U.S. NAVAL RESEARCH LABORATORY

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Bayesian Learning for Classification using a Uniform Dirichlet Prior

Paul Rademacher[1] **Miloš Doroslovački**[2]

[1] U.S. Naval Research Laboratory
Radar Division

[2] The George Washington University
Department of Electrical and Computer Engineering

November 13, 2019

Introduction

Part I

Bayesian approaches to machine learning attempt to make better decisions by exploiting *prior knowledge* regarding the data-generating distribution:

Informative

- If the prior is localized around the true data-generating model, low-risk decisions can be made even with limited training data
- Priors that assign low weighting to the true model may not be able to realize satisfactory performance

Non-Informative

- Learners designed with approximately uniform priors will not perform as well as those made with well-selected informative priors
- Avoid high risk inherent to learners made by mismatched informative priors

Introduction

Part II

- Often, priors are termed non-informative as long as they are approximately uniform over their limited support. For example, a parametric regression function might use a high covariance Gaussian vector prior to characterize a subset of probability distributions
- 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* for independent, identically distributed observations and leads to a closed-form model posterior distribution

U.S. NAVAL RESEARCH LABORATORY

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Objective and Bayesian Setup

Data Model

Observable random element: $\mathrm{x} \in \mathcal{X}$
Unobservable random element: $\mathrm{y} \in \mathcal{Y}$
Observable training data: $\mathrm{D} \in \mathcal{D} = \{\mathcal{Y} \times \mathcal{X}\}^N$

Independently, identically 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 $\mathrm{P}_{\mathrm{y},\mathrm{x}\,|\theta}(y, x|\theta) = \mathrm{P}_{\mathrm{D}_n\,|\theta}(y, x|\theta) = \theta(y, x)$.

Alternate Notation: $\theta \Leftrightarrow \left(\theta', \tilde{\theta}\right)$

- Marginal model $\theta' \equiv \sum_{y \in \mathcal{Y}} \theta(y, \cdot) = \mathrm{P}_{\mathrm{x}\,|\theta}$ over the set $\mathcal{X}$
- Conditional models $\tilde{\theta}(\mathrm{x}) \equiv \theta(\cdot, \mathrm{x})/\theta'(x) = \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\theta}$ over the set $\mathcal{Y}$

Sufficient Statistic

Training data PMF:

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

is expressed via
$\bar{N}(y,x;D) = \sum_{n=1}^{N} \delta\big[(y,x), D_n\big]$, with range

$$\bar{\mathcal{N}} = \left\{ \bar{n} \in \mathbb{Z}_{\geq 0}{}^{\mathcal{Y}\times\mathcal{X}} : \sum_{y\in\mathcal{Y}} \sum_{x\in\mathcal{X}} \bar{n}(y,x) = N \right\}$$

- Empirical count $\bar{N}(\mathrm{D})$ is a *sufficient statistic* for the model θ
- $|\bar{\mathcal{N}}| = \binom{N+|\,\mathcal{Y}\,||\,\mathcal{X}\,|-1}{|\,\mathcal{Y}\,||\,\mathcal{X}\,|-1} \leq |\,\mathcal{D}\,|$

  $\Rightarrow$ **Efficient Transform**

$\Downarrow \quad \Downarrow$

Compact Representation

Express distributions using new random process $\bar{\mathrm{n}} \equiv \bar{N}(\mathrm{D}) \in \bar{\mathcal{N}}$

Objective

Design Metric

User-selected function $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}_{\geq 0}$. Most commonly selected for classification is the 0–1 loss:

$$\mathcal{L}(h, y) = 1 - \delta[h, y]$$

- All misclassifications penalized equally

Design Task

Create a classification function $f : \bar{\mathcal{N}} \mapsto \mathcal{Y}^{\mathcal{X}}$ that minimizes the conditional expected loss, or conditional "risk",

$$\begin{aligned}\mathcal{R}_{\Theta}(f; \theta) &= \mathrm{E}_{\mathrm{x},\bar{\mathrm{n}}\,|\,\theta}\left[\mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\theta}\left[\mathcal{L}\left(f(\mathrm{x}; \bar{\mathrm{n}}), \mathrm{y}\right)\right]\right] \\ &= 1 - \mathrm{E}_{\mathrm{x},\bar{\mathrm{n}}\,|\,\theta}\left[\tilde{\theta}\big(f(\mathrm{x}; \bar{\mathrm{n}}); \mathrm{x}\big)\right]\end{aligned}$$

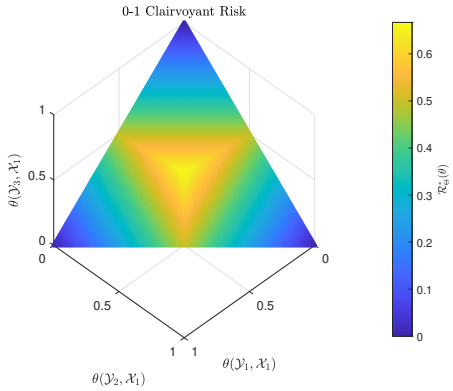
Clairvoyant Hypothesis and Risk

The "clairvoyant" classifier $f_\theta : \Theta \mapsto \mathcal{Y}^{\mathcal{X}}$ is maximum *a posteriori* (MAP):

$$f_\Theta(\mathrm{x};\theta) = \arg\min_{h\in\mathcal{Y}} \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\theta}\left[1-\delta[h,y]\right]$$
$$\equiv \arg\max_{y\in\mathcal{Y}} \tilde{\theta}(y;\mathrm{x})$$

$$\Downarrow \quad \Downarrow$$

$$\mathcal{R}^*_\Theta(\theta) \equiv 1 - \sum_{x\in\mathcal{X}} \theta'(x) \max_{y\in\mathcal{Y}} \tilde{\theta}(y;x)$$



Clairvoyant Risk = Lower Bound for Conditional Risk

Bayesian Inference

⇓ ⇓ ***Model Unknown. Select Prior*** p_{θ} ⇓ ⇓

$$\mathcal{R}(f) = \mathrm{E}_{\theta}\left[\mathcal{R}_{\Theta}(f;\theta)\right] = 1 - \mathrm{E}_{\mathrm{x},\bar{\mathrm{n}}}\left[\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}\left(f(\mathrm{x};\bar{\mathrm{n}})\right)\right]$$

Optimal classifier: Bayesian MAP

$$f^{*}(\mathrm{x};\bar{\mathrm{n}}) = \arg\max_{y\in\mathcal{Y}} \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}(y|\,\mathrm{x},\bar{\mathrm{n}})$$

Minimum Bayes Probability of Error:

$$\mathcal{R}(f^{*}) = 1 - \mathrm{E}_{\mathrm{x},\bar{\mathrm{n}}}\left[\max_{y\in\mathcal{Y}} \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}\left(y|\,\mathrm{x},\bar{\mathrm{n}}\right)\right]$$

Predictive Distributions

Bayesian PMF is the conditional expectation of the true PMF $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\theta} \equiv \tilde{\theta}(\mathrm{x})$:

$$\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}} = \mathrm{E}_{\theta|\,\mathrm{x},\bar{\mathrm{n}}}\left[\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\theta}\right] \equiv \mu_{\tilde{\theta}(\mathrm{x})|\,\mathrm{x},\bar{\mathrm{n}}}$$

U.S. NAVAL RESEARCH LABORATORY

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Distributions: Prior to Predictive

Dirichlet Prior PDF

The probability density function (PDF) for the model random process $\theta \in \Theta$ is Dirichlet with parameters $\alpha(\cdot,\cdot) = 1$:

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

$\Rightarrow$

$\Rightarrow$

Conditional PDF of $\tilde{\theta}$ given θ' is

$$\begin{aligned} \mathrm{p}_{\tilde{\theta}|\theta'}\left(\tilde{\theta}|\theta'\right) &= \mathrm{p}_{\tilde{\theta}}\left(\tilde{\theta}\right) \\ &= \prod_{x\in\mathcal{X}}\mathrm{Dir}\left(\tilde{\theta}(x);\alpha(\cdot,x)\right)\Big|_{\alpha(\cdot,x)=1} \\ &= \left[\left(|\,\mathcal{Y}\,|-1\right)!\right]^{|\,\mathcal{X}\,|} \end{aligned}$$

True predictive distribution $\tilde{\theta}$ is independent of the marginal model and inherits a uniform PDF

Model Posterior PDF

Closed-Form

- Independence of $\tilde{\theta}$ from θ' implies conditional independence of $\tilde{\theta}$ from x given $\bar{\mathrm{n}}$
- As the data PMF $\mathrm{P}_{\bar{\mathrm{n}}\,|\theta}$ has exponential form, the Dirichlet PDF is a conjugate prior

$$\begin{aligned}\mathrm{p}_{\tilde{\theta}|\,\mathrm{x},\bar{\mathrm{n}}}\left(\tilde{\theta}|x,\bar{n}\right) &= \mathrm{p}_{\tilde{\theta}|\,\bar{\mathrm{n}}}\left(\tilde{\theta}|\bar{n}\right) \\ &= \prod_{x'\in\mathcal{X}} \mathrm{Dir}\left(\tilde{\theta}(x');\alpha(\cdot,x')\right)\Big|_{\alpha(\cdot,x')=1+\bar{n}(\cdot,x')}\end{aligned}$$

Posterior for model $\tilde{\theta}(x)$ is Dirichlet and dependent solely on the sufficient statistic elements $\bar{\mathrm{n}}(\cdot,x)$

Model Posterior PDF

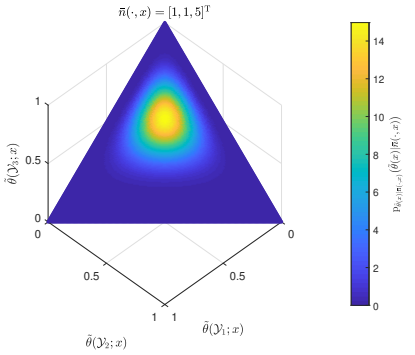
Asymptotic Trend

- Covariance of Dirichlet $\mathrm{p}_{\tilde{\theta}(x)|\,\bar{\mathrm{n}}(\cdot,x)}$ decreases monotonically with concentration $\alpha'(x) \equiv \sum_{y\in\mathcal{Y}} \alpha(y,x)$
- Localized around $\mu_{\tilde{\theta}(x)|\,\bar{\mathrm{n}}(\cdot,x)} = \alpha(\cdot,x)/\alpha'(x)$

$$\Downarrow \quad \Downarrow$$

As $n'(x) \equiv \sum_{y\in\mathcal{Y}} \bar{n}(y,x) \to \infty$, the posteriors converge to

$$\mathrm{p}_{\tilde{\theta}(x)|\,\bar{\mathrm{n}}(\cdot,x)}\left(\tilde{\theta}(x)|\bar{n}(\cdot,x)\right) \to \delta\left(\tilde{\theta}(x) - \frac{\bar{n}(\cdot,x)}{n'(x)}\right)$$



Asymptotically consistent estimation of $\tilde{\theta}$ due to full support of prior

Bayesian Predictive PMF

The Bayesian predictive PMF is a convex combination of two conditional distributions:

$$\begin{aligned} \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}(\cdot|x,\bar{n}) &= \mu_{\tilde{\theta}(x)|\,\bar{\mathrm{n}}(\cdot,x)}\big(\bar{n}(\cdot,x)\big) \\ &\equiv \left(\frac{|\,\mathcal{Y}\,|}{n'(x)+|\,\mathcal{Y}\,|}\right)\frac{1}{|\,\mathcal{Y}\,|} + \left(\frac{n'(x)}{n'(x)+|\,\mathcal{Y}\,|}\right)\frac{\bar{n}(\cdot,x)}{n'(x)} \end{aligned}$$

Prior Mean Uniform PMF $\mu_{\tilde{\theta}(x)} = |\,\mathcal{Y}\,|^{-1}$ dependent only on the Dirichlet parameterization

Conditional Empirical PMF $\bar{n}(\cdot,x)/n'(x)$ dependent only on training data

As the number of training data $n'(x)$ increases relative to the number of classes $|\,\mathcal{Y}\,|$, the predictive distribution tends toward the empirical PMF

U.S. NAVAL RESEARCH LABORATORY

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Bayesian Classifier and Error Trends

Bayesian Classifier and Risk

Optimal Hypothesis: the *conditional majority decision*

$$f^*(x;\bar{n}) = \arg\max_{y\in\mathcal{Y}} \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\bar{\mathrm{n}}}\left(y|x,\bar{n}\right) = \arg\max_{y\in\mathcal{Y}} \bar{n}(y,x)$$

Minimum Expected Probability of Error:

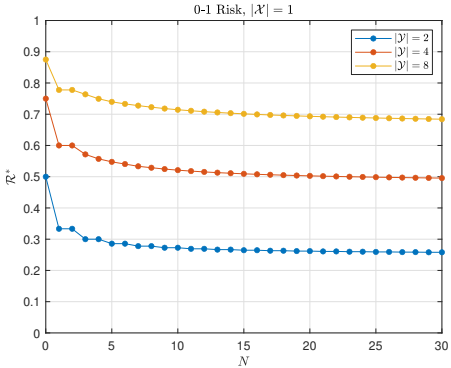
$$\mathcal{R}^* = 1 - \mathrm{E}_{\mathrm{x,D}}\left[\max_{y\in\mathcal{Y}} \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x,D}}(y|\,\mathrm{x},\mathrm{D})\right] = 1 - \sum_{x\in\mathcal{X}} \frac{\mathrm{E}_{\bar{\mathrm{n}}}\left[\max_{y\in\mathcal{Y}} \bar{\mathrm{n}}(y,x)\right] + 1}{|\,\mathcal{Y}\,||\,\mathcal{X}\,|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

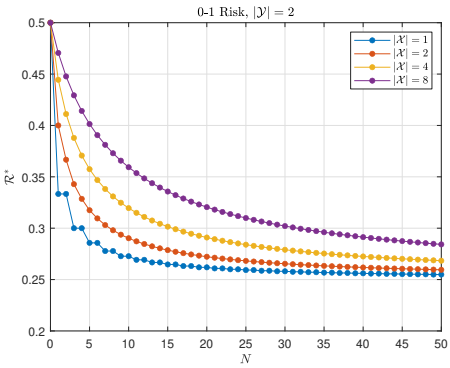
$$= 1 - \frac{\sum_{m=1}^{|\mathcal{Y}|}\binom{|\mathcal{Y}|}{m}(-1)^{m-1}\sum_{n=0}^{\left\lfloor\frac{N}{m}\right\rfloor}\prod_{l=1}^{|\mathcal{Y}||\mathcal{X}|-1}\left(1-\frac{mn}{N+l}\right)}{|\,\mathcal{Y}\,| + N/|\,\mathcal{X}\,|}$$

∗ *Efficient formula derived using Inclusion-Exclusion principle*

Probability of Error Trends

with Class/Data Set Sizes





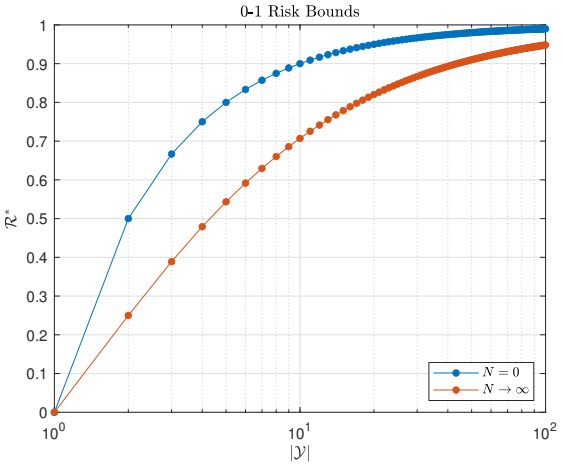
- Larger class sets $\mathcal{Y}$ raise the lower bound on probability of error
- Larger observation set $\mathcal{X} \longrightarrow$ more data N required to achieve the same level of performance

Probability of Error Trends

with Training Data Volume

- For binary classification, infinite training data only reduces the expected probability of error from 0.5 to 0.25
- As $|\mathcal{Y}|$ increases, the probability of error tends to unity and any improvement due to training data becomes negligible

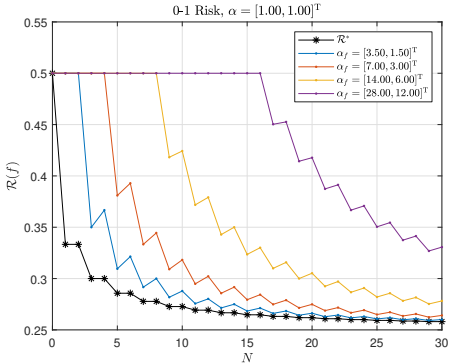
N	$\mathcal{R}^*$
0	$1-\|\mathcal{Y}\|^{-1}$
$\to\infty$	$1-\|\mathcal{Y}\|^{-1}\sum_{m=1}^{\|\mathcal{Y}\|} m^{-1}$

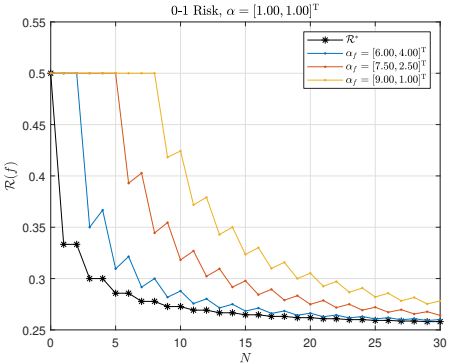


Comparison to Informative Classifiers

Bayes Risk

Classifiers derived from informative Dirichlet priors with $\alpha'_f(x) \gg |\mathcal{Y}||\mathcal{X}|$ prioritize the prior mean $\mu_{\tilde{\theta}} = \alpha_f(\cdot, x)/\alpha'_f(x)$ for prediction

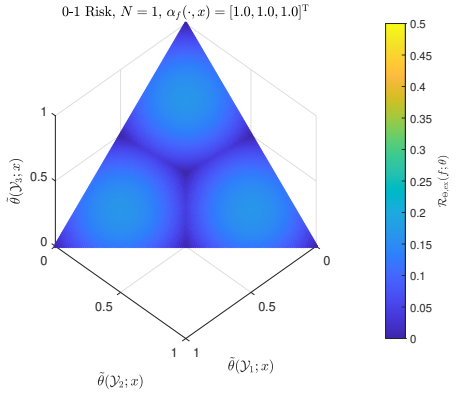


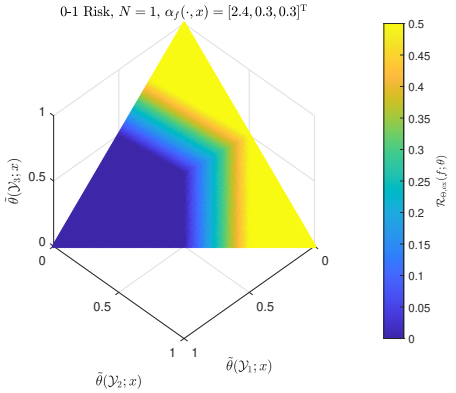


Slower adaptation ⇒ Higher Bayes probability of error

Comparison to Informative Classifiers

Excess Conditional Risk





Trade-Off

Uniform prior provides a robust classifier for all unknown models θ, limiting the space of high error models. However, fewer models achieve the clairvoyant risk.

Conclusions

- The majority decision classifier designed with a uniform Dirichlet prior minimizes the possibility of maximal error for applications where the data distribution can not be adequately modeled
- Full support of Dirichlet priors guarantees minimal risk in the limit of training data volume
- Efficient closed-form for Bayesian probability of error provides a lower-bound for general classifiers
- Low localization Dirichlet priors result in the same classifier - these priors are sensible for recognition applications where humans perform with low-error