Bayesian Learning for Regression using Dirichlet Prior Distributions of Varying Localization

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

July 11, 2021

Introduction

Bayesian Learning

Bayesian approaches to statistical 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 minimally localized priors respond strongly to training data, avoiding the drawbacks of misinformed prior knowledge
- If the data volume is limited, high variance “overfit” solutions can occur

The Dirichlet Prior

Dirichlet prior distributions have a number of desirable properties:

- Full support over the space of data-generating distributions, guaranteeing *consistent estimation* of the true data model
- They are conjugate priors for independent, identically distributed observations[1], leading to *closed-form* posterior distributions
- Flexible parameterization enabling *both* maximally and minimally informative priors and thus a wide range of learning solutions

[1] Thomas S. Ferguson. "A Bayesian Analysis of Some Nonparametric Problems". In: *The Annals of Statistics* (1973).

U.S. NAVAL RESEARCH LABORATORY

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Data Model and Regression Objective

Data Representation

Observable random variable: $\mathrm{x} \in \mathcal{X} \subset \mathbb{R}$
Unobservable random variable: $\mathrm{y} \in \mathcal{Y} \subset \mathbb{R}$
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: $\theta \Leftrightarrow \left(\theta_{\mathrm{m}}, \theta_{\mathrm{c}} \right)$

- Marginal model $\theta_{\mathrm{m}} \equiv \sum_{y \in \mathcal{Y}} \theta(y, \cdot) = \mathrm{P}_{\mathrm{x}\,|\,\theta} \equiv \mathrm{P}_{\mathrm{x}\,|\,\theta_{\mathrm{m}}}$
- Conditional models $\theta_{\mathrm{c}}(x) \equiv \theta(\cdot, x) / \theta_{\mathrm{m}}(x) = \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\theta} \equiv \mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\theta_{\mathrm{c}}}$

Sufficient Statistic

Transform

Using the i.i.d. assumption,

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

where data are represented using $\Psi : \mathcal{D} \mapsto \Psi \subset \Theta$, defined as

$$\Psi(y,x;D) = N^{-1}\sum_{n=1}^{N}\delta\big[(y,x),D_n\big] \; .$$

- Empirical distribution $\Psi(\mathrm{D})$ is a sufficient statistic[2] for the model θ
- Efficient: $|\Psi| = \binom{N+|\mathcal{Y}||\mathcal{X}|-1}{|\mathcal{Y}||\mathcal{X}|-1} \leq |\mathcal{D}|$

$\Rightarrow$ **Represent data using new random process** $\psi \equiv \Psi(\mathrm{D}) \in \Psi$

[2] Bernardo et al., *Bayesian Theory*.

Sufficient Statistic

Distribution

- Conditioned on the true model, the data statistic is an "Empirical" random process $\psi|\theta \sim \mathrm{Emp}(N, \theta)$
  – Equivalent to a normalized multinomial random process[3]
- As $N \to \infty$, the random process converges to $\psi|\theta \xrightarrow{p} \theta$
  $\Rightarrow$ Use enables consistent estimation of model

Alternate Notation: $\psi \Leftrightarrow (\psi_{\mathrm{m}}, \psi_{\mathrm{c}})$

- Marginal $\psi_{\mathrm{m}} \equiv \sum_{y \in \mathcal{Y}} \psi(y, \cdot)$
- Conditional $\psi_{\mathrm{c}}(x) \equiv \psi(\cdot, x) / \psi_{\mathrm{m}}(x)$

By the aggregation property [4],

- $\psi_{\mathrm{m}} \,|\, \theta_{\mathrm{m}} \sim \mathrm{Emp}(N, \theta_{\mathrm{m}})$
- $\psi_{\mathrm{c}}(x) |\, \psi_{\mathrm{m}}(x), \theta_{\mathrm{c}}(x) \sim \mathrm{Emp}\left(N \psi_{\mathrm{m}}(x), \theta_{\mathrm{c}}(x)\right)$ are mutually independent

---

[3] Thomas P. Minka. *Bayesian inference, entropy, and the multinomial distribution*. Tech. rep. Microsoft Research, 2003.

[4] Johnson et al., *Discrete Multivariate Distributions*.

Objective

- Design a regression function $f : \Psi \mapsto \mathbb{R}^{\mathcal{X}}$ to minimize the expected squared-error with respect to θ:

$$\mathcal{R}_\Theta(f;\theta) = \mathrm{E}_{y,x,\psi|\theta}\left[\left(f(x;\psi) - y\right)^2\right] \equiv \underbrace{\mathrm{E}_{x\,|\,\theta_m}\left[\Sigma_{y\,|\,x,\theta_c}\right]}_{\mathcal{R}^*_\Theta(\theta)} + \underbrace{\mathrm{E}_{x,\psi|\theta}\left[\left(f(x;\psi) - \mu_{y\,|\,x,\theta_c}\right)^2\right]}_{\mathcal{R}_{\Theta,\mathrm{ex}}(f;\theta)}$$

- Clairvoyant[5] regressor $f_\Theta(x;\theta_c) = \mu_{y\,|\,x,\theta_c}$ achieves *irreducible* squared-error $\mathcal{R}^*_\Theta(\theta)$
- Excess squared-error can be decomposed into bias and variance terms:

$$\mathcal{R}_{\Theta,\mathrm{ex}}(f;\theta) \equiv \mathrm{E}_{x\,|\,\theta_m}\left[\left(\mathrm{E}_{\psi|\theta}\left[f(x;\psi)\right] - f_\Theta(x;\theta_c)\right)^2 + \mathrm{C}_{\psi|\theta}\left[f(x;\psi)\right]\right]$$

[5] Steven M. Kay. *Fundamentals of Statistical Signal Processing: Detection Theory*. Vol. 2. Prentice-Hall, 1998.

Bayesian Inference

Model unknown. Select prior p_θ and formulate Bayesian risk:

$$\mathcal{R}(f) = \mathrm{E}_\theta\left[\mathcal{R}_\Theta(f;\theta)\right] = \mathrm{E}_{\mathrm{y,x},\psi}\left[\left(f(\mathrm{x};\psi) - \mathrm{y}\right)^2\right]$$

$$\Downarrow \quad \Downarrow \quad \Downarrow \quad \Downarrow$$

Bayes optimal regressor:

$$f^*(\mathrm{x};\psi) = \arg\min_{y'\in\mathbb{R}} \mathrm{E}_{\mathrm{y}\,|\,\mathrm{x},\psi}\left[(y'-\mathrm{y})^2\right] = \mu_{\mathrm{y}\,|\,\mathrm{x},\psi}$$

- Observe that $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\psi} = \mathrm{E}_{\theta|\,\mathrm{x},\psi}\left[\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x},\theta}\right] \equiv \mu_{\theta_c(\mathrm{x})|\,\mathrm{x},\psi}$

Bayesian distribution is the posterior mean[6] of the predictive model θ_c

[6] Kevin P. Murphy. *Binomial and multinomial distributions*. Tech. rep. University of British Columbia, 2006.

Distributions: Prior to Predictive

Dirichlet Prior

The probability density function of the model $\theta \in \Theta$ is Dirichlet:

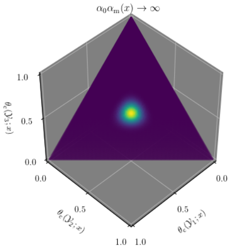
$$\mathrm{p}_\theta(\theta) = \mathrm{Dir}(\theta; \alpha_0, \alpha) \equiv \beta(\alpha_0 \alpha)^{-1} \prod_{y \in \mathcal{Y}} \prod_{x \in \mathcal{X}} \theta(y, x)^{\alpha_0 \alpha(y,x) - 1}$$

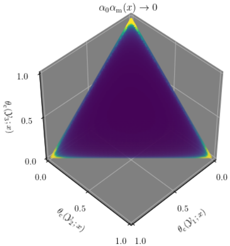
- Parameter α_0 controls localization around mean α

Alternate Notation:

- Marginal $\alpha_\mathrm{m} \equiv \sum_{y \in \mathcal{Y}} \alpha(y, \cdot)$
- Conditional $\alpha_\mathrm{c}(x) \equiv \alpha(\cdot, x) / \alpha_\mathrm{m}(x)$

∗ By the aggregation property[7], $\theta_\mathrm{m} \sim \mathrm{Dir}(\alpha_0, \alpha_\mathrm{m})$ and $\theta_\mathrm{c}(x) \sim \mathrm{Dir}\left(\alpha_0\, \alpha_\mathrm{m}(x), \alpha_\mathrm{c}(x)\right)$ are mutually independent





[7] Ferguson, "A Bayesian Analysis of Some Nonparametric Problems".

Predictive Model Posterior

Closed-Form

Since $\perp\!\!\!\perp_x \theta_c(x)$ and $\theta_c \perp\!\!\!\perp \theta_m$, and since the Empirical process $\theta_c(x) | \psi_m(x), \psi_c(x)$ has exponential form, Dirichlet process $\theta_c(x)$ is conjugate[8] and thus

$$\theta_c(x) | \psi_m(x), \psi_c(x) \sim \mathrm{Dir}\left(\alpha_0\, \alpha_m(x) + N\, \psi_m(x), \mu_{\theta_c(x) | \psi_m(x), \psi_c(x)}\right) ,$$

with mean functions

$$\mu_{\theta_c(x) | \psi_m(x), \psi_c(x)} = \gamma(x; \psi_m)\, \alpha_c(x) + \big(1 - \gamma(x; \psi_m)\big)\, \psi_c(x) \equiv \mathrm{P}_{y\,|\,x,\psi} ,$$

where $\gamma(x; \psi_m) = \Big(1 + N\, \psi_m(x) / \big(\alpha_0\, \alpha_m(x)\big)\Big)^{-1} \in (0, 1]$.

Bayesian predictions mix prior mean α_c with empirical distribution ψ_c

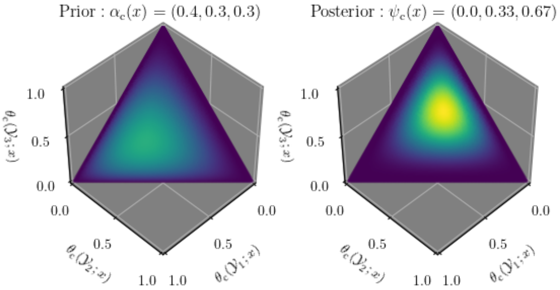
[8] Sergios Theodoridis. *Machine Learning: A Bayesian and Optimization Perspective*. Elsevier, 2015.

Predictive Model Posterior

Trends

- As localization α_0 increases, $\theta_c(x) | \psi_m(x), \psi_c(x) \xrightarrow{p} \alpha_c(\mathrm{x})$ and the prior is emphasized

- As training volume N increases, $\theta_c(x) | \psi_m(x), \psi_c(x) \xrightarrow{p} \psi_c(x)$ and data is emphasized
  - Since $\psi_c \,|\, \theta_c \xrightarrow{p} \theta_c$, the true predictive model is identified



Full support prior ensures consistent estimation of model

Bayes Optimal Regressor

Regressor: $f^*(\mathrm{x};\psi) \equiv \gamma(\mathrm{x};\psi_\mathrm{m})\mu_{\mathrm{y}\,|\,\mathrm{x}} + \big(1-\gamma(\mathrm{x};\psi_\mathrm{m})\big)\sum_{y\in\mathcal{Y}}\psi_\mathrm{c}(y;\mathrm{x})\ y$

- $*$ Convexly combines first moment of $\mathrm{P}_{\mathrm{y}\,|\,\mathrm{x}} = \mu_{\theta_\mathrm{c}(\mathrm{x})} = \alpha_\mathrm{c}(\mathrm{x})$ with empirical mean
- $*$ Inherits trends from posterior distribution, allowing maximal or minimal confidence in the prior

---

Excess SE: $\mathcal{R}_{\Theta,\mathrm{ex}}(f^*;\theta) \equiv \mathrm{E}_{\mathrm{x}\,|\,\theta_\mathrm{m}}\left[\lambda_{\text{Bias}}(\mathrm{x};\theta_\mathrm{m})\left(\mu_{\mathrm{y}\,|\,\mathrm{x}} - \mu_{\mathrm{y}\,|\,\mathrm{x},\theta_\mathrm{c}}\right)^2 + \lambda_{\text{Var}}(\mathrm{x};\theta_\mathrm{m})\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta_\mathrm{c}}\right]$

- $*$ $\lambda_{\text{Bias}}(x;\theta_\mathrm{m}) = \mathrm{E}_{\psi_\mathrm{m}\,|\,\theta_\mathrm{m}}\left[\gamma(x;\psi_\mathrm{m})^2\right]$ and $\lambda_{\text{Var}}(x;\theta_\mathrm{m}) = \mathrm{E}_{\psi_\mathrm{m}\,|\,\theta_\mathrm{m}}\left[\frac{\left(1-\gamma(x;\psi_\mathrm{m})\right)^2}{N\ \psi_\mathrm{m}(x)}\right]$

- Bias: proportionate to squared-difference between data-independent regressor $\mu_{\mathrm{y}\,|\,\mathrm{x}}$ and clairvoyant regressor
- Variance: proportionate to the predictive variance, adding to the irreducible risk

Trends and Results

Example

Data Model:

- $\mathcal{X} = \mathcal{Y} = \{i/127 : i = 0, \ldots, 127\}$
- $\theta_{\mathrm{m}} = |\,\mathcal{X}\,|^{-1}$
- Clairvoyant: $\mu_{\mathrm{y}\,|\,\mathrm{x},\theta_{\mathrm{c}}} = 1/\big(2 + \sin(2\pi\,\mathrm{x}\,)\big)$
- $\Sigma_{\mathrm{y}\,|\,\mathrm{x},\theta} = 0.2\,\mu_{\mathrm{y}\,|\,\mathrm{x},\theta}(1 - \mu_{\mathrm{y}\,|\,\mathrm{x},\theta})$
  $\Rightarrow\ \mathcal{R}^*_{\Theta}(\theta) \approx 0.039$

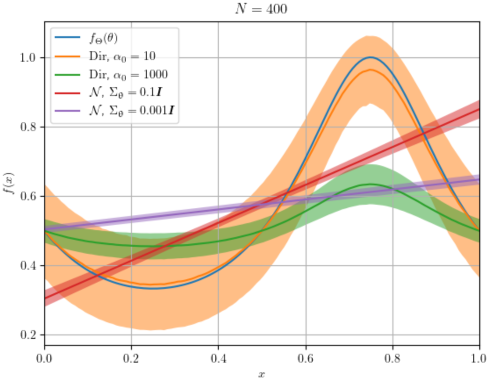
Learners:

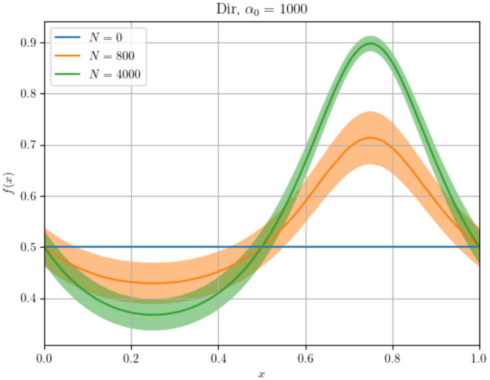
- Dirichlet:
  - $\alpha_{\mathrm{m}} = |\,\mathcal{X}\,|^{-1}$
  - Prior $\mu_{\mathrm{y}\,|\,\mathrm{x}} = 0.5$
- Normal[9]:
  - $\mathrm{y}\,|\,\mathrm{x}, \theta \sim \mathcal{N}\left([1, \mathrm{x}]\theta, 0.1\right)$
  - $\theta \sim \mathcal{N}\left([0.5, 0], \Sigma_{\theta}\right)$

- *Prior confidence* of Dirichlet and Normal learners varied using α_0 and Σ_θ
- Both learners effect the same biased untrained regressor to approximate the non-linear clairvoyant regressor

[9] Theodoridis, *Machine Learning: A Bayesian and Optimization Perspective*.

Prediction Statistics





- Lines show bias $\mathrm{E}_{\psi|\theta}\left[\mu_{\mathrm{y}\,|\,\mathrm{x},\psi}\right]$, fill regions shows variance $\mathrm{C}_{\psi|\theta}\left[\mu_{\mathrm{y}\,|\,\mathrm{x},\psi}\right]$
- Python simulation results average 50,000 learning iterations

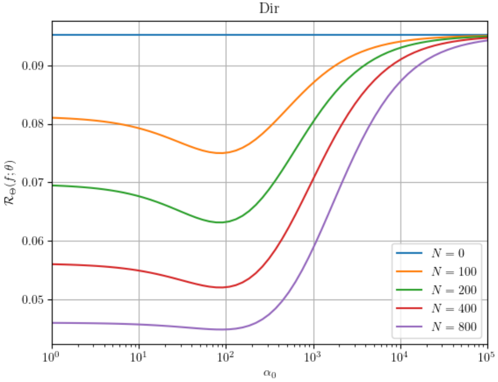
Optimal Localization

For a given conditional mean α_c, localization $\bar{\alpha}_0(x) \equiv \alpha_0\, \alpha_m(x)$ controls a Bias-Variance trade-off:

$\bar{\alpha}_0(x)$	$\lambda_{\text{Bias}}(x;\theta_m)$	$\lambda_{\text{Var}}(x;\theta_m)$
$\to \infty$	1	0
$\to 0$	$\left(1-\theta_m(x)\right)^N$	$E_{\psi_m \mid \theta_m}\left[\left(N\,\psi_m(x)\right)^{-1}\right]$

Optimal:

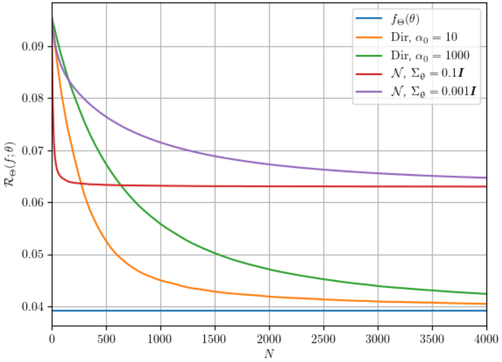
$$\bar{\alpha}_0(x) = \frac{\Sigma_{y \mid x,\theta_c}}{\left(\mu_{y \mid x} - \mu_{y \mid x,\theta_c}\right)^2}$$



Training Volume Trends

- As $N \to \infty$, both $\lambda_{\text{Bias}}(\mathrm{x}; \theta_{\mathrm{m}}) \to 0$ and $\lambda_{\text{Var}}(\mathrm{x}; \theta_{\mathrm{m}}) \to 0$
  $\Rightarrow \ \mathcal{R}_{\Theta,\mathrm{ex}}(f^*; \theta) \to 0$ for any model θ

- Note that $f^*(\mathrm{x}; \psi)$ converges to the clairvoyant regressor regardless of how biased the prior conditional mean α_{c} is, or how much confidence in α_{c} is indicated through the localization α_0



Conclusions

Summary

Full-support Bayesian learning with a Dirichlet prior enables:

- Asymptotically optimal performance for data-rich applications
- Maximal prior knowledge required for data-limited applications

Future Work

- Generalize these concepts for more general data models using the continuous Dirichlet process[10]
  - Practical necessity motivates the use of discretization to realize the demonstrated benefits
- Use the Dirichlet prior with different likelihood functions (e.g., mixture model) to effect limited-support priors that may be best suited for data-limited applications

[10] Samuel J. Gershman et al. "A tutorial on Bayesian nonparametric models". In: *Journal of Mathematical Psychology* 56 (2012).