Bayesian Learning for Classification using a Uniform Dirichlet Prior

Paul Rademacher
Radar Division
U.S. Naval Research Laboratory
Washington, DC, USA
paul.rademacher@nrl.navy.mil

Miloš Doroslovački
Department of Electrical and Computer Engineering
The George Washington University
Washington, DC, USA
doroslov@gwu.edu

***Abstract*—In Bayesian learning, designs based on non-informative priors are appropriate when the user cannot confidently identify the data-generating distribution. While such learners cannot achieve the performance of those based on a well-matched subjective prior, they impart a robustness against poor prior selection. The uniform Dirichlet distribution is the true non-informative prior as it has full support over the space of candidate distributions; additionally, it leads to closed-form posteriors. This work applies such a prior to classification using the 0–1 loss, determines the optimal Bayes classifier and the corresponding minimum probability of error, and analyzes the results.**

***Index Terms*—Bayesian learning, machine learning, classification, Dirichlet distribution, predictive distribution**

I. Introduction

Bayesian approaches to machine learning attempt to make better decisions by exploiting prior knowledge regarding the data-generating distribution. A prior probability distribution weights the different data distributions and defines the mechanism for prediction of unknown quantities using independent training data. When highly concentrated "subjective" priors are used, the performance of the learned functions can vary widely [1]. If the prior is localized around the true data-generating model, low-risk decisions can be made even with limited training data; conversely, if the prior assigns low weighting to the true model, satisfactory performance may not be realized.

If the designer does not have prior confidence in any specific model, a non-informative prior distribution can be used to weight the different models equally. Although learners designed with such priors will not perform as well as those made with well-selected subjective priors, they are more robust against poor prior selection. Often, priors are termed non-informative as long as they are approximately uniform over their support. The uniform Dirichlet distribution is unique in that it has full support over the space of data-generating distributions and is thus truly non-informative. Additionally, it is a conjugate prior [2] for independent, identically distributed observations and leads to a closed-form model posterior distribution.

This work assumes that joint observed/unobserved random variables are drawn from finite sets, enabling the use of the uniform Dirichlet prior. After discussing the relevant probability distributions, they will be applied to Bayesian classification using the 0–1 loss function [3]. The effects of the data set cardinalities and the volume of training data on the minimum probability of error will be analyzed and discussed.

II. Objective

Consider an observable random element $\mathrm{x} \in \mathcal{X}$ and and unobservable random element $\mathrm{y} \in \mathcal{Y}$ which are jointly distributed according to an unknown probability mass function (PMF) $\theta \in \Theta = \left\{\theta \in \mathbb{R}_{\geq 0}{}^{\mathcal{Y}\times\mathcal{X}} : \sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}}\theta(y,x) = 1\right\}$, such that $\mathrm{P}_{\mathrm{y},\mathrm{x}\,|\,\theta}(y,x|\theta) = \theta(y,x)$.

Also observed is a random sequence of N samples drawn from θ, denoted $\mathrm{D} = (\mathrm{Y},\mathrm{X}) \in \mathcal{D} = \{\mathcal{Y}\times\mathcal{X}\}^N$. The N data pairs are identically distributed as $\mathrm{P}_{\mathrm{D}_n\,|\,\theta}(y,x|\theta) = \theta(y,x)$ and are conditionally independent from one another and from the novel pair (y,x).

The objective is to design a decision function $f : \mathcal{D} \mapsto \mathcal{H}^{\mathcal{X}}$ which outputs a mapping from the space of the observed random element x to a decision space $\mathcal{H}$. The metric guiding the design is a loss function $\mathcal{L} : \mathcal{H}\times\mathcal{Y} \mapsto \mathbb{R}_{\geq 0}$ which penalizes the decision $h \in \mathcal{H}$ based on the value of y. The conditional expected loss, or conditional "risk", is defined as

$$\mathcal{R}_\Theta(f;\theta) = \mathrm{E}_{\mathrm{D}\,|\,\theta}\left[\mathrm{E}_{\mathrm{y},\mathrm{x}\,|\,\theta}\left[\,\mathcal{L}\left(f(\mathrm{x};\mathrm{D}),\mathrm{y}\right)\right]\right]. \tag{1}$$

As the model θ is not observed, $\mathcal{R}_\Theta$ is not yet a feasible objective function for optimization. If the designer selects a probability density function (PDF) p_θ, the Bayes risk is formulated as

$$\begin{aligned}\mathcal{R}(f) &= \mathrm{E}_\theta\left[\,\mathcal{R}_\Theta(f;\theta)\right] \\ &= \mathrm{E}_{\mathrm{y},\mathrm{x},\mathrm{D}}\left[\,\mathcal{L}(f(\mathrm{x};\mathrm{D}),\mathrm{y})\right]\end{aligned} \tag{2}$$

and y, x, and D are treated as jointly distributed random elements. The optimal learning function is expressed as

$$f^*(\mathrm{x};\mathrm{D}) = \arg\min_{h\in\mathcal{H}} \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}\left[\,\mathcal{L}(h,\mathrm{y})\right] \tag{3}$$

and the corresponding minimum Bayes risk is

$$\mathcal{R}(f^*) = \mathrm{E}_{\mathrm{x},\mathrm{D}}\left[\min_{h\in\mathcal{H}} \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}\left[\,\mathcal{L}(h,\mathrm{y})\right]\right]. \tag{4}$$

III. Probability Distributions

In this section, the joint PMF $\mathrm{P}_{\mathrm{y,x,D}}$ is determined using the uniform Dirichlet prior p_{θ}. Other distributions of interest will be provided, including the training data PMF P_{D} and the predictive distribution $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x,D}}$.

A. Model PDF, p_{θ}

The uniform PDF of the model random process $\theta \in \Theta$ is Dirichlet [4] with parameters $\alpha(y,x) = 1$, such that

$$\begin{aligned}\mathrm{p}_{\theta}(\theta) &= \left[\beta(\alpha)^{-1} \prod_{y\in\mathcal{Y}} \prod_{x\in\mathcal{X}} \theta(y,x)^{\alpha(y,x)-1}\right]_{\alpha(\cdot,\cdot)=1} \\ &= \big(|\,\mathcal{Y}\,||\,\mathcal{X}\,|-1\big)!\end{aligned} \tag{5}$$

is uniform over Θ. The operator β is the generalized beta function.

For convenience, the Dirichlet concentration parameter is defined as $\alpha_0 \equiv \sum_{y\in\mathcal{Y}} \sum_{x\in\mathcal{X}} \alpha(y,x)$. The mean function of the model is

$$\mu_{\theta}(y,x) = \left.\frac{\alpha(y,x)}{\alpha_0}\right|_{\alpha(\cdot,\cdot)=1} = \big(|\,\mathcal{Y}\,||\,\mathcal{X}\,|\big)^{-1}. \tag{6}$$

The marginal data PMF $\mathrm{P}_{\mathrm{y,x}} = \mu_{\theta}$ is uniform over $\mathcal{Y}\times\mathcal{X}$.

B. Training Set PMF, P_{D}

Next, the conditional distribution $\mathrm{P}_{\mathrm{D}\,|\,\theta}$ will be used to determine the marginal training data PMF, P_{D}. The distribution of D conditioned on the model can be formulated as

$$\mathrm{P}_{\mathrm{D}\,|\,\theta}(D|\theta) = \prod_{y\in\mathcal{Y}} \prod_{x\in\mathcal{X}} \theta(y,x)^{\bar{N}(y,x;D)}, \tag{7}$$

where the dependency on the training data D is expressed though a transform function $\bar{N} : \mathcal{D} \mapsto \bar{\mathcal{N}}$, defined as $\bar{N}(y,x;D) = \sum_{n=1}^{N} \delta[y,Y_n]\,\delta[x,X_n]$, which counts the number of occurences of the pair (y,x) in the training set D. The range of the transform is the function space $\bar{\mathcal{N}} = \left\{\bar{n} \in \mathbb{Z}_{\geq 0}{}^{\mathcal{Y}\times\mathcal{X}} : \sum_{y\in\mathcal{Y}} \sum_{x\in\mathcal{X}} \bar{n}(y,x) = N\right\}$. The cardinality of the set is $|\bar{\mathcal{N}}| = \binom{N+|\,\mathcal{Y}\,||\,\mathcal{X}\,|-1}{|\,\mathcal{Y}\,||\,\mathcal{X}\,|-1}$; this can be shown using the stars-and-bars method [5].

Note that $\mathrm{P}_{\mathrm{D}|\theta}$ depends on the training data D only through the transform $\bar{N}$; $\bar{N}(\mathrm{D})$ is thus a sufficient statistic [6] for the model θ. Consequently, other distributions of interest P_{D} and $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x,D}}$ will also depend on D via $\bar{N}(\mathrm{D})$. As such, it is useful to define a new random process $\bar{\mathrm{n}} \equiv \bar{N}(\mathrm{D}) \in \bar{\mathcal{N}}$.

The conditional PMF $\mathrm{P}_{\mathrm{n}\,|\,\theta}$ is easily shown to be Multinomial. As a Dirichlet distribution p_{θ} characterizes the parameters of this distribution, the marginal PMF of $\bar{\mathrm{n}}$ is a Dirichlet-Multinomial distribution [7] parameterized by $\alpha = 1$,

$$\begin{aligned}\mathrm{P}_{\bar{\mathrm{n}}}(\bar{n}) &= \mathcal{M}(\bar{n})\beta(\alpha)^{-1}\beta(\alpha+\bar{n})\big|_{\alpha=1} \\ &= |\bar{\mathcal{N}}|^{-1} = \binom{N+|\,\mathcal{Y}\,||\,\mathcal{X}\,|-1}{|\,\mathcal{Y}\,||\,\mathcal{X}\,|-1}^{-1},\end{aligned} \tag{8}$$

which is uniform over the set $\bar{\mathcal{N}}$. The operator $\mathcal{M}$ represents the multinomial coefficient.

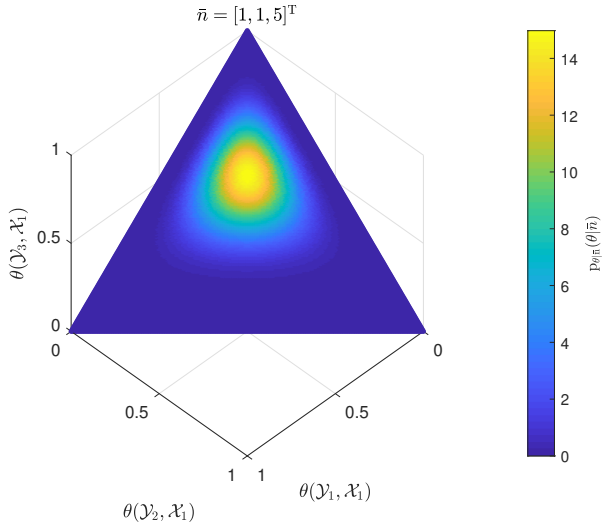

Fig. 1. Model Posterior for an example training set $\bar{\mathrm{n}}$

C. Predictive PMF, $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x,D}}$

As shown in Equation (3), the decision selected by the optimally designed function depends on the Bayesian predictive PMF $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x,D}}$. First, note that as $\mathrm{P}_{\mathrm{D}\,|\theta}$ is of exponential form, the Dirichlet PDF p_{θ} is its conjugate prior [2]; thus, the posterior PDF $\mathrm{p}_{\theta|\,\mathrm{D}}$ is a Dirichlet distribution with parameter function $\alpha(y,x) = \bar{N}(y,x;\mathrm{D}) + 1$,

$$\begin{aligned}\mathrm{p}_{\theta|\,\mathrm{D}}(\theta|D) = \big(N+|\,\mathcal{Y}\,||\,\mathcal{X}\,|-1\big)! \\ \times \prod_{y\in\mathcal{Y}} \prod_{x\in\mathcal{X}} \frac{\theta(y,x)^{\bar{N}(y,x;\mathrm{D})}}{\bar{N}(y,x;\mathrm{D})!}.\end{aligned} \tag{9}$$

The posterior concentration parameter is $\alpha_0 = N + |\,\mathcal{Y}\,||\,\mathcal{X}\,|$; Figure 1 shows how the posterior is localized around the empirical PMF $\bar{N}(\mathrm{D})/N$ when the model is conditioned on the training data.

The joint PMF of y and x conditioned on the training data is expressed as $\mathrm{P}_{\mathrm{y,x}\,|\,\mathrm{D}} = \mu_{\theta|\,\mathrm{D}}$ [8]. The predictive distribution of interest is generated via Bayes rule as

$$\begin{aligned}\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x,D}}(y,x|D) &= \frac{\bar{N}(y,x;D)+1}{N'(x;D)+|\,\mathcal{Y}\,|} \\ &= \left(\frac{|\,\mathcal{Y}\,|}{N'(x;D)+|\,\mathcal{Y}\,|}\right)\frac{1}{|\,\mathcal{Y}\,|} \\ &\quad + \left(\frac{N'(x;D)}{N'(x;D)+|\,\mathcal{Y}\,|}\right)\frac{\bar{N}(y,x;D)}{N'(x;D)},\end{aligned} \tag{10}$$

where $N'(x;D) = \sum_{n=1}^{N} \delta[x,X_n]$.

The second form represents the predictive PMF as a convex combination of two conditional distributions. The first PMF $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x}} = |\,\mathcal{Y}\,|^{-1}$ is uniform and independent of the training data; the second distribution is the conditional empirical PMF. As the number of matching training data $N'(x;\mathrm{D})$ increases relative to the number of classes $|\,\mathcal{Y}\,|$, the predictive PMF tends toward the empirical PMF.

IV. Classification using the 0–1 Loss

In this section, the developed framework is applied to classification. The 0–1 loss function is the most widely used for these problems; it is represented as $\mathcal{L}(h, y) = 1 - \delta[h, y]$ with hypothesis space $\mathcal{H} = \mathcal{Y}$.

A. Optimal Hypothesis: the Conditional Majority Decision

To determine the optimal learning function, the 0–1 loss is substituted into (3) to find

$$\begin{aligned} f^*(\mathrm{x}; \mathrm{D}) &= \arg\min_{h \in \mathcal{Y}} \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}} \left[1 - \delta[h, \mathrm{y}]\right] \\ &= \arg\max_{y \in \mathcal{Y}} \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}(y|\,\mathrm{x}, \mathrm{D}) \\ &= \arg\max_{y \in \mathcal{Y}} \bar{N}(y, \mathrm{x}; \mathrm{D}) \ . \end{aligned} \tag{11}$$

The optimal classifier chooses the value $y \in \mathcal{Y}$ that maximizes the predictive PMF given the observed values of x and D. It is a conditional majority decision which chooses the class from $\mathcal{Y}$ most often represented among training set samples D with a matching input value $\mathrm{x} = x$.

B. Minimum Bayes Probability of Error

Substituting the 0–1 loss into (4), the minimum Bayes 0–1 risk is

$$\begin{aligned} \mathcal{R}^* &= 1 - \mathrm{E}_{\mathrm{x},\mathrm{D}} \left[\max_{y \in \mathcal{Y}} \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}(y|\,\mathrm{x}, \mathrm{D})\right] \\ &= 1 - \sum_{x \in \mathcal{X}} \frac{\mathrm{E}_{\bar{\mathrm{n}}}\left[\max_{y \in \mathcal{Y}} \bar{\mathrm{n}}(y, x)\right] + 1}{|\,\mathcal{Y}\,||\,\mathcal{X}\,| + N} \ . \end{aligned} \tag{12}$$

The expectation operates on the maximum value from a subset of the random process $\bar{\mathrm{n}}$ – it is found in Appendix A. Substituting, the minimum 0–1 Bayes risk is

$$\begin{aligned} \mathcal{R}^* &= 1 - \left(|\,\mathcal{Y}\,| + N/|\,\mathcal{X}\,|\right)^{-1} \\ &\times \sum_{m=1}^{|\mathcal{Y}|} \binom{|\,\mathcal{Y}\,|}{m} (-1)^{m-1} \sum_{n=0}^{\left\lfloor \frac{N}{m} \right\rfloor} \prod_{l=1}^{|\,\mathcal{Y}\,||\,\mathcal{X}\,|-1} \left(1 - \frac{mn}{N+l}\right) . \end{aligned} \tag{13}$$

It is informative to express the risk for minimal and maximal volumes of training data. Using an identity for alternating binomial sums of polynomials [9], it can be shown that for $N = 0$, the minimum risk is $\mathcal{R}^* = 1 - |\,\mathcal{Y}\,|^{-1}$.

To find the risk for $N \to \infty$, note that

$$\begin{aligned} &\lim_{N \to \infty} \left(|\,\mathcal{Y}\,| + N/|\,\mathcal{X}\,|\right)^{-1} \sum_{n=0}^{\left\lfloor \frac{N}{m} \right\rfloor} \prod_{l=1}^{|\,\mathcal{Y}\,||\,\mathcal{X}\,|-1} \left(1 - \frac{mn}{N+l}\right) \\ &= \lim_{N/m \to \infty} \frac{|\,\mathcal{X}\,|}{N} \sum_{n=0}^{\left\lfloor \frac{N}{m} \right\rfloor} \left(1 - \frac{mn}{N}\right)^{|\,\mathcal{Y}\,||\,\mathcal{X}\,|-1} \\ &= \frac{|\,\mathcal{X}\,|}{m} \int_0^1 (1-t)^{|\,\mathcal{Y}\,||\,\mathcal{X}\,|-1} \mathrm{d}t = \frac{1}{m|\,\mathcal{Y}\,|} \ . \end{aligned} \tag{14}$$

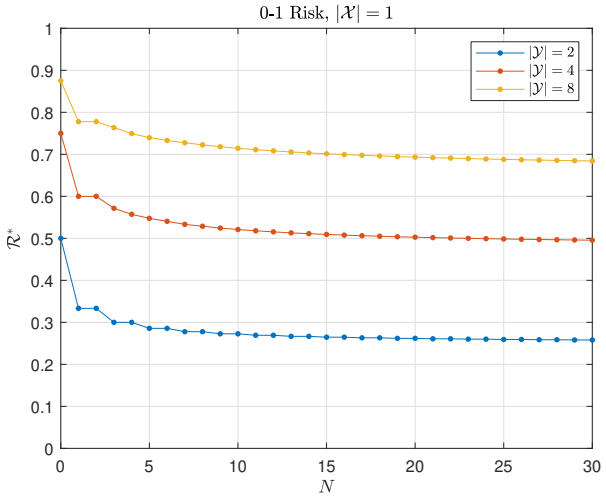

Fig. 2. Minimum 0–1 Risk for different numbers of classes

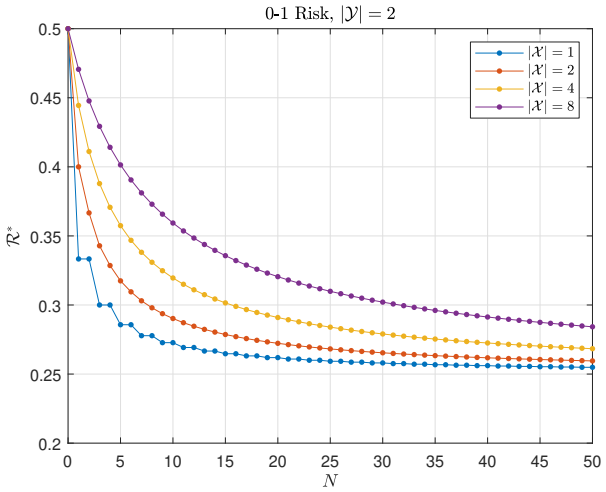

Fig. 3. Minimum 0–1 Risk for different numbers of possible observations

Thus, the 0–1 Bayes risk for the uniform prior tends toward

$$\begin{aligned} \mathcal{R}^* &\to 1 - |\,\mathcal{Y}\,|^{-1} \sum_{m=1}^{|\mathcal{Y}|} \binom{|\,\mathcal{Y}\,|}{m} (-1)^{m-1} m^{-1} \\ &= 1 - |\,\mathcal{Y}\,|^{-1} \sum_{m=1}^{|\mathcal{Y}|} m^{-1} \ , \end{aligned} \tag{15}$$

providing a lower bound for the achievable Bayes probability of error. The above formulation has made use of the alternating summation identity from [10] to express the risk using the $|\,\mathcal{Y}\,|^{\text{th}}$ harmonic number $H_{|\,\mathcal{Y}\,|} = \sum_{m=1}^{|\mathcal{Y}|} m^{-1}$. Observe that the risk does not depend on the cardinality $|\,\mathcal{X}\,|$.

Figures 2 and 3 demonstrate how the minimum 0–1 risk decreases with training volume N for different cardinalities $|\,\mathcal{Y}\,|$ and $|\,\mathcal{X}\,|$ respectively. With a larger set of classes, accurate classification becomes more difficult; similarly, the probability of error increases with the number of possible observations $\mathrm{x} = x$.

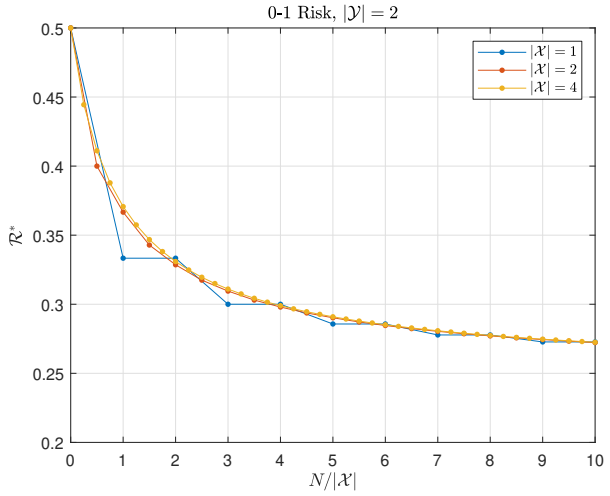

Fig. 4. Minimum 0–1 Risk as a function of $N/|\mathcal{X}|$

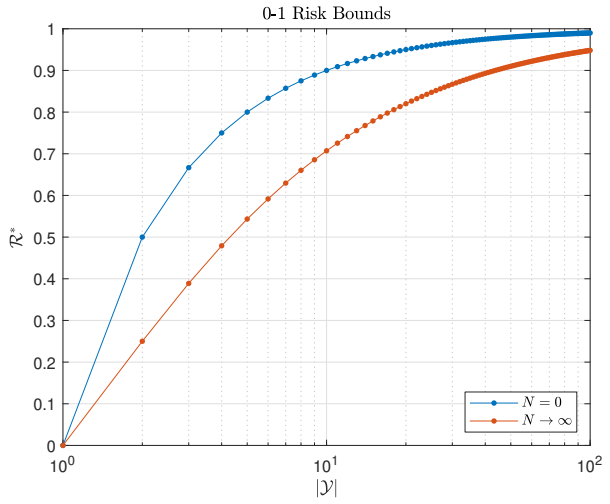

Fig. 5. Minimum 0–1 Risk for zero and infinite number of training data

Further insight into how the risk depends on $|\mathcal{X}|$ is obtained by plotting the risk as a function of $N/|\mathcal{X}|$. In Figure 4, it is shown that the optimal risk can be approximated by a function dependent only on $N/|\mathcal{X}|$; of the series plotted, only the series for $|\mathcal{X}| = 1$ shows non-negligible deviation from the others.

It is also informative to graph the $N = 0$ and $N \to \infty$ minimum risk values as a function of $|\mathcal{Y}|$. Figure 5 displays these bounds; note the margin in the probability of error between the optimal $N = 0$ and $N \to \infty$ classifiers. For binary classification ($|\mathcal{Y}| = 2$), both sequences are at their minimum and infinite training data reduces the expected probability of error from 0.5 to 0.25. As $|\mathcal{Y}|$ increases, the classification risk of both sequences tends to unity and the improvement using training data becomes negligible.

V. Conclusion

This work has used the non-informative uniform Dirichlet prior for Bayesian learning, discussed the Bayes predictive distribution, and applied the results to classification. The optimal majority decision learner and the minimum 0–1 risk have been analyzed. Graphical examples illustrate how the probability of error increases with the number of classes and the number of possible observations. Additionally, the asympotic Bayes risk using infinite training data has been found and its deviation from the risk of an untrained classifier has been discussed.

Future work will expand the results presented for a general Dirichlet prior. The effect of the subjectivity of the prior on the conditional risk (1) will be of specific interest. Additionally, the work will be generalized for an infinite number of possible observations using Dirichlet processes

Appendix A
Evaluation of $\mathrm{E}_{\bar{\mathrm{n}}}\left[\max_{y\in\mathcal{Y}} \bar{\mathrm{n}}(y,x)\right]$

To evaluate the expectation, new random variables $\bar{\mathrm{n}}_{\max}(x) \equiv \max_{y\in\mathcal{Y}} \bar{\mathrm{n}}(y,x)$ are introduced and characterized by the event probabilities $\mathrm{P}\left(\bar{\mathrm{n}}_{\max}(x) \geq n\right) = \mathrm{P}\left(\cup_{y\in\mathcal{Y}} \{\bar{\mathrm{n}}(y,x) \geq n\}\right)$. As the distribution of $\bar{\mathrm{n}}$ is uniform, the event probability is proportionate to the cardinality of the set $\cup_{y\in\mathcal{Y}}\{\bar{n} : \bar{n}(y,x) \geq n\}$. Using the inclusion-exclusion principle [11], the cardinality is represented as

$$\left|\cup_{y\in\mathcal{Y}} \{\bar{n} : \bar{n}(y,x) \geq n\}\right| \tag{16}$$
$$= \begin{cases} \binom{N+|\mathcal{Y}||\mathcal{X}|-1}{|\mathcal{Y}||\mathcal{X}|-1} & \text{if } n < 0, \\ \sum_{m=1}^{|\mathcal{Y}|} \binom{|\mathcal{Y}|}{m}(-1)^{m-1} & \\ \quad \times \binom{N-mn+|\mathcal{Y}||\mathcal{X}|-1}{|\mathcal{Y}||\mathcal{X}|-1} H\left(\left\lfloor \frac{N}{m} \right\rfloor - n\right) & \text{if } 0 \leq n \leq N, \\ 0 & \text{if } n > N, \end{cases}$$

where $H : \mathbb{Z} \mapsto \{0,1\}$ is the discrete Heaviside step function. Note the independence from x.

For $n < 0$, the cardinality is equivalent to $|\bar{\mathcal{N}}|$. For $0 \leq n < N$, the cardinality is an alternating binomial summation where the m^{th} term accounts for the different intersections of m of the $|\mathcal{Y}|$ individual sets $\{\bar{n} : \bar{n}(y,x) \geq n\}$. Observe that the cardinality of the intersections is only dependent on the number of contributing sets m and not on which sets intersect. Furthermore, note the dependency of the intersection cardinalities on the argument n. The step function contributes such that if $n > \left\lfloor \frac{N}{m} \right\rfloor$, only up to $m-1$ individual sets will intersect. The binomial coefficient calculates the intersection cardinality for a given m; note the similarity to the cardinality $|\bar{\mathcal{N}}|$ - the only difference is the number of points characterizing the $|\mathcal{Y}||\mathcal{X}| - 1$ dimensional region.

The probability of interest can thus be expressed as $\mathrm{P}\left(\bar{\mathrm{n}}_{\max}(x) \geq n\right) = |\bar{\mathcal{N}}|^{-1} \left|\cup_{y\in\mathcal{Y}} \{\bar{n} : \bar{n}(y,x) \geq n\}\right|$. Since the PMF of $\bar{\mathrm{n}}_{\max}(x)$ has support on $n \in [0,\ldots,N]$, the expectation over $\bar{\mathrm{n}}$ is evaluated as

$$\mathrm{E}_{\bar{\mathrm{n}}}\left[\bar{\mathrm{n}}_{\max}(x)\right] = \sum_{n=0}^{N} n\,\mathrm{P}\left(\bar{\mathrm{n}}_{\max}(x) = n\right) \tag{17}$$
$$= -1 + \sum_{n=0}^{N} \mathrm{P}\left(\bar{\mathrm{n}}_{\max}(x) \geq n\right)$$
$$= -1 + \sum_{m=1}^{|\mathcal{Y}|} \binom{|\mathcal{Y}|}{m}(-1)^{m-1} \sum_{n=0}^{\left\lfloor \frac{N}{m} \right\rfloor} \prod_{l=1}^{|\mathcal{Y}||\mathcal{X}|-1} \left(1 - \frac{mn}{N+l}\right).$$

REFERENCES

[1] G. E. Box and G. C. Tiao, *Bayesian Inference in Statistical Analysis*. Wiley, 1973.

[2] S. Theodoridis, *Machine Learning: A Bayesian and Optimization Perspective*. Elsevier, 2015.

[3] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. Springer, 1980.

[4] C. M. Bishop, *Pattern Recognition and Machine Learning*, ser. Information Science and Statistics. Springer, 2006.

[5] W. Feller, *An Introduction to Probability Theory and Its Applications*, 2nd ed., ser. Probability and Mathematical Statistics. New York, New York: John Wiley & Sons, 1971, vol. 2.

[6] J. M. Bernardo and A. F. M. Smith, *Bayesian Theory*, ser. Wiley Series in Probability and Statistics. John Wiley & Sons, 2000.

[7] N. L. Johnson, S. Kotz, and N. Balakrishnan, *Discrete Multivariate Distributions*, ser. Probability and Statistics. John Wiley & Sons, 1997.

[8] K. P. Murphy, "Binomial and multinomial distributions," University of British Columbia, Tech. Rep., 2006.

[9] R. L. Graham, D. E. Knuth, and O. Patashnik, *Concrete Mathematics*, 2nd ed. Reading, Massachusetts: Addison-Wesley, 1994.

[10] S. Roman, "The logarithmic binomial formula," *American Mathematical Monthly*, vol. 99, no. 7, 1992.

[11] R. A. Brualdi, *Introductory Combinatorics*, 5th ed. Pearson, 2010.