Bayesian Approaches to Distribution Regression

Ho Chung Leon Law*, Dougal J. Sutherland*, Dino Sejdinovic, Seth Flaxman

ho.law@spc.ox.ac.uk, dougals@gatsby.ucl.ac.uk, dino.sejdinovic@stats.ox.ac.uk, s.flaxman@imperial.ac.uk; arXiv:1705.04293

Distribution regression

Inputs are sample sets from distributions [e.g. 7]:

\[ y_i = f^*(P_i) + \varepsilon \]

Examples:

- Multivariate distribution [7]: \( \mathbb{P} \)
- Sample from distribution
- Entropy of a projection

| Galaxy velocities in a cluster [4, 5] | Cluster mass: \( 7 \times 10^{5.3} \) |

Bayesian linear regression

- Standard BLR model (similar to e.g. [3]):
- Handles regression weight \( \beta \) uncertainty
- Assumes \( \hat{\mu}_i \) known exactly
- Normal prior over regression weights \( \beta \sim \mathcal{N}(0, \rho^2) \)
- Observations \( y_i \mid X_i, \beta \sim \mathcal{N}(\beta^T \mu_i, \sigma^2) \)
- Gives normal \( y_i \mid X_i \) with hyperparameters \( \sigma, \rho \)

Shrinkage

- New model
- Handles \( \hat{\mu}_i \) uncertainty
- Point estimate for \( \beta \) to retain conjugacy
- Uses Bayesian nonparametric model for \( \hat{\mu}_i \) [1]:
  - Prior: \( \mu_i \sim \mathcal{GP}(m_0, \eta_0(\cdot, \cdot)) \)
  - Likelihood: “observed” at points \( u_i \), CLT:
    \( \hat{\mu}_i(u) \mid \mu_i(u) \sim \mathcal{N}(\mu_i(u), \Sigma_i) \)
- Closed-form GP posterior for \( \hat{\mu}_i \mid X_i \)
- Similar to James-Stein shrinkage
- Observations \( y_i \mid \mu_i, f \sim \mathcal{N}(f(\mu_i), \sigma^2) \)
- Say \( f(\cdot) = \sum_k \alpha_k k(\cdot, z_k) \) (representer theorem)
- Predictive: \( y_i \mid X_i, \alpha, \Sigma \sim \mathcal{N}(\alpha^T \Sigma^{-1} R_{zz} R_{zz}^{-1}(\Sigma + \Sigma_i)^{-1})(\mu_i - m_0) + \Sigma^{-1} m_0 \)
- MAP estimate for \( \alpha, \sigma \), maybe \( k \), kernel params...
- Optimise with backprop (TensorFlow)

Bayesian Distribution Regression

- Full uncertainty model (BDR)
- Shrinkage posterior for \( \hat{\mu}_i \)
- BLR-like posterior for weights \( \alpha \)
- Not conjugate
- MCMC for inference about \( \alpha \) (Stan)

Results on synthetic data

Toy problem to examine input uncertainty:
- Labels \( y_i \) uniform over \([0, 4.8]\)
- 5d data points: \( [x_j^T] \mid y_i \sim \mathcal{N}(y_j^T \mu_i, \sigma^2) \)
- \( \{x_{(s, 25, 100, 100 - s)}/\% \} \) have \( N_i = (5, 20, 100, 100) \)
- BDR \( \approx \) shrinkage < BLR in NLL, MSE

Age prediction from images

- IMDb database from [6]
- Very noisy labels in the dataset
- Group pictures of actors, predict mean age
- Features: last hidden layer from [6]'s CNN
- Lots of variation in \( N_i \)

<table>
<thead>
<tr>
<th>Method</th>
<th>NLL</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN</td>
<td>3.80 (0.03)</td>
<td>10.25 (0.22)</td>
</tr>
<tr>
<td>RBF network</td>
<td>9.51 (0.20)</td>
<td></td>
</tr>
<tr>
<td>BLR</td>
<td>3.68 (0.02)</td>
<td>9.55 (0.19)</td>
</tr>
<tr>
<td>shrshrinkage</td>
<td>3.54 (0.02)</td>
<td>9.28 (0.20)</td>
</tr>
</tbody>
</table>

References