Dirichlet Process Mixtures of Generalized Linear Models

Lauren Hannah
David Blei
Warren Powell

Princeton University
1Department of Operations Research and Financial Engineering
2Department of Computer Science

May 13, 2010
Nonparametric Regression
Nonparametric Regression

- Covariates $X$ and response $Y$
- $X$ and $Y$ may have different forms (continuous, count, categorical)
- Goal: prediction, ie compute $E[Y|X = x]$
- Parametric regression restricts shape (a straight line, polynomial, etc)
- Nonparametric regression tries to fit a function
Motivation

Nonparametric Regression Goals

- Flexible model
- Accommodate input/output types
- Be successfully applied to data with different characteristics
- Theoretical assurances, like asymptotic unbiasedness
- Computational tractability
Idea!

*Locally*, a complex model can be represented by a simpler model

- Dirichlet process mixture models:
  - Cluster observations probabilistically
  - Can accommodate many data types
- Cluster data so that a GLM fits well in each cluster
  - Clusters *and* local GLM parameters are latent variables
  - Predict mean response by averaging posterior draws
What am I going to talk about?

- Abbreviation: DP-GLM
- General regression method for all input types accommodated by DP and output types accommodated by GLM
- Continuous, categorical, count, circular, etc covariates/response
- Generalization of existing special case methods (eg Shahbaba and Neal (2009))
- We give conditions for asymptotic unbiasedness
Start with training data.
Cluster and fit regression probabilistically.
Observe testing data—we want to predict a mean function.
Fit testing covariates into clustered model; average to get mean function.
Properties of the Dirichlet Process

- A distribution over distributions—i.e. a draw from a DP is a random measure
- Random measures from DPs are almost surely discrete
  - When used as a distribution on hidden parameters, this produces a clustering effect
- Parameterized by base probability measure $G_0$ and scale $\alpha$
- If $\theta_1, \ldots, \theta_n \sim P$, $P \sim DP(\alpha G_0)$, then
  \[
  \theta_{n+1}|\theta_1:n \sim \frac{1}{\alpha + n} \sum_{i=1}^{n} \delta_{\theta_i} + \frac{\alpha}{\alpha + n} G_0
  \]
- Use as prior on distribution for hidden parameters $\theta_i$
Dirichlet Process Mixtures of Generalized Linear Models (DP-GLM) for covariates $X$ and response $Y$:

\[ P \sim DP(\alpha G_0) \]
\[ \theta_i | P \sim P \]
\[ X_i | \theta_i \sim f_X(x | \theta_i, x) \]
\[ Y_i | \theta_i, X_i \sim f_Y(y | X_i, \theta_i, y) \]

**Example: Gaussian Model: $X, Y \in \mathbb{R}$**

\[ P \sim DP(\alpha G_0) \]
\[ \theta_i = (\mu_{i,x}, \sigma_{i,x}, \beta_{i,0}, \beta_{i,1}, \sigma_{i,y}) | P \sim P \]
\[ X_i | \mu_{i,x}, \sigma_{i,x} \sim N(\mu_{i,x}, \sigma_{i,x}^2) \]
\[ Y_i | \beta_{i,0}, \beta_{i,1}, \sigma_{i,y}, X_i \sim N(\beta_{i,0} + \beta_{i,1}X_i, \sigma_{i,y}^2) \]
DP-GLM: Gaussian Model

Training Data

Testing Data

Lauren Hannah (Princeton University)
Given data $D = (X_i, Y_i)_{1:n}$, we want to compute $\mathbb{E}[Y|X = x, D]$

1) Choose the GLM
2) Choose DP base measure $G_0$
3) Estimate posterior of $\theta_{1:n}$ given $(X_i, Y_i)_{1:n}$
   - We use Gibbs sampling, Neal (2000) Algorithms 3, 6 or 8
   - Obtain $M$ i.i.d. samples of $\theta_{1:n}^{(m)}$ from the posterior
4) Compute predicted value $\mathbb{E}[Y|X = x, D]$

$$\mathbb{E}[Y|X = x] = \mathbb{E} [\mathbb{E}[Y|X = x, D, \theta_{1:n}]]$$
Computational Procedure

Computing the prediction \( \mathbb{E}[Y|X = x, D] \)

- Given \( \theta_{1:n} \), we can compute expectation:

\[
\mathbb{E}[Y|x, \theta_{1:n}] = \frac{1}{b} \sum_{i=1}^{n} \mathbb{E}[Y|x, \theta_i] f_x(x|\theta_i) + \frac{\alpha}{b} \int \mathbb{E}[Y|x, \theta] f_x(x|\theta) G_0(d\theta),
\]

\[
b = \alpha \int f_x(x|\theta) G_0(d\theta) + \sum_{i=1}^{n} f_x(x|\theta_i).
\]

- Get \( M \) observations of \( \theta_{1:n} \)
- But \( \theta_{1:n} \) is unknown, so we average over samples \( (\theta_{1:n}^{(m)})_{m=1}^{M} \)

\[
\mathbb{E}[Y|X = x, D] \approx \sum_{m=1}^{M} \mathbb{E}\left[Y|X = x, D, \theta^{(m)}\right]
\]
Asymptotic Unbiasedness

- Want our estimate of the mean function to converge to the true mean function as we get more observations.
- This is not a given with Dirichlet process priors (Diaconis and Freedman, 1986).
- Asymptotic unbiasedness depends on:
  - True distribution of $X, Y$, denoted $f_0(x, y)$.
  - Model (i.e. DP-GLM parametric functions).
  - Base measure $G_0$. 

Theoretical Properties of the DP-GLM

Theorem

The DP-GLM is asymptotically unbiased in a compact set of covariates $C$ if:

(i) (K-L Condition) for every $\delta > 0$, prior puts positive measure on

$$\left\{ f : \int f_0(x, y) \log \frac{f_0(x, y)}{f(x, y)} dx dy < \delta, \right. \nonumber$$

$$\left. \int f_0(x, y) \left( \log \frac{f_0(x, y)}{f(x, y)} \right)^2 dx dy < \delta \right\}, \nonumber$$

(ii) $\int |y|^2 f_0(y|x) dy < \infty$ for every $x \in C$, and

(iii) there exists an $\epsilon > 0$ such that for every $x \in C$,

$$\int \int |y|^{1+\epsilon} f_y(y|x, \theta) G_0(d\theta) < \infty. \nonumber$$
Theoretical Properties of the DP-GLM

Satisfying Main Theorem

- K-L condition is hard to show.
- When is it satisfied?
  - Gaussian Model: conjugate base measures, shown in slide.
  - Continuous and categorical covariates/response can be used as well with conjugate base measures.
- The rest is an open question.
Empirical Analysis
DP-GLM Comparison: Heteroscedastic Data

CMB Dataset

<table>
<thead>
<tr>
<th>DPGLM</th>
<th>Gaussian Process</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Wavelength</td>
<td></td>
</tr>
<tr>
<td>DP-GLM</td>
<td>Treed Linear Model</td>
</tr>
<tr>
<td>Treed GP</td>
<td></td>
</tr>
<tr>
<td>Treed Linear Model</td>
<td></td>
</tr>
</tbody>
</table>

Lauren Hannah (Princeton University)
Cosmic Microwave Background (CMB) Bennett et al. (2003)
- Power spectrum vs. multipole moments.
- One continuous covariate, continuous response.
- Heteroscedastic noise.

Concrete Compressive Strength (CCS) Yeh (1998)
- Concrete compressive strength against composition covariates (cement, water, fly ash, etc).
- Eight continuous covariates, one continuous response.
- Low noise, moderate dimensionality.

Solar Flare (Solar) Bradshaw (1989)
- Number of solar flares vs. sun features (solar spots, etc).
- Eleven categorical covariates, count response.
- Moderate dimensionality, atypical covariate/response types.
Competitors

- Least squares linear regression (for CMB, CCS)
- Tree regression (CART), treed linear models
- Gaussian process prior regression, treed Gaussian processes
- Dirichlet process regression \textit{without} GLM
- Poisson regression (for Solar)
Numerical Results: Cosmic Microwave Background

Covariates:
- 1 continuous

Response:
- continuous

Other:
- heteroscedastic

CMB Dataset

Number of Observations

Error

Algorithm
- DP–GLM
- OLS
- CART
- Gaussian Process
- Bayesian Cart
- Treed
- Gaussian Process

Algorithm

DP–GLM
May 13, 2010
Numerical Results: Concrete Compressive Strength

Covariates:
- 8 continuous
Response:
- continuous

CCS Dataset

Mean Absolute Error
Mean Square Error

Algorithm
- DP−GLM
- OLS
- CART
- Gaussian
- Process
- Scale DP
- Bayesian
- CART

Number of Observations

Lauren Hannah (Princeton University)
Numerical Results: Solar Flare

Covariates:
- 11 categorical

Response:
- count

Solar Dataset

Mean Absolute Error

Mean Square Error

Algorithm
- DP-GLM
- Poisson
- GLM
- Tree
- Regression
- Bayesian
- CART

Number of Observations

50 100 200 500 800
Summary

DP-GLM Issues/Future Work:
- Automate choice of $G_0$, hyperparameters
- Investigate balance between modeling covariates and response

DP-GLM Pros:
- Flexible nonparametric regression method; can be used in many settings
- Generally competitive with state of the art regression methods
- Generally stable outputs
- Can accommodate heteroscedasticity, overdispersion in a natural manner
Thank You!

Lauren Hannah
lhannah@princeton.edu