Distributional Functional Estimation and Applications

Kevin R. Moon

May 25, 2016
Iris data set with added noise and KDE level sets

1. How similar are each of the classes to each other?
2. What is the intrinsic dimension of the data?
3. What is the best possible error rate any classifier can asymptotically achieve (i.e., the Bayes error)?
4. Are the variables dependent?
5. Are any of the points anomalies?
6. Which features are most relevant for classification?
Motivation

1. How similar are each of the classes to each other?
2. What is the intrinsic dimension of the data?
3. What is the best possible error rate any classifier can asymptotically achieve (i.e. the Bayes error)?
4. Are the variables dependent?
5. Are any of the points anomalies?
6. Which features are most relevant for classification?

**Answer:** Distributional functional estimation
Distributional Functionals

Distributional functional: integral functional of one or more densities, e.g.,

- **Divergence Functionals**

  \[ G(f_1, f_2) = \int g(f_1(x), f_2(x)) f_2(x) dx \]

  - Includes Kullback-Leibler divergence \( g(x, y) = -\ln \left( \frac{x}{y} \right) \)

- **Mutual Information Functionals**

  \[ G(X; Y) = \int g \left( \frac{f_X(x)f_Y(y)}{f_{XY}(x, y)} \right) f_{XY}(x, y) dx dy \]

- **Entropy Functionals**

  \[ G(f) = \int g(f(x)) f(x) dx \]
1. How similar are each of the classes to each other?
   - Divergence

2. What is the intrinsic dimension of the data?
   - Entropy

3. What is the best possible error rate any classifier can asymptotically achieve (i.e. the Bayes error)?
   - Divergence, Mutual Information

4. Are the variables dependent?
   - Mutual Information

5. Are any of the points anomalies?
   - Entropy

6. Which features are most relevant for classification?
   - Mutual Information
Smoothness is wrt Hölder condition
Goal is to estimate these functionals when only a finite population of iid samples is available from each distribution.

Kernel density plug-in methods are **highly biased** for large dimension $d$.

**Heat map of predicted bias of divergence functional plug-in estimator**

$$\tilde{G}_h = \frac{1}{N} \sum_{i=1}^{N} g \left( \tilde{f}_1, h(X_i), \tilde{f}_2, h(X_i) \right)$$

**Use weighted ensemble estimation to improve bias**

$$\tilde{G}_{w} = \sum_{\ell \in \ell} w(\ell) \tilde{G}_{h(\ell)}$$
\[ \ell = \{\ell_1, \ell_2, \ldots, \ell_L\} \] a set of index values, \( N \) the number of samples

- An ensemble of estimators \( \left\{ \hat{E}_\ell \right\}_{\ell \in \ell} \) of parameter \( E \) and weights \( w \) with \( \sum_{\ell \in \ell} w(\ell) = 1 \)
Ensemble Estimation Procedure

1. Derive the bias: $\mathbb{B}\left[\hat{E}_\ell\right] = \sum_{i \in J} c_i \psi_i(\ell) \phi_i, d(N) + O\left(\frac{1}{\sqrt{N}}\right)$
   - $c_i$ are constants, $J$ a finite index set, $\psi_i(\ell)$ are basis functions independent of $N$

2. Derive the variance: $\nabla\left[\hat{E}_\ell\right] = c_\nu \left(\frac{1}{N}\right) + o\left(\frac{1}{N}\right)$

3. Calculate (offline) optimal weight $w_0$ to zero out lower order bias terms (Moon et al, 2016a):

   $$\begin{align*}
   &\min_w \quad ||w||_2 \\
   &\text{subject to} \quad \sum_{\ell \in \ell^-} w(\ell) = 1,
   \\
   &\quad \gamma_w(i) = \sum_{\ell \in \ell^-} w(\ell) \psi_i(\ell) = 0, \ i \in J.
   \end{align*}$$

4. MSE of $\hat{E}_{w_0} = \sum_{\ell \in \ell^-} w_0(\ell)\hat{E}_\ell$ is $O\left(\frac{1}{N}\right)$
Divergence Functional Example (Moon et al, 2016)

- **Divergence functional:**
  \[ G(f_1, f_2) = \int g(f_1(x), f_2(x)) f_2(x) \, dx \]

- **Plug-in estimator**
  \[ \tilde{G}_h(\ell) = \frac{1}{N} \sum_{i=1}^{N} g(\tilde{f}_{1,h}(X_i), \tilde{f}_{2,h}(X_i)) \]

  - \( f_j \) are densities, \( X_i \) drawn from \( f_2 \), \( \tilde{f}_{i,h} \) are KDEs with bandwidth \( h(\ell) = \ell N^{-1/(2d)} \)

- **Bias**
  \[ \mathbb{B}[\tilde{G}_h(\ell)] = \sum_{i=1}^{d} c_i \ell^i N^{-i/(2d)} + O\left(\frac{1}{\sqrt{N}}\right) \]
Advantages of our Estimators

They

- Apply to general distributional functionals
- Are simpler to implement than competing estimators
- Achieve the parametric MSE rate ($O(1/T)$) if densities are at least $(d + 1)/2$ times differentiable
  - Competitive with other estimators
- Are computationally tractable
- Have a central limit theorem for performing hypothesis testing
  (Moon and Hero, 2014a; Moon et al, 2016)
Applications

- Intrinsic dimension estimation via entropy estimation for dimensionality reduction (Carter et al, 2010)
  - Applied to sunspot images (Moon et al, 2014, 2016a) and neural data (Gliske et al, 2016)

From (L to R) NASA Goddard Space Flight Center, William Stacey Lab
Applications

- Extending machine learning applications (e.g. classification and clustering) to probability distributions as features (Poczos and Schneider, 2011; Oliva et al, 2013)
  - Use divergence functionals as a measure of dissimilarity
  - Applied to cluster sunspot images (Moon et al, 2016b)

![Sunspot Image](image)

NASA Goddard Space Flight Center

- Testing whether two collections of samples come from the same distribution (Moon et al, 2016a)
Other Applications

- Estimating the best probability of error for a classification problem, the Bayes error (Moon and Hero, 2014a; Moon et al, 2015; Berisha et al, 2015)
- Text/multimedia clustering (Dhillon et al, 2003)
- Feature selection for machine learning (Peng et al, 2005)
- fMRI data processing (Chai et al, 2009)
- Clustering (Lewi et al, 2006)
- Neuron classification (Schneidman et al, 2002)
- Anomaly detection (Sricharan et al, 2010)
- Texture classification and image registration (Hero et al, 2002)


Other References

Other References