \text{if }\kappa=\pm\infty\\ P\left(\text{mdl}\,|\,(x,y)\right)e^{-\kappa\widehat{S}_{\alpha}(\text{mdl})} & \text{if }-\infty<\kappa<\infty\end{cases}\tag{4}$$

and the *estimated* (α,κ)*-sharpened posterior predictive distribution*,

$$\widehat{f}^{(\alpha,\kappa)}\left(\widehat{y}_{n+t}|\,(x,y)\,,x_{n+t}\right)=\sum_{\text{mdl}}\widehat{P}^{(\alpha,\kappa)}\left(\text{mdl}\,|\,(x,y)\,,x_{n+t}\right)f_{\text{mdl}}^{\text{prdct}}\left(\widehat{y}_{n+t}|\,(x,y)\,,x_{n+t}\right).\tag{5}$$

2.3 Averages of non-Bayesian models given the complexity of predictive distributions

This section extends the framework of Section 2.2 beyond Bayesian models to more general *prediction models*, which are mathematical algorithms that transform samples into predictive distributions (Bickel, 2019e, following Breiman, 2001). In the case that y is a categorical variable, each prediction

model is a *classification model*.

With $f_{\text{mdl}}^{\text{prdct}}\left(\widehat{y}_{n+t}|\left(x,y\right),x_{n+t}\right)$ as the *model predictive distribution* (Bickel, 2019e), the predictive distribution from a prediction model and from the sample (x,y), and with $P\left(\text{mdl}\,|\,(x,y)\right)$ as the posterior probability according to Eklund and Karlsson (2007)'s method of applying Bayesian model averaging to non-Bayesian models, equations (4)-(5) still hold. However, since $P\left(\text{mdl}\,|\,(x,y)\right)$ may not be available without assuming the observations in the (x,y) are independent and since that assumption can lead to inaccurate predictions, it is often safer to average the models with respect to the prior probability $P\left(\text{mdl}\right)$ in place of the posterior probability $P\left(\text{mdl}\,|\,(x,y)\right)$, as if the likelihood function were constant (Kittler et al., 1998). In that case, the estimates for equation (5) are given by

$$\widehat{P}^{(\alpha,\kappa)}\left(\text{mdl}\,|\,(x,y),x_{n+t}\right)\propto P\left(\text{mdl}\right)e^{-\kappa\widehat{S}_{\alpha}(\text{mdl})}\text{ for all}-\infty<\kappa<\infty. \tag{6}$$

Assuming $P\left(\text{mdl}\right)>0$ for each mdl, the $\kappa=\pm\infty$ part of equation (4) provides the model predictive distributions that minimize or maximize the Rényi entropy:

$$\widehat{f}^{(\alpha,\pm\infty)}\left(\widehat{y}_{n+t}|\left(x,y\right),x_{n+t}\right)=f_{\text{mdl}(\pm\infty)}^{\text{prdct}}\left(\widehat{y}_{n+t}|\left(x,y\right),x_{n+t}\right),$$

where $\text{mdl}\left(+\infty\right)=\arg\min_{\text{mdl}}\widehat{S}_{\alpha}\left(\text{mdl}\right)$ and $\text{mdl}\left(-\infty\right)=\arg\max_{\text{mdl}}\widehat{S}_{\alpha}\left(\text{mdl}\right)$.

In the case of classification into $y\left(1\right),\ldots,y\left(K\right)$ as the possible categories, $\widehat{f}^{(\alpha,\kappa)}\left(\widehat{y}_{n+t}|\left(x,y\right),x_{n+t}\right)$ specifies the predictive probability of category $y\left(k\right)$ for $k=1,\ldots,K$. That classification probability of $y\left(k\right)$ is called the *estimated* (α,κ)*-sharpened probability* and is denoted by $\widehat{p}^{(\alpha,\kappa)}\left(k\right)$, which depends on x, y, and x_{n+t} as well as on α, κ, and k.

3 Other methods of ensemble learning

In the case of classifying the dependent variable y_{n+t} into one of K categories $y\left(1\right),\ldots,y\left(K\right)$, the model predictive distribution can be written as the vector

$$f_{\text{mdl}(\pm\infty)}^{\text{prdct}}\left(\bullet|\left(x,y\right),x_{n+t}\right)=p_{\text{mdl}}=\left(p_{\text{mdl}}\left(1\right),\ldots,p_{\text{mdl}}\left(K\right)\right),$$

where $p_{\text{mdl}}\left(k\right)$ is the probability that $y_{n+t}=y\left(k\right)$ according to mdl for $k=1,\ldots,K$. The dependence of each $p_{\text{mdl}}\left(k\right)$ on x, y, and x_{n+t} is suppressed to keep the notation concise. The methods of this section can be extended to more general predictive distributions by replacing sums

with integrals.

Let $\mathcal{M}$ denote the set of prediction models and $|\mathcal{M}|$ is the number of prediction models. The $|\mathcal{M}|$ probabilities corresponding to category $y\left(k\right)$ are encoded in the vector $p\left(k\right)$ of dimension $|\mathcal{M}|$. For example, if the models are **logistic** (a logistic regression model) and **forest** (a random forest model), then $\mathcal{M} = \{\textbf{logistic}, \textbf{forest}\}$ and $p\left(k\right) = \left(p_{\textbf{logistic}}\left(k\right), p_{\textbf{forest}}\left(k\right)\right)$.

3.1 Ensemble learning by central tendency

3.1.1 Central tendencies of probabilities

One of the simplest ensemble methods uses the *arithmetic mean*

$$\text{mean}\left(p\left(k\right)\right) = \frac{\sum_{\text{mdl}\in\mathcal{M}} p_{\text{mdl}}\left(k\right)}{|\mathcal{M}|}$$

as the probability that $y_{n+t} = y\left(k\right)$ (Kittler et al., 1998). The *normalized geometric mean* is

$$\text{gmean}_{\text{norm}}\left(p\left(k\right)\right) \propto \left(\prod_{\text{mdl}\in\mathcal{M}} p_{\text{mdl}}\left(k\right)\right)^{\frac{1}{|\mathcal{M}|}},$$

where the constant of proportionality normalizes each geometric mean to ensure that $\sum_{k=1}^{K} \text{gmean}_{\text{norm}}\left(p\left(k\right)\right) = 1$ (Kittler et al., 1998). Using only the highest and lowest probabilities at each category, the *normalized extreme geometric mean* is

$$\text{gmean}_{\text{norm}}\left(\text{ext}\, p\left(k\right)\right) \propto \sqrt{\left(\min_{\text{mdl}\in\mathcal{M}} p_{\text{mdl}}\left(k\right)\right)\left(\max_{\text{mdl}\in\mathcal{M}} p_{\text{mdl}}\left(k\right)\right)},$$

where $\text{ext}\, p\left(k\right)$ is the pair $\left(\min_{\text{mdl}\in\mathcal{M}} p_{\text{mdl}}\left(k\right), \max_{\text{mdl}\in\mathcal{M}} p_{\text{mdl}}\left(k\right)\right)$. It, too, is normalized: $\sum_{k=1}^{K} \text{gmean}_{\text{norm}}\left(\text{ext}\, p\left(k\right)\right) = 1$.

3.1.2 Central tendencies of log odds

Many measures of central tendency tend to lack sensitivity when applied directly to probabilities close to 0 or 1. That is often addressed by transforming the probabilities to numbers between $-\infty$ and ∞. For example, the log odds of $p_{\text{mdl}}\left(k\right)$ is

$$\text{logit}\, p_{\text{mdl}}\left(k\right) = \log \frac{p_{\text{mdl}}\left(k\right)}{1 - p_{\text{mdl}}\left(k\right)}$$

for each $k = 1, \ldots, K$. The vector of the $|\mathcal{M}|$ log odds corresponding to category $y(k)$ is denoted by $\operatorname{logit} p(k)$. Then $C(\operatorname{logit} p(k))$ is a central tendency of $\operatorname{logit} p(k)$, where C is a measure of central tendency such as the mean, $\text{gmean}_{\text{norm}}$, or $\text{gmean}^{\text{ext}}_{\text{norm}}$ of Section 3.1.1. In the $\mathcal{M} = \{\textbf{logistic}, \textbf{forest}\}$ example with $C = \text{gmean}_{\text{norm}}$, that would be

$$C(\operatorname{logit} p(k)) = \text{gmean}_{\text{norm}}\left(\left(\operatorname{logit} p_{\textbf{logistic}}(k), \operatorname{logit} p_{\textbf{forest}}(k)\right)\right).$$

To obtain an ensemble predictive distribution, $C(\operatorname{logit} p(k))$ needs to be transformed to a probability by the *C-probability*

$$p^{C}(\operatorname{logit} p(k)) = \operatorname{logit}^{-1} C(\operatorname{logit} p(k)) = \left(1 + e^{-C(\operatorname{logit} p(k))}\right)^{-1}.$$

The *normalized C-probability* is $p^{C}_{\text{norm}}(\operatorname{logit} p(k)) \propto p^{C}(\operatorname{logit} p(k))$ with the constant of proportionality such that $\sum_{k=1}^{K} p^{C}_{\text{norm}}(\operatorname{logit} p(k)) = 1$.

For example, $p^{\text{median}}_{\text{norm}}(\operatorname{logit} p(k))$ is the normalized median-probability for category $y(k)$. We could also consider a normalized mode-probability $p^{\text{mode}}_{\text{norm}}(\operatorname{logit} p(k))$, where mode is the half-sample mode, an estimator of the mode of a unimodal distribution (Bickel and Frühwirth, 2006). The normalized mean-probability $p^{\text{mean}}_{\text{norm}}(\operatorname{logit} p(k))$ is called the *natural odds-based probability* since $\text{mean}(\operatorname{logit} p(k))$ is a monotonic function of the geometric mean of the odds $p(k)/(1-p(k))$. Similarly, Alshemali et al. (2020) took a mean of real-valued classifier latent variable before the transforming it to classification probabilities by the softmax function, which is an alternative to $\operatorname{logit}^{-1}$ that is commonly used in deep neural networks.

The *extreme C-probability* and the *normalized extreme C-probability* are $p^{C}(\operatorname{logit}\operatorname{ext} p(k))$ and $p^{C}_{\text{norm}}(\operatorname{logit}\operatorname{ext} p(k))$. Taking $C = \text{mean}$ would use the geometric mean of the most extreme odds, which corresponds to what Bickel (2019d, Corollary 2) derived as "the strength of inferential evidence." Accordingly, $p^{\text{mean}}(\operatorname{logit}\operatorname{ext} p(k))$ and $p^{\text{mean}}_{\text{norm}}(\operatorname{logit}\operatorname{ext} p(k))$ are called the *unnormalized inferential probability* and the *normalized inferential probability*.

3.2 Ensemble learning by optimization

The methods of this section determine the classification probabilities by selecting one of the prediction models. Since the model is selected on the basis of x_{n+t} separately for each $t = 1, 2, \ldots$, these procedures are considered ensemble methods rather than methods of model selection.

3.2.1 Most confident and least confident classifiers

Given x, y, and x_{n+t}, the *most confident classifier* (Dam et al., 2006) in $\mathcal{M}$ is

$$\overline{\mathrm{mdl}} = \arg \max_{\mathrm{mdl} \in \mathcal{M}} \max_{k=1,\ldots,K} p_{\mathrm{mdl}}(k),$$

whereas the *least confident classifier* in $\mathcal{M}$ is

$$\underline{\mathrm{mdl}} = \arg \min_{\mathrm{mdl} \in \mathcal{M}} \max_{k=1,\ldots,K} p_{\mathrm{mdl}}(k).$$

Their corresponding predictive probabilities of the K categories are denoted by $p_{\overline{\mathrm{mdl}}}(k)$ and $p_{\underline{\mathrm{mdl}}}(k)$ for $k = 1, \ldots, K$.

3.2.2 Hurwicz criterion

We are again given x, y, and x_{n+t}. Consider a loss function $\ell\left(y_{n+t}, \widehat{\mathrm{mdl}}\right)$, where $\widehat{\mathrm{mdl}} \in \mathcal{M}$ is a prediction model chosen as an action. For example, the *log loss function* is

$$\ell_{\log}\left(y_{n+t}, \widehat{\mathrm{mdl}}\right) = -\log p_{\widehat{\mathrm{mdl}}}(i_{n+t}),$$

where i_{n+t} is the category index such that $y(i_{n+t}) = y_{n+t}$, and the *Brier loss function* is

$$\ell_{\mathrm{Brier}}\left(y_{n+t}, \widehat{\mathrm{mdl}}\right) = \sum_{k=1}^{K} \left(p_{\widehat{\mathrm{mdl}}}(k) - \chi(k)\right)^2,$$

where $\chi(k) = 1$ if $y(k) = y_{n+t}$ but $\chi(k) = 0$ if $y(k) \neq y_{n+t}$. If y_{n+t} were generated from a target prediction model $\mathrm{mdl} \in \mathcal{M}$, then the expected loss would be

$$\mathrm{E}_{\mathrm{mdl}}\, \ell\left(\bullet, \widehat{\mathrm{mdl}}\right) = \sum_{k=1}^{K} p_{\mathrm{mdl}}(k)\, \ell\left(y(k), \widehat{\mathrm{mdl}}\right).$$

According to the Hurwicz criterion (Hurwicz, 1951), the optimal prediction model minimizes a linear combination of worst-case and best-case expected losses:

$$\widehat{\mathrm{mdl}}_{\ell}(c) = \arg \inf_{\widehat{\mathrm{mdl}} \in \mathcal{M}} \left(c \sup_{\mathrm{mdl} \in \mathcal{M}} \mathrm{E}_{\mathrm{mdl}}\, \ell\left(\bullet, \widehat{\mathrm{mdl}}\right) + (1-c) \inf_{\mathrm{mdl} \in \mathcal{M}} \mathrm{E}_{\mathrm{mdl}}\, \ell\left(\bullet, \widehat{\mathrm{mdl}}\right) \right),$$

where c is a caution parameter between 0 and 1 and ℓ is a loss function such as $\ell_{\log}$ or ℓ_{Brier}. Whereas $\widehat{\mathrm{mdl}}_{\ell}(1)$ minimizes the maximum expected loss, $\widehat{\mathrm{mdl}}_{\ell}(0)$ minimizes the minimum expected

loss. Although $\widehat{\text{mdl}}_\ell(0)$ is widely regarded as too optimistic, it has been derived under a framework of idempotent probability (Bickel, 2019c, App. A).

The optimal predictive probabilities of the K categories are those of the optimal prediction model: $p_{\widehat{\text{mdl}}_\ell(c)}(k)$ for $k = 1, \ldots, K$. As usual, the dependence on x, y, and x_{n+t} is implicit.

4 Application to deep learning

4.1 The handwriting recognition data

In *unsupervised learning*, algorithms provide information on the structure of data without access to observations of any dependent variable. The classic example of unsupervised learning is cluster analysis.

In *supervised learning*, prediction models such as neural networks *learn* by adapting weights or other parameter values to measurements called *training data*, called the "sample" in Section 2.1. The performance of the algorithms is assessed on the basis of other measurements, called *non-training data*. The reason the training data are separated from the other data is that algorithms that *overfit* accurately classify and predict with respect to the training data but not with respect to other data.

Choosing the prediction model or ensemble of such models that performs best on non-training data can overfit the non-training data in the same way that an individual model can overfit the training data. For that reason, a non-training set is often divided into a *validation data* set, which is used in method selection, and a *test data* set, which is instead used to evaluate the performance of the chosen method. Overfitting due to method selection could otherwise be problematic when optimizing the performance over many ensemble methods, as when fitting a continuous hyperparameter, each value of which corresponds to a different ensemble method. That hyperparameter is (α, κ) in the case of the sharpening methods of Section 2.

"THE MNIST DATABASE of handwritten digits" (`http://yann.lecun.com/exdb/mnist/`, accessed 21 November 2019) has $n = 60,000$ observations in its training set and $10,000$ observations in its non-training set, where the ith observation is x_i, an image of a handwritten digit, and y_i, the correct digit, a category in $\{y(1), \ldots, y(10)\}$, where $y(k) = k-1$ for $k = 1, \ldots, 10$. A validation set and test set were created for Section 4.3 by randomly dividing the observations of the non-training set into two sets of 5000 observations each.

4.2 The prediction models

Three prediction models were used to analyze the data in Section 4.1: a logistic regression model, a random forest model, and a deep neural network. The set of models is denoted by $\mathcal{M}=\{\textbf{logistic}, \textbf{forest}, \textbf{LeNet}\}$. Each model was trained on the training set of Section 4.1 ($n = 60,000$).

The first two models are those of the `Classify` function of Wolfram Research, Inc. (2019) with all options at their defaults except `Method`, which is `LogisticRegression` for **logistic** and `RandomForest` for **forest**. The model **logistic** performed maximum likelihood estimation with the likelihood function penalized by L_2 regularization after using cross validation with the training data of Section 4.1 to choose 10 as the value of the regularization coefficient. Similarly using cross validation to set hyperparameters, **forest** settled on a random forest of 50 trees with leaves of size 5, training each tree on 1/28 of the observations in the training set.

The model **LeNet** is LeNet-5, the neural network described in LeCun et al. (1998). It is *deep* in the sense that it has multiple hidden layers of neurons. By contrast, **logistic** and **forest** are considered *shallow*. For an expository description of **LeNet**, see Wani et al. (2020, §4.2).

4.3 Sharpening versus other ensemble methods: Predictive performance

The predictive performance of the three prediction models of Section 4.2 and of several ensemble methods of combining the trained models was assessed by the mean loss

$$\widehat{\ell}\left(\bullet, \widehat{\text{mdl}}\right) = \frac{1}{5000}\sum_{t=1}^{5000} \ell\left(y_{n+t}, \widehat{\text{mdl}}\right),$$

where $n = 60,000$, $\ell \in \{\ell_{\log}, \ell_{\text{Brier}}\}$ (§3.2.2), either $\widehat{\text{mdl}} \in \{\textbf{logistic}, \textbf{forest}, \textbf{LeNet}\}$ or $\widehat{\text{mdl}}$ is an ensemble of $\{\textbf{logistic}, \textbf{forest}, \textbf{LeNet}\}$, and $\left(y_{n+1}, \ldots, y_{n+5000}\right)$ is either the validation set or the test set of Section 4.1.

The validation set was used to determine the lowest-$\widehat{\ell}\left(\bullet, \widehat{\text{mdl}}\right)$ values of α and κ of the sharpening methods of Section 2.3 with equation (6) and $P(\textbf{logistic}) = P(\textbf{forest}) = P(\textbf{LeNet}) = 1/3$, as follows. First, the mean losses of the prediction models corresponding to the classification probability $\widehat{p}^{(\alpha,\kappa)}(k)$ of the kth category for $\alpha = 1$ and the values of κ indicated in Figure 1 indicated that $\kappa = 4$ performed best. For that reason, κ was then fixed at 4 while trying the values of α appearing in Figure 1, with the finding that $\alpha = 1/8$ had the lowest mean losses. Finally, κ was varied while holding α at 1/8, resulting in $(\alpha, \kappa) = (1/8, 4)$ and $(\alpha, \kappa) = (1/8, 3)$ as the pairs that minimized $\widehat{\ell}_{\log}\left(\bullet, \widehat{\text{mdl}}\right)$ and $\widehat{\ell}_{\text{Brier}}\left(\bullet, \widehat{\text{mdl}}\right)$, respectively, as seen in Figure 1.

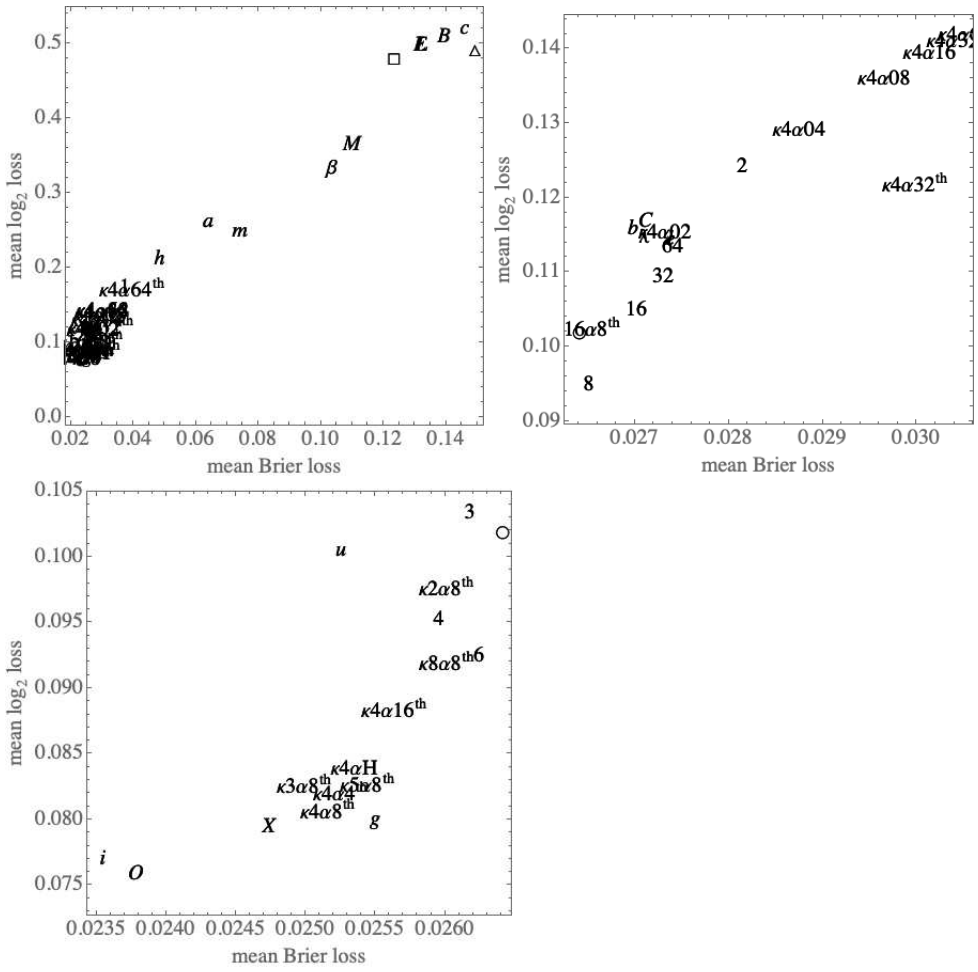

Figure 1: The validation set's $\widehat{\ell}_{\log}\left(\bullet, \widehat{\text{mdl}}\right)$ and $\widehat{\ell}_{\text{Brier}}\left(\bullet, \widehat{\text{mdl}}\right)$ for $\widehat{\text{mdl}} \in \{\triangle = \textbf{logistic}, \square = \textbf{forest}, \circ = \textbf{LeNet}\}$ and for each ensemble $\widehat{\text{mdl}}$ labeled by a plot symbol of Table 1. Each plot zooms to a different range of mean losses.

The validation set was also used to determine the lowest-$\widehat{\ell}\left(\bullet, \widehat{\text{mdl}}\right)$ methods among the methods in Section 3 that appear in Table 1. Figure 1 shows that whereas the natural odds-based probability minimized $\widehat{\ell}_{\log}\left(\bullet, \widehat{\text{mdl}}\right)$, the normalized inferential probability minimized $\widehat{\ell}_{\text{Brier}}\left(\bullet, \widehat{\text{mdl}}\right)$. Those ensemble methods only differ by whether they use non-extreme probabilities, which is clear when writing them, in the notation in Section 3.1.1, as $p^{\text{mean}}_{\text{norm}}\left(\text{logit}\, p\left(k\right)\right)$ and $p^{\text{mean}}_{\text{norm}}\left(\text{logit}\,\text{ext}\, p\left(k\right)\right)$.

To eliminate selection bias, the test set was used to measure the performance of the ensemble methods that performed best according to the validation set. The results appear in Figure 2.

Symbol	Key term	Probability of $y\,(k)$	Section
2	estimated $(1,2)$-sharpened probability	$\widehat{p}^{(1,2)}\,(k)$	2.3
h	estimated $\left(1,\frac{1}{2}\right)$-sharpened probability	$\widehat{p}^{\left(1,\frac{1}{2}\right)}\,(k)$	2.3
$\kappa 4\alpha 02$	estimated $(2,4)$-sharpened probability	$\widehat{p}^{(2,4)}\,(k)$	2.3
$\kappa 4\alpha 64^{\text{th}}$	estimated $\left(\frac{1}{64},4\right)$-sharpened probability	$\widehat{p}^{\left(\frac{1}{64},4\right)}\,(k)$	2.3
$\kappa 4\alpha \mathrm{H}$	estimated $\left(\frac{1}{2},4\right)$-sharpened probability	$\widehat{p}^{\left(\frac{1}{2},4\right)}\,(k)$	2.3
e	minimum Shannon $(\alpha=1)$ entropy	$\widehat{p}^{(1,\infty)}\,(k)$	2.3
E	maximum Shannon $(\alpha=1)$ entropy	$\widehat{p}^{(1,-\infty)}\,(k)$	2.3
a	arithmetic mean	$\operatorname{mean}(p\,(k))$	3.1.1
g	normalized geometric mean	$\operatorname{gmean}_{\text{norm}}(p\,(k))$	3.1.1
X	normalized extreme geometric mean	$\operatorname{gmean}_{\text{norm}}(\operatorname{ext} p\,(k))$	3.1.1
O	natural odds-based probability	$p^{\text{mean}}_{\text{norm}}(\operatorname{logit} p\,(k))$	3.1.2
m	normalized median-probability	$p^{\text{median}}_{\text{norm}}(\operatorname{logit} p\,(k))$	3.1.2
M	normalized mode-probability	$p^{\text{mode}}_{\text{norm}}(\operatorname{logit} p\,(k))$	3.1.2
u	unnormalized inferential probability	$p^{\text{mean}}(\operatorname{logit}\operatorname{ext} p\,(k))$	3.1.2
i	normalized inferential probability	$p^{\text{mean}}_{\text{norm}}(\operatorname{logit}\operatorname{ext} p\,(k))$	3.1.2
C	most confident classifier	$p_{\overline{\text{mdl}}}\,(k)$	3.2.1
c	least confident classifier	$p_{\underline{\text{mdl}}}\,(k)$	3.2.1
L	minimax classifier, log loss	$p_{\widehat{\text{mdl}}_{\ell_{\log}(1)}}\,(k)$	3.2.2
B	minimax classifier, Brier loss	$p_{\widehat{\text{mdl}}_{\ell_{\text{Brier}}(1)}}\,(k)$	3.2.2
l	minimin classifier, log loss	$p_{\widehat{\text{mdl}}_{\ell_{\log}(0)}}\,(k)$	3.2.2
b	minimin classifier, Brier loss	$p_{\widehat{\text{mdl}}_{\ell_{\text{Brier}}(0)}}\,(k)$	3.2.2
λ	classifier with $c=\frac{1}{2}$, log loss	$p_{\widehat{\text{mdl}}_{\ell_{\log}\left(\frac{1}{2}\right)}}\,(k)$	3.2.2
β	classifier with $c=\frac{1}{2}$, Brier loss	$p_{\widehat{\text{mdl}}_{\ell_{\text{Brier}}\left(\frac{1}{2}\right)}}\,(k)$	3.2.2

Table 1: The ensemble methods of Figures 1-2. The beginning rows have examples of $\widehat{p}^{(\alpha,\kappa)}\,(k)$ for various values of α and κ. For instance, $\widehat{p}^{(1,2)}\,(k)$ is the estimated sharpened probability of $y\,(k)$ given $\alpha=1$ and $\kappa=2$. Whereas the notation in the probabilities column agrees with the notation of the main text, the much shorter symbols in the plots do not.

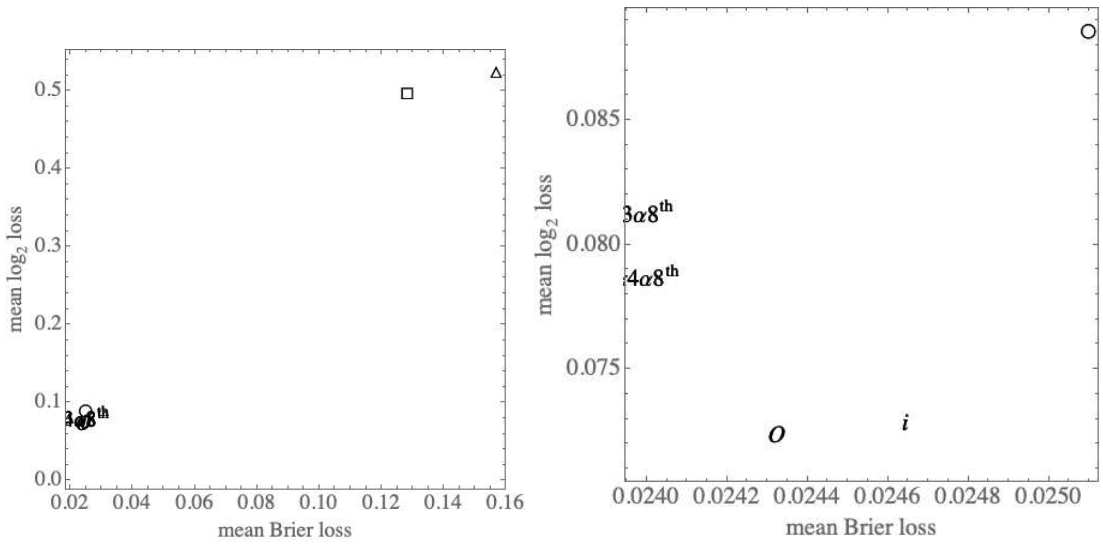

Figure 2: The test set's $\widehat{\ell}_{\log}\left(\bullet, \widehat{\text{mdl}}\right)$ and $\widehat{\ell}_{\text{Brier}}\left(\bullet, \widehat{\text{mdl}}\right)$ for $\widehat{\text{mdl}} \in \{\triangle = \textbf{logistic}, \square = \textbf{forest}, \circ = \textbf{LeNet}\}$ and for the ensemble methods with these classification probabilities: $\kappa 3\alpha 8^{\text{th}} \to \widehat{p}^{\left(\frac{1}{8},3\right)}$, $\kappa 4\alpha 8^{\text{th}} \to \widehat{p}^{\left(\frac{1}{8},4\right)}(k)$, $O \to p_{\text{norm}}^{\text{mean}}(\text{logit}\, p(k))$, $i \to p_{\text{norm}}^{\text{mean}}(\text{logit ext}\, p(k))$; see Table 1. Each plot zooms to a different range of mean losses.

5 Conclusions

According to the mean losses of ensemble methods according to the validation set (Figure 1), there is a tendency for ensemble methods considered conservative to outperform those considered optimistic: the most confident model, the minimin models (cf. Bickel, 2019c, App. A), and the minimum entropy model outperform the least confident model, the minimax models, and the maximum entropy model. That may be explained by the boldness (closeness to 0 or 1) of the classification probabilities of **LeNet**, the deep learner, compared to the less bold probabilities of **logistic** and **forest**, the shallow learners. The boldness of a learner in this case may indicate the predictive quality of its training. That could explain why the best ensemble methods among the sharpening methods (§2.3) have high sharpness ($\kappa = 3, 4$) and why the best of the other ensemble methods (§3) are based on means of log odds.

Using the test set to quantify the performance of those best ensemble methods (Figure 2) measures the extent to which sharpening can combine shallow models with a deep model to outperform the deep model alone. However, it also shows that sharpening is not unique in that ability, for the ensemble methods based on mean log odds also outperform **LeNet**. Interestingly, the sharpening methods have the lowest mean Brier loss, while the logit-based methods have the lowest mean log loss.

That advantage of the logit-based methods for the handwriting data set must be weighed against

their non-robustness, for a single model's classification probability too close to 0 or 1 can exert an undue influence on a mean log odds. Kittler et al. (1998) argued against the geometric mean as an ensemble method on the grounds of its similar sensitivity to probabilities close to 0.

Acknowledgments

This research was partially supported by the Natural Sciences and Engineering Research Council of Canada (RGPIN/356018-2009).

References

Alshemali, B., Graham, A., Kalita, J., 2020. Toward robust image classification. In: Bi, Y., Bhatia, R., Kapoor, S. (Eds.), Intelligent Systems and Applications. Springer International Publishing, Cham, pp. 483–489.

Bickel, D. R., 2019a. An explanatory rationale for priors sharpened into Occam's razors. Bayesian Analysis, DOI: 10.1214/19-BA1189.
URL https://doi.org/10.1214/19-BA1189

Bickel, D. R., 2019b. Confidence intervals, significance values, maximum likelihood estimates, etc. sharpened into Occam's razors. Communications in Statistics - Theory and Methods, DOI: 10.1080/03610926.2019.1580739.
URL https://doi.org/10.1080/03610926.2019.1580739

Bickel, D. R., 2019c. Maximum entropy derived and generalized under idempotent probability to address Bayes-frequentist uncertainty and model revision uncertainty, working paper, DOI: 10.5281/zenodo.2645555.
URL https://doi.org/10.5281/zenodo.2645555

Bickel, D. R., 2019d. Reporting Bayes factors or probabilities to decision makers of unknown loss functions. Communications in Statistics - Theory and Methods 48, 2163–2174.

Bickel, D. R., 2019e. Testing prediction algorithms as null hypotheses: Application to assessing the performance of deep neural networks, working paper, DOI: 10.5281/zenodo.3525401.
URL https://doi.org/10.5281/zenodo.3525401

Bickel, D. R., Frühwirth, R., 2006. On a fast, robust estimator of the mode: comparisons to other robust estimators with applications. Computational Statistics and Data Analysis 50, 3500–3530.

Breiman, B., 2001. Statistical modeling: The two cultures (with comments and a rejoinder). Statistical Science 16, 199–231.

Dam, H., Abbass, H., Lokan, C., 2006. Bcs: a bayesian learning classifier system. Tech. rep.

Eklund, J., Karlsson, S., 2007. Forecast combination and model averaging using predictive measures. Econometric Reviews 26 (2-4), 329–363.

Fraser, D. A. S., 2004. Ancillaries and conditional inference. Statistical Science 19, 333–351.

Geisser, S., 1971. The inferential use of predictive distributions. Foundations of Statistical Inference. Holt, Rinehart, and Winston, Toronto, pp. 456–469.

Hurwicz, L., 1951. Optimality criteria for decision making under ignorance. Cowles Commission Discussion Paper 370.

Kass, R. E., Wasserman, L., 1996. The selection of prior distributions by formal rules. Journal of the American Statistical Association 91, 1343–1370.

Kittler, J., Hatef, M., Duin, R. P. W., Matas, J., March 1998. On combining classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (3), 226–239.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., et al., 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE 86 (11), 2278–2324.

Rényi, A., 1965. On the foundations of information theory. Revue de l'Institut International de Statistique / Review of the International Statistical Institute 33, 1–14.

Wani, M., Bhat, F., Afzal, S., Khan, A., 2020. Advances in Deep Learning. Studies in big data. Springer, New York.

Wolfram Research, Inc., 2019. Mathematica, version 12.0.0.0.
URL `https://www.wolfram.com`

Zhou, Z.-H., 2012. Ensemble methods: foundations and algorithms. Chapman and Hall/CRC, New York.