Predictive Distribution Estimation for Bayesian Machine Learning using a Dirichlet Process Prior

Paul Rademacher
U.S. Naval Research Laboratory
Radar Division
Washington, DC 20375, USA

Miloš Doroslovački
The George Washington University
Department of Electrical and Computer Engineering
Washington, DC 20052, USA

***Abstract*—In Bayesian treatments of machine learning, the success or failure of the estimator/classifier hinges on how well the prior distribution selected by the designer matches the actual data-generating model. This paper assumes that the model distribution is a realization of a Dirichlet process and assesses the mismatch between the true predictive distribution and the predictive distribution approximated using the training data. It is shown that highly localized Dirichlet priors can overcome the burden of a limited training set when the prior mean is well matched to the true distribution, but will degrade the approximation if the match is poor. A bias/variance trade-off will be demonstrated with illustrative examples.**

I. Introduction

This article investigates how a Bayesian perspective influences the predictive distributions used to make decisions in machine learning applications. The efficacy of Bayesian learning methods depends on how well the prior knowledge imparted by the designer matches reality. The chosen prior distribution over the set of data-generating probability mass functions (PMF) reflects the users confidence that different PMF's are responsible for generating the observed/unobserved random elements. If a highly informative prior is chosen that strongly weights the actual data PMF, low risk learning is possible even with a limited amount of training data; however, if the prior is poorly selected, a good solution may not be achieved. Conversely, a non-informative prior with high variance will always be able to adapt with enough training data; if the amount of data is limited, however, the learning function may not deliver the required performance.

This work assumes that the prior distribution is Dirichlet. The class of Dirichlet probability density functions (PDF) has the desirable properties of full support over the set of possible PMF's and a tractable posterior distribution for independently and identically distributed data [1]. The full support is necessary to ensure that the true underlying model can be identified with enough training samples. Also, control of the Dirichlet parameters can enable both non-informative and informative prior distributions.

Once the Bayesian assumption is used to form the predictive distribution conditioned on the training data, it will be compared to the true predictive distribution given knowledge of the model PMF. Specifically, the mean and covariance functions of the difference between the two PMF's will be determined. Specific attention will be given to various asymptotic cases of the Dirichlet distribution to illustrate the bias/variance trade-off for both non-informative and informative priors.

Throughout this article, italic, roman, and calligraphic fonts, e.g. x, x, and $\mathcal{X}$, are reserved for specific values, random elements, and sets, respectively. The notation $\mathcal{Y}^{\mathcal{X}}$ represents the set of functions $g : \mathcal{X} \mapsto \mathcal{Y}$. For functions $g : \mathcal{X} \mapsto \mathcal{Z}^{\mathcal{Y}}$, $g(x) \in \mathcal{Z}^{\mathcal{Y}}$ is a function and $g(y; x) \in \mathcal{Z}$ is a scalar.

II. Objective

Consider an observable discrete random element $\mathrm{x} \in \mathcal{X}$ and and unobservable discrete random element $\mathrm{y} \in \mathcal{Y}$ which are jointly distributed according to an unknown 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 $\mathrm{P}_{\mathrm{y,x}\,|\,\theta}(y, x|\theta) = \theta(y, x)$. Also observed is a random sequence of N samples generated from θ, denoted $\mathrm{D} \in \mathcal{D} = \{\mathcal{Y} \times \mathcal{X}\}^N$. 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 (y, x) and from one another, such that

$$\mathrm{P}_{\mathrm{y,x,D}\,|\,\theta}(y, x, D|\theta) = \mathrm{P}_{\mathrm{y,x}\,|\,\theta}(y, x|\theta) \prod_{n=1}^{N} \mathrm{P}_{\mathrm{D}_n\,|\,\theta}\left(D_n|\theta\right). \tag{1}$$

Define the decision function $f : \mathcal{D} \mapsto \mathcal{H}^{\mathcal{X}}$, where $\mathcal{H}$ is the decision space. The metric guiding the design of f 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

$$\mathcal{R}_\theta(f) = \mathrm{E}_{\mathrm{x,D}\,|\,\theta}\left[\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\theta}\left[\mathcal{L}\left(f(\mathrm{x};\mathrm{D}), \mathrm{y}\right)\right]\right]. \tag{2}$$

If the model θ were known, a decision $h = \arg\min_{h \in \mathcal{H}} \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x,D},\theta}\left[\mathcal{L}(h, \mathrm{y})\right]$ could be made to minimize the objective for a given set of observations. It can be shown that given the model θ, the unobserved element y is conditionally independent of the training data D. As such, a "clairvoyant" decision function $f_\theta : \Theta \mapsto \mathcal{H}^{\mathcal{X}}$ could be designed that is dependent only on the true predictive distribution $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\theta}$.

However, as the model θ is not observed, this predictive PMF is unknown and the conditional risk objective is not

This work was supported by U.S. Naval Research Laboratory and Office of Naval Research.

feasible for optimization. To remove the dependency on θ, the model is treated as a random process Θ with PDF p_Θ and a Bayesian approach is adopted. The Bayes risk can be formulated as

$$\mathcal{R}(f) = \mathrm{E}_\Theta\left[\mathcal{R}_\Theta(f)\right] = \mathrm{E}_{\mathrm{x},\mathrm{D}}\left[\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}\left[\mathcal{L}\left(f(\mathrm{x};\mathrm{D}),\mathrm{y}\right)\right]\right] \quad (3)$$

and y, x, and D are treated as jointly distributed random elements.

For a given set of observations, the Bayes optimal decision is $h = \arg\min_{h\in\mathcal{H}} \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}\left[\mathcal{L}(h,\mathrm{y})\right]$, dependent on the Bayesian predictive PMF $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}$. Using Bayes theorem, it can be shown that $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}} = \mathrm{E}_{\Theta|\,\mathrm{x},\mathrm{D}}\left[\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\Theta}\right]$, a combination of the possible predictive PMF's using the model posterior PDF $\mathrm{p}_{\Theta|\,\mathrm{x},\mathrm{D}}$. As such, the Bayesian predictive distribution can be interpreted as an estimate of the true predictive distribution $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\Theta}$.

III. Bayesian Prediction

This section introduces the Dirichlet prior distribution for the model Θ and uses it to formulate the estimated predictive distribution $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}$.

A. Model PDF, p_Θ

The Dirichlet PDF for the model $\Theta \in \Theta$ is [2]

$$\mathrm{p}_\Theta(\theta) = \beta(\alpha)^{-1}\prod_{y\in\mathcal{Y}}\prod_{x\in\mathcal{X}}\theta(y,x)^{\alpha(y,x)-1}\,, \quad (4)$$

where the user-selected PDF parameters $\alpha : \mathcal{Y}\times\mathcal{X} \mapsto \mathbb{R}^+$ are introduced and β is the generalized beta function.

The parameter α controls around which models θ the PDF concentrates and how strongly. For convenience, introduce the concentration parameter $\alpha_0 \equiv \sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}}\alpha(y,x)$.

The mean and covariance functions of the model are [2]

$$\mu_\Theta(y,x) = \frac{\alpha(y,x)}{\alpha_0} \quad (5)$$

and

$$\Sigma_\Theta(y,x,y',x') = \frac{\mu_\Theta(y,x)\delta[y,y']\delta[x,x'] - \mu_\Theta(y,x)\mu_\Theta(y',x')}{\alpha_0+1}\,, \quad (6)$$

where $\delta[\cdot,\cdot]$ represents the Kronecker delta function.

1) Aggregation Properties: As x is observable and y is not, it is important to characterize the marginal model $\Theta' \equiv \sum_{y\in\mathcal{Y}}\Theta(y,\cdot)$ over the set $\mathcal{X}$ and the conditional models $\tilde{\Theta}(x) \equiv \Theta(\cdot,x)/\Theta'(x),\ \forall x\in\mathcal{X}$ over the set $\mathcal{Y}$.

By the aggregation property [1], Θ' is a Dirichlet random process parameterized by $\alpha' \equiv \sum_{y\in\mathcal{Y}}\alpha(y,\cdot)$. Additionally, it can be proven that the conditional PDF of $\tilde{\Theta}$ given Θ' is

$$\mathrm{p}_{\tilde{\Theta}|\Theta'}\left(\tilde{\theta}|\theta'\right) = \mathrm{p}_{\tilde{\Theta}}\left(\tilde{\theta}\right) = \prod_{x\in\mathcal{X}}\left[\beta\big(\alpha(\cdot,x)\big)^{-1}\prod_{y\in\mathcal{Y}}\tilde{\theta}(y;x)^{\alpha(y,x)-1}\right], \quad (7)$$

a product of Dirichlet distributions defined for $\tilde{\theta} \in \tilde{\Theta}^{\mathcal{X}}$, where $\tilde{\Theta} = \left\{p\in\mathbb{R}_{\geq 0}{}^{\mathcal{Y}} : \sum_{y\in\mathcal{Y}}p(y) = 1\right\}$. As shown, the normalized functions $\tilde{\Theta}(x)$ are independent of one another and of the aggregation Θ'. Note that $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\Theta}(y|x,\theta) \equiv \tilde{\theta}(y;x)$ compactly notates the true predictive distribution.

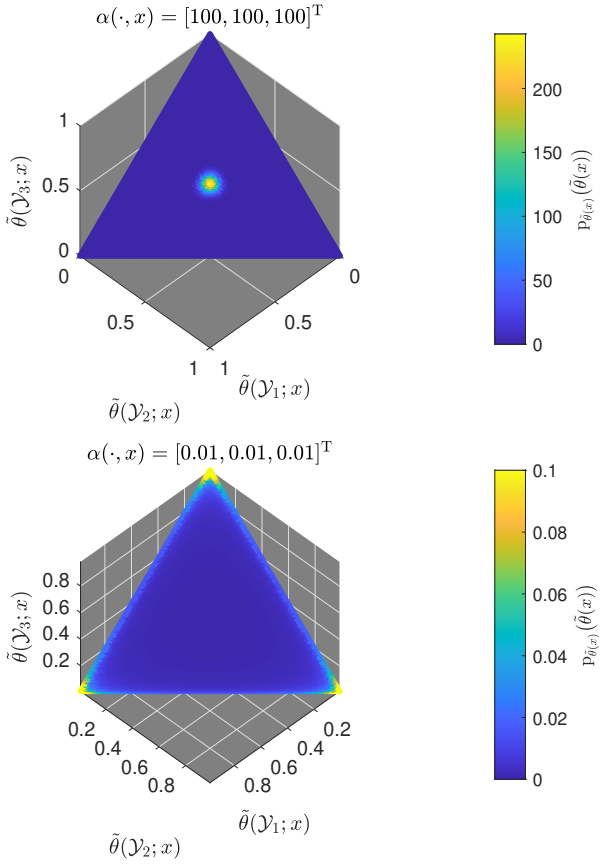

Fig. 1. Model prior PDF for different concentrations $\alpha'(x)$

Of specific interest is how $\mathrm{p}_{\tilde{\Theta}(x)}$ changes as its concentration parameter $\alpha'(x)$ approaches its limiting values. For $\alpha'(x)\to\infty$, the PDF concentrates at its mean, resulting in

$$\mathrm{p}_{\tilde{\Theta}(x)}\left(\tilde{\theta}(x)\right) \to \delta\left(\tilde{\theta}(x) - \frac{\alpha(\cdot,x)}{\alpha'(x)}\right)\,, \quad (8)$$

where $\delta(\cdot)$ denotes the Dirac delta function on the set $\tilde{\Theta}$. Conversely, for $\alpha'(x)\to 0$, the PDF tends toward

$$\mathrm{p}_{\tilde{\Theta}(x)}\left(\tilde{\theta}(x)\right) \to \sum_{y\in\mathcal{Y}}\frac{\alpha(y,x)}{\alpha'(x)}\delta\Big(\tilde{\theta}(x) - \delta[\cdot,y]\Big)\,, \quad (9)$$

which distributes its weight among the $|\,\mathcal{Y}\,|$ models $\delta[\cdot,y]\in\tilde{\Theta}$, each having an ℓ_0 norm equal to one. These trends are demonstrated in Figure 1.

B. Training Set conditional PMF, $\mathrm{P}_{\mathrm{D}\,|\Theta}$

Next, properties of the conditional distribution $\mathrm{P}_{\mathrm{D}\,|\Theta}$ will be discussed. The distribution of D conditioned on the model can be formulated as

$$\mathrm{P}_{\mathrm{D}\,|\Theta}\left(D|\theta\right) = \prod_{y\in\mathcal{Y}}\prod_{x\in\mathcal{X}}\theta(y,x)^{\bar{N}(y,x;D)}\,, \quad (10)$$

where the dependency on the training data D is expressed though a transform function $\bar{N}(y,x;D) = \sum_{n=1}^{N}\delta\big[(y,x),D_n\big]$ which counts the number of occurrences of each pair (y,x) in the training set D.

Note that $\mathrm{P}_{\mathrm{D}\,|\,\theta}$ depends on the training data D only through the transform $\bar{N}$; consequently, $\bar{N}(\mathrm{D})$ is a sufficient statistic for the model θ. As such, it is useful to define a new random process $\bar{\mathrm{n}} \equiv \bar{N}(\mathrm{D})$.

Conditioned on the model θ, the PMF of $\bar{\mathrm{n}}$ is a multinomial distribution

$$\begin{aligned}\mathrm{P}_{\bar{\mathrm{n}}\,|\,\theta}(\bar{n}|\theta) &= \sum_{D:\bar{N}(D)=\bar{n}} \mathrm{P}_{\mathrm{D}\,|\,\theta}(D|\theta) \\ &= \mathcal{M}(\bar{n}) \prod_{y\in\mathcal{Y}}\prod_{x\in\mathcal{X}} \theta(y,x)^{\bar{n}(y,x)}\ ,\end{aligned} \tag{11}$$

where the multinomial operator $\mathcal{M}$ is used. The mean and covariance functions of this multinomial distribution are [3]

$$\mu_{\bar{\mathrm{n}}|\,\theta}(y,x|\theta) = N\theta(y,x) \tag{12}$$

and

$$\begin{aligned}&\Sigma_{\bar{\mathrm{n}}\,|\,\theta}(y,x,y',x'|\theta) \\ &\quad = N\big(\theta(y,x)\delta[y,y']\delta[x,x'] - \theta(y,x)\theta(y',x')\big)\ .\end{aligned} \tag{13}$$

1) Aggregation Properties: As performed for the model θ, a characterization of the function $\bar{\mathrm{n}}$ integrated over the set $\mathcal{Y}$ will be performed. Introduce the function $N'(x;D) = \sum_{y\in\mathcal{Y}}\bar{N}(y,x;D)$ and define the "marginalized" random process $\mathrm{n}' \equiv \sum_{y\in\mathcal{Y}}\bar{\mathrm{n}}(y,\cdot) \equiv N'(\mathrm{D})$ over the set $\mathcal{X}$. By the aggregation property of Multinomial random processes [4], the aggregation conditioned on the model is only dependent on θ via the marginal process θ' and is distributed as $\mathrm{n}'\,|\theta' \sim \mathrm{Multi}(N,\theta')$.

Also of interest is the distribution of $\bar{\mathrm{n}}$ conditioned on its aggregation n'. It can be shown that when conditioned on the model θ as well, the dependency of the distribution on θ is only expressed through the true predictive models $\tilde{\theta}(x)$. The PMF of interest is thus

$$\mathrm{P}_{\bar{\mathrm{n}}|\,\mathrm{n}',\tilde{\theta}}\left(\bar{n}|n',\tilde{\theta}\right) = \prod_{x\in\mathcal{X}}\left[\mathcal{M}\left(\bar{n}(\cdot,x)\right)\prod_{y\in\mathcal{Y}}\tilde{\theta}(y;x)^{\bar{n}(y,x)}\right] \tag{14}$$

for $\left\{\bar{n}\in{\mathbb{Z}_{\geq 0}}^{\mathcal{Y}\times\mathcal{X}} : \sum_{y\in\mathcal{Y}}\bar{n}(y,x) = n'(x),\quad \forall x\in\mathcal{X}\right\}$. Observe that conditioning on the aggregation process n' renders the function segments $\bar{\mathrm{n}}(\cdot,x)$ independent of one another and that they are also Multinomial, such that $\bar{\mathrm{n}}(\cdot,x)|\,\mathrm{n}'(x),\tilde{\theta} \sim \mathrm{Multi}\left(\mathrm{n}'(x),\tilde{\theta}(x)\right)$.

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

In this section, the Bayesian predictive PMF $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}$ is provided. As $\bar{N}(\mathrm{D})$ is a sufficient statistic for the training data, $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}$ will be derived instead, simplifying the analysis. Note that $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}(y|x,D) = \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}\left(y|x,\bar{N}(D)\right)$.

As mentioned in Section II, the Bayesian predictive PMF can be interpreted as the expectation of the true predictive PMF with respect to the model posterior distribution $\mathrm{p}_{\theta|\,\mathrm{x},\mathrm{D}}$. Using the notation detailed in Section III-A1, the PMF is now expressed as $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}} = \mathrm{E}_{\theta|\,\mathrm{x},\bar{\mathrm{n}}}\left[\,\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\theta}\,\right] = \mu_{\tilde{\theta}(\mathrm{x})|\,\mathrm{x},\bar{\mathrm{n}}}$, the conditional expectation of the predictive model $\tilde{\theta}(\mathrm{x})$.

It can be shown that since the conditional models $\tilde{\theta}(x)$ are independent of the marginal model θ', they are also conditionally independent of x given the statistic $\bar{\mathrm{n}}$. Thus, the conditional model posterior PDF is represented as

$$\begin{aligned}&\mathrm{p}_{\tilde{\theta}|\,\mathrm{x},\bar{\mathrm{n}}}\left(\tilde{\theta}|x,\bar{n}\right) = \mathrm{p}_{\tilde{\theta}|\,\bar{\mathrm{n}}}\left(\tilde{\theta}|\bar{n}\right) \\ &= \frac{\mathrm{P}_{\bar{\mathrm{n}}\,|\,\mathrm{n}',\tilde{\theta}}\left(\bar{n}|\sum_y\bar{n}(y,\cdot),\tilde{\theta}\right)}{\mathrm{P}_{\bar{\mathrm{n}}\,|\,\mathrm{n}'}\left(\bar{n}|\sum_y\bar{n}(y,\cdot)\right)}\,\mathrm{p}_{\tilde{\theta}}\left(\tilde{\theta}\right) \\ &= \prod_{x'\in\mathcal{X}}\left[\beta\big(\alpha(\cdot,x')+\bar{n}(\cdot,x')\big)^{-1}\prod_{y\in\mathcal{Y}}\tilde{\theta}(y;x')^{\alpha(y,x')+\bar{n}(y,x')-1}\right]\ .\end{aligned} \tag{15}$$

As the $|\,\mathcal{X}\,|$ Multinomial components of $\mathrm{P}_{\bar{\mathrm{n}}\,|\,\mathrm{n}',\tilde{\theta}}$ have exponential form, the corresponding Dirichlet components of the conditional model PDF $\mathrm{p}_{\tilde{\theta}}$ are their conjugate priors [5], [3]. Thus, the posterior is a product of Dirichlet distributions $\tilde{\theta}(x)|\,\mathrm{x},\bar{\mathrm{n}} \sim \mathrm{Dir}\left(\alpha(\cdot,x)+\bar{\mathrm{n}}(\cdot,x)\right)$, with each model $\tilde{\theta}(x)$ dependent solely on the sufficient statistic elements $\bar{\mathrm{n}}(\cdot,x)$.

The concentration parameters increase proportionately with the volume of training data. Thus as each $n'(x)\to\infty$, the posteriors converges to $\mathrm{p}_{\tilde{\theta}(x)|\,\bar{\mathrm{n}}(\cdot,x)}\left(\tilde{\theta}(x)|\bar{n}(\cdot,x)\right) \to \delta\big(\tilde{\theta}(x) - \bar{n}(\cdot,x)/\sum_y\bar{n}(y,x)\big)$ and the conditional models are identified. Conversely, as $\alpha'(x)\to\infty$, the confidence in the prior model increases and the posterior tends toward $\mathrm{p}_{\tilde{\theta}(x)|\,\bar{\mathrm{n}}(\cdot,x)}\left(\tilde{\theta}(x)|\bar{n}(\cdot,x)\right) \to \delta\big(\tilde{\theta}(x) - \alpha(\cdot,x)/\alpha'(x)\big)$, independent of the training data.

Figure 2 shows the influence of the training data on the model distribution; after conditioning on the training data (via $\bar{\mathrm{n}}$), the PDF concentration shifts away from the models favored by the prior knowledge and towards other models that better account for the observations.

Taking the mean with respect to the conditional model posterior, the Bayesian predictive PMF is

$$\begin{aligned}\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}} &= \mu_{\tilde{\theta}(\mathrm{x})|\,\mathrm{x},\bar{\mathrm{n}}} = \mu_{\tilde{\theta}(\mathrm{x})|\,\bar{\mathrm{n}}(\cdot,\mathrm{x})} \\ &= \left(\frac{\alpha'(\mathrm{x})}{\alpha'(\mathrm{x})+\sum_y\bar{\mathrm{n}}(y,\mathrm{x})}\right)\frac{\alpha(\cdot,\mathrm{x})}{\alpha'(\mathrm{x})} \\ &\quad + \left(\frac{\sum_y\bar{\mathrm{n}}(y,\mathrm{x})}{\alpha'(\mathrm{x})+\sum_y\bar{\mathrm{n}}(y,\mathrm{x})}\right)\frac{\bar{\mathrm{n}}(\cdot,\mathrm{x})}{\sum_y\bar{\mathrm{n}}(y,\mathrm{x})}\ .\end{aligned} \tag{16}$$

The last representation views the distribution as a convex combination of two conditional distributions. The first distribution $\alpha(\cdot,\mathrm{x})/\alpha'(\mathrm{x})$ is independent of the training data and based on the prior knowledge implied via the model PDF parameter; the second distribution is the conditional empirical PMF and depends on $\bar{\mathrm{n}}$, not on α. For both, only those values α and $\bar{\mathrm{n}}$ corresponding to the observed value x influence the distribution.

The weighting factors are dependent on these values as well. As $n'(x)/\alpha'(x)\to 0$, the PMF tends toward the data-independent distribution; as $n'(x)/\alpha'(x)\to\infty$, $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}$ tends towards the empirical conditional distribution.

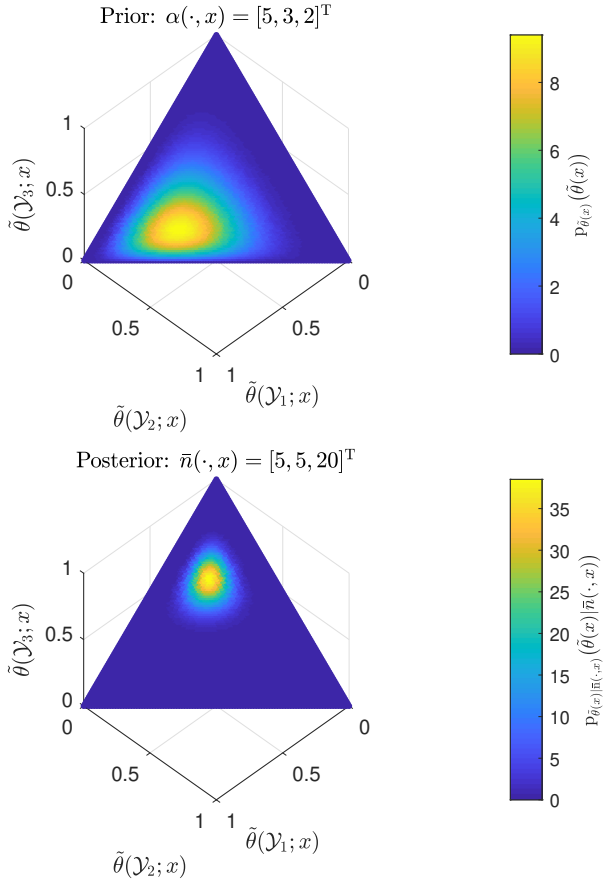

Fig. 2. Model $\tilde{\theta}(x)$ PDF, prior and posterior

IV. Density Estimation Perspective

This section compares the Bayesian predictive distribution $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}$ to the true predictive PMF $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\theta}$ and investigates the effects of prior knowledge. For a given x and corresponding number of training samples $\mathrm{n}'(\mathrm{x})$, the expected value of the estimate conditioned on the true model θ is

$$\begin{aligned} &\mathrm{E}_{\bar{\mathrm{n}}\,|\,\mathrm{n}',\theta}\left[\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}\right] \qquad (17)\\ &= \left(\frac{\alpha'(\mathrm{x})}{\alpha'(\mathrm{x})+\mathrm{n}'(\mathrm{x})}\right)\frac{\alpha(\cdot,\mathrm{x})}{\alpha'(\mathrm{x})} + \left(\frac{\mathrm{n}'(\mathrm{x})}{\alpha'(\mathrm{x})+\mathrm{n}'(\mathrm{x})}\right)\tilde{\theta}(\mathrm{x})\ , \end{aligned}$$

where the properties of a multinomial distribution conditioned on its aggregation have been used. The result is a convex combination of the conditional data-independent distribution $\alpha(\cdot, x)/\alpha'(x)$ and the true conditional distribution $\tilde{\theta}(x)$. The convex coefficients are inherited from $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}$; note that as the number of matching training samples $\mathrm{n}'(x)$ increases relative to $\alpha'(x)$, the estimate tends towards the true conditional PMF.

To aid characterization of the predictive PMF estimator, define the random process $\Delta(\mathrm{x}, \bar{\mathrm{n}}, \theta) \equiv \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}} - \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\theta} \in \mathbb{R}^{\mathcal{Y}}$. For a given x and corresponding number of training samples $\mathrm{n}'(\mathrm{x})$, the bias of the conditional PMF estimate is

$$\begin{aligned} \mathrm{Bias}(\mathrm{x}, \mathrm{n}', \theta) &= \mathrm{E}_{\bar{\mathrm{n}}\,|\,\mathrm{n}',\theta}\left[\Delta(\mathrm{x}, \bar{\mathrm{n}}, \theta)\right] \qquad (18)\\ &= \frac{\alpha'(\mathrm{x})}{\alpha'(\mathrm{x})+\mathrm{n}'(\mathrm{x})}\left(\frac{\alpha(\cdot,\mathrm{x})}{\alpha'(\mathrm{x})} - \tilde{\theta}(\mathrm{x})\right) \end{aligned}$$

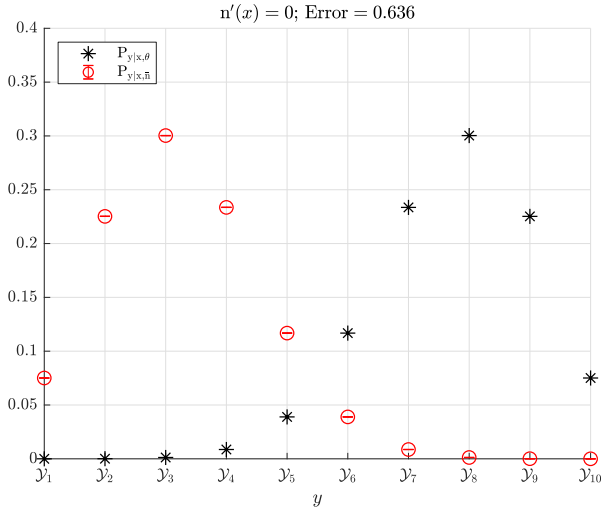

Fig. 3. Model $\tilde{\theta}(x)$ estimate, no training data

and its covariance function is

$$\begin{aligned} \mathrm{Cov}(y, y'; \mathrm{x}, \mathrm{n}', \theta) &= \mathrm{C}_{\bar{\mathrm{n}}\,|\,\mathrm{n}',\theta}\left[\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}(\cdot\,|\,\mathrm{x}, \bar{\mathrm{n}})\right](y, y') \qquad (19)\\ &= \frac{\Sigma_{\bar{\mathrm{n}}(\cdot,\mathrm{x})|\,\mathrm{n}'(\mathrm{x}),\tilde{\theta}(\mathrm{x})}(y, y')}{\left(\alpha'(\mathrm{x})+\mathrm{n}'(\mathrm{x})\right)^2}\\ &= \frac{\mathrm{n}'(\mathrm{x})}{\left(\alpha'(\mathrm{x})+\mathrm{n}'(\mathrm{x})\right)^2}\left(\tilde{\theta}(y;\mathrm{x})\delta[y, y'] - \tilde{\theta}(y;\mathrm{x})\tilde{\theta}(y';\mathrm{x})\right)\ , \end{aligned}$$

where the properties of multinomial random processes have been used. Note that the bias is proportionate to the difference between the true conditional model and the data-independent estimate. The scaling factor tends to zero as $\mathrm{n}'(x)/\alpha'(x) \to \infty$; as such, more informative priors (large $\alpha'(x)$) will lead to PMF estimates that are prone to bias. Conversely, the variance of the PMF estimate decreases with increasing $\alpha'(x)$.

Combining the estimator bias and variance, the conditional second moments of $\Delta(\mathrm{x}, \bar{\mathrm{n}}, \theta)$ are

$$\begin{aligned} \mathcal{E}(y, y'; \mathrm{x}, \mathrm{n}', \theta) &= \mathrm{E}_{\bar{\mathrm{n}}\,|\,\mathrm{n}',\theta}\left[\Delta(y; \mathrm{x}, \bar{\mathrm{n}}, \theta)\Delta(y'; \mathrm{x}, \bar{\mathrm{n}}, \theta)\right] \qquad (20)\\ &= \mathrm{Bias}(y; \mathrm{x}, \mathrm{n}', \theta)\mathrm{Bias}(y'; \mathrm{x}, \mathrm{n}', \theta) + \mathrm{Cov}(y, y'; \mathrm{x}, \mathrm{n}', \theta)\ . \end{aligned}$$

As $\mathrm{n}'(x) \to \infty$, this function tends to zero and thus the underlying model $\tilde{\theta}(x)$ is determined exactly. A more practical case is estimation with a finite volume of training data. Specification of the Dirichlet model prior can be interpreted as providing a distribution estimate $\alpha(\cdot, x)/\alpha'(x)$ and a degree of confidence $\alpha'(x)$. Higher confidence reduces error due to the variance of the estimator, but increases the error due to bias between the true model and its estimate; low confidence renders the estimate unbiased, but increases the estimator variance.

To exemplify how the model estimate $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\mathrm{D}}$ approximates $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\theta}$, consider a scenario with $|\mathcal{Y}| = 10$. The data-independent PMF $\alpha(\cdot, x)/\alpha'(x)$ and true model $\tilde{\theta}(x)$ are shown in Figure 3 - note the significant mismatch.

Figures 4 and 5 show how the bias and variance of the estimate change for different values of $\mathrm{n}'(x)$ and $\alpha'(x)$. The

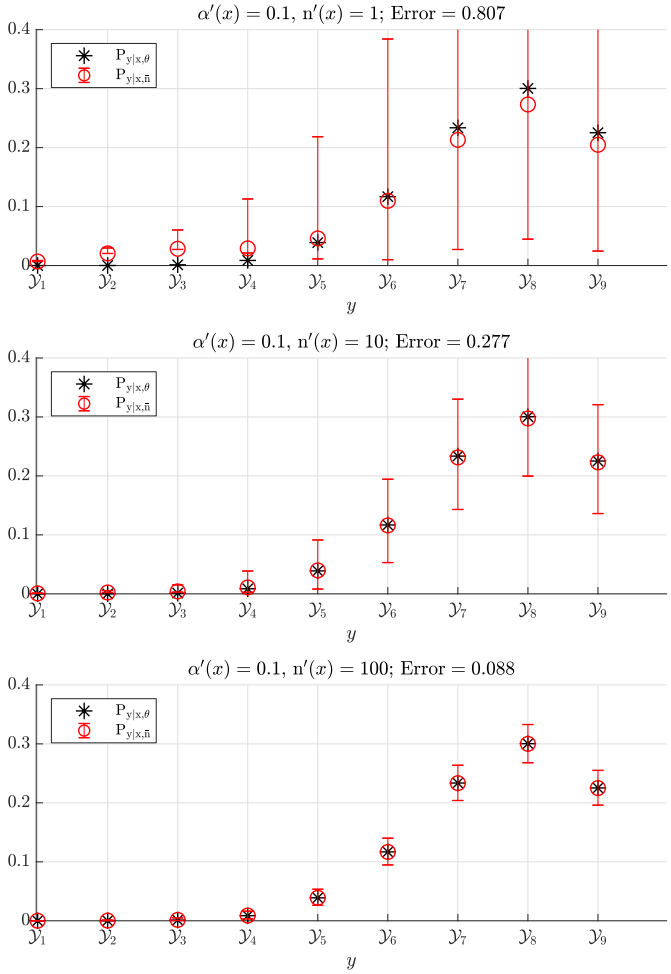

Fig. 4. Model $\tilde{\theta}(x)$ estimates, $\alpha'(x) = 0.1$

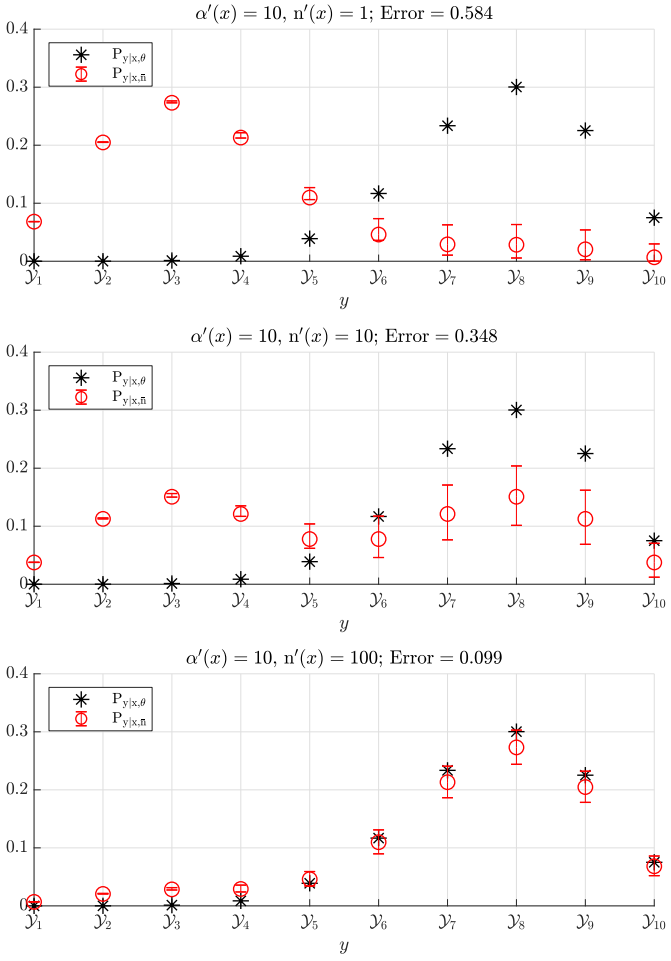

Fig. 5. Model $\tilde{\theta}(x)$ estimates, $\alpha'(x) = 10$

blue markers indicate the conditional mean of the estimator, $\mathrm{E}_{\bar{\mathrm{n}}\,|\,\mathrm{n}',\theta}\left[\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}(y|\,\mathrm{x},\bar{\mathrm{n}})\right]$; the upper and lower error bars indicate the square-root of the expected squared deviation above and below the conditional mean, respectively. Each individual plot heading provides the error $\sqrt{\sum_{y\in\mathcal{Y}}\mathcal{E}(y,y;\mathrm{x},\mathrm{n}',\theta)}$ to assess the quality of the PMF estimate.

Observe that for $\mathrm{n}'(x) = 1$, the high variance of the $\alpha'(x) = 0.1$ estimate (favoring the empirical PMF) renders it worse than the $\alpha'(x) = 10$ estimate; in fact, the variance is so high that the error exceeds that of the data-independent estimate $\alpha(\cdot,x)/\alpha'(x)$ (Figure 3). Conversely, for $\mathrm{n}'(x) = 10$, the confidence in the $\alpha'(x) = 10$ estimate leads to high bias and the $\alpha'(x) = 0.1$ estimate is superior. For $\mathrm{n}'(x) = 100$, both the $\alpha'(x) = 0.1$ and $\alpha'(x) = 10$ estimates begin converging to the true distribution - this is guaranteed due to the full support of the Dirichlet prior.

V. Conclusions

This article has shown how a Dirichlet prior distribution may be used for a Bayesian approach to machine learning prediction. An analysis of how well the Bayesian predictive distributions match the true predictive distributions has been performed for a variety of different non-informative and informative priors.

The conditional second moments of the difference between the two predictive PMF's have important applications for Bayesian regression, specifically for determining the expected squared-error loss. This will be a primary focus of future work.

References

[1] T. S. Ferguson, "A Bayesian analysis of some nonparametric problems," *The Annals of Statistics*, vol. 1, no. 2, pp. 209–230, 1973.

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

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

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

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