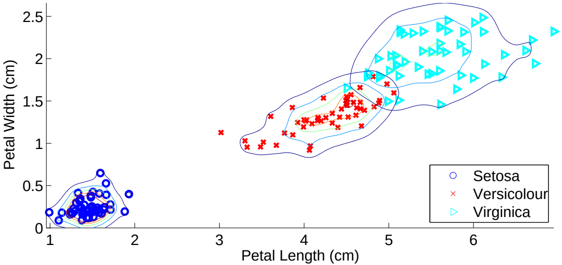
Distributional Functional Estimation and Applications

Kevin R. Moon

May 25, 2016

Motivation



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

$$G\left(f_1, f_2\right) = \int g\left(f_1(x), f_2(x)\right) f_2(x) dx$$

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

- Mutual Information Functionals

$$G\left(\mathbf{X};\mathbf{Y}\right) = \int g\left(\frac{f_X(x) f_Y(y)}{f_{XY}(x,y)}\right) f_{XY}(x,y) dxdy$$

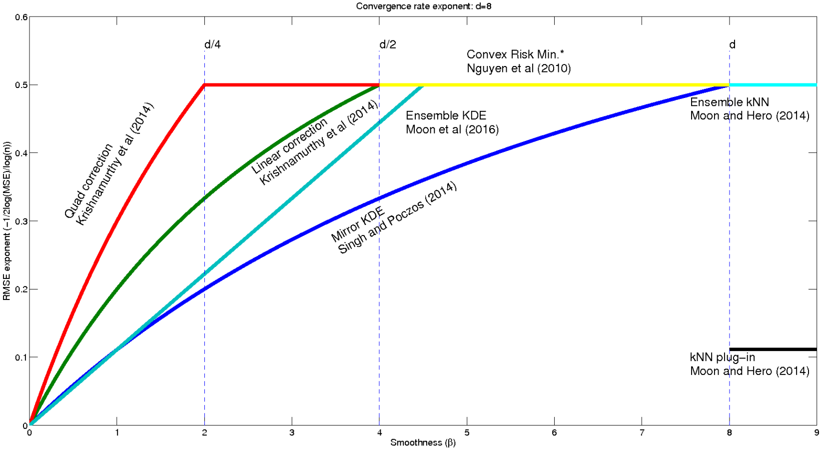
- Entropy Functionals

$$G\left(f\right) = \int g\left(f(x)\right) f(x) dx$$

Distributional Functionals

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

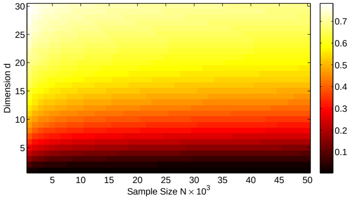
Related Work



Smoothness is wrt Hölder condition

Nonparametric Estimation of Distributional Functionals

- 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
$$\tilde{\mathbf{G}}_h = \frac{1}{N}\sum_{i=1}^{N} g\left(\tilde{\mathbf{f}}_{1,h}(\mathbf{X}_i), \tilde{\mathbf{f}}_{2,h}(\mathbf{X}_i)\right)$$

- Use **weighted ensemble** estimation to improve bias

$$\tilde{\mathbf{G}}_w = \sum_{\ell \in \bar{\ell}} w(\ell)\tilde{\mathbf{G}}_{h(\ell)}$$

Ensemble Estimation Setup

- $\bar{\ell} = \{\ell_1, \ell_2, \ldots, \ell_L\}$ a set of index values, N the number of samples
- An ensemble of estimators $\left\{\hat{\mathbf{E}}_\ell\right\}_{\ell \in \bar{\ell}}$ of parameter E and weights w with $\sum_{\ell \in \bar{\ell}} w(\ell) = 1$

Ensemble Estimation Procedure

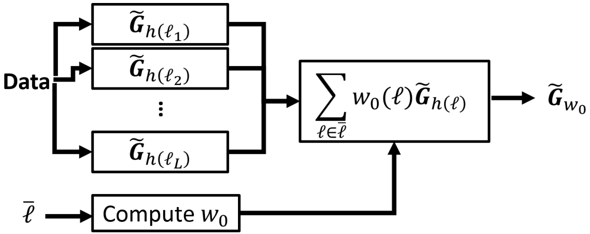
1. Derive the **bias**: $\mathbb{B}\left[\hat{\mathbf{E}}_\ell\right] = \sum_{i\in J} c_i\psi_i(\ell)\phi_{i,d}(N) + O\left(\frac{1}{\sqrt{N}}\right)$
   - c_i are constants, J a finite index set, $\psi_i(\ell)$ are basis functions independent of N
2. Derive the **variance**: $\mathbb{V}\left[\hat{\mathbf{E}}_\ell\right] = c_v\left(\frac{1}{N}\right) + o\left(\frac{1}{N}\right)$
3. Calculate (offline) **optimal weight** w_0 to zero out lower order bias terms (Moon et al, 2016a):

$$\begin{array}{rl} \min_w & ||w||_2 \\ \textit{subject to} & \sum_{\ell\in\bar{\ell}} w(\ell) = 1, \\ & \gamma_w(i) = \sum_{\ell\in\bar{\ell}} w(\ell)\psi_i(\ell) = 0,\ i \in J. \end{array}$$

4. **MSE** of $\hat{\mathbf{E}}_{w_0} = \sum_{\ell\in\bar{\ell}} w_0(\ell)\hat{\mathbf{E}}_\ell$ 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 $\tilde{\mathbf{G}}_{h(\ell)} = \frac{1}{N} \sum_{i=1}^{N} g\left(\tilde{\mathbf{f}}_{1,h(\ell)}(\mathbf{X}_i), \tilde{\mathbf{f}}_{2,h(\ell)}(\mathbf{X}_i)\right)$
  - f_j are densities, $\mathbf{X}_i$ drawn from f_2, $\tilde{\mathbf{f}}_{i,h}$ are KDEs with bandwidth $h(\ell) = \ell N^{-1/(2d)}$
- $\mathbb{B}\left[\tilde{\mathbf{G}}_{h(\ell)}\right] = \sum_{i=1}^{d} c_i \ell^i N^{-i/(2d)} + O\left(\frac{1}{\sqrt{N}}\right)$



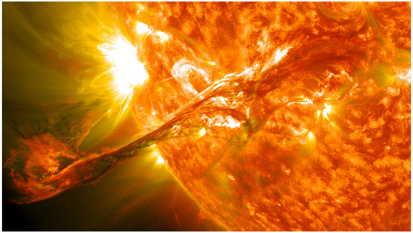
Advantages of our Estimators

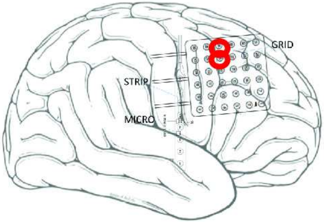
They

- 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)

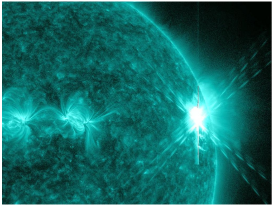




From (L to R) NASA Goddard Space Flight Center, William Stacey Lab

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)



NASA Goddard Space Flight Center

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

- 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)

References

- K. Moon and A. Hero, "Multivariate f-divergence estimation with confidence," *Advances in Neural Information Processing Systems (NIPS)*, 2014a.
- K. Moon and A. Hero, "Ensemble estimation of multivariate f-divergence," *IEEE Int. Symp. on Inf. Theory (ISIT)*, 2014b.
- K. Moon, J. Li, V. Delouille, F. Watson, and A. Hero, "Image patch analysis and clustering of sunspots: a dimensionality reduction approach," *IEEE Int. Conf. Image Processing (ICIP)*, 2014.
- K. Moon, V. Delouille, and A. Hero, "Meta learning of bounds on the Bayes classifier error," *IEEE Signal Processing and SP Education Workshop (SPW)*, 2015.
- K. Moon, K. Sricharan, K. Greenewald, and A. Hero, "Improving convergence of divergence functional ensemble estimators," To appear in *IEEE International Symposium on Information Theory*, 2016.

References

- K. Moon, J. Li, V. Delouille, R. De Visscher, F. Watson, and A. Hero, "Image patch analysis of sunspots and active regions. I. Intrinsic dimension and correlation analysis," *Journal of Space Weather and Space Climate*, 2016a.
- K. Moon, V. Delouille, J. Li, R. De Visscher, F. Watson, and A. Hero, "Image patch analysis of sunspots and active regions. II. Clustering via matrix factorization," *Journal of Space Weather and Space Climate,* 2016b.
- S. Gliske, K. Moon, W. Stacey, and A. Hero, "The intrinsic value of HFO features as a biomarker of epileptic activity," *IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP),* 2016.

Other References

- B Poczos and JG Scheider, "On the estimation of alpha-divergences," in *International Conference on Machine Learning*, 2013.
- IS Dhillon, S Mallela, and R Kumar, "A divisive information theoretic feature clustering algorithm for text classification," *The Journal of Machine Learning Research*, 2003.
- H Peng, F Long, and C Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005.
- B Chai, D Walther, D Beck, and L Fei-Fei, "Exploring functional connectivities of the human brain using multivariate information analysis," *In Adv. Neural Inf. Process. Syst.*, 2009.
- J Lewi, R Butera, and L Paninski, "Real-time adaptive information-theoretic optimization of neurophysiology experiments," in *Adv. Neural Inf. Process. Syst.*

Other References

- E Schneidman, W Bialek, and MJ Berry, "An information theoretic approach to the functional classification of neurons," in *Adv. Neural Inf. Process Syst.*, 2002.
- KM Carter, R Raich, and AO Hero, "On local intrinsic dimension estimation and its applications," *IEEE Transactions on Signal Processing*, 2010.
- AO Hero, B Ma, J Michel, and J Gorman, "Applications of entropic spanning graphs," *IEEE Signal Processing Magazine*, 2002.
- K Sricharan and AO Hero, "Efficient anomaly detection using bipartite k-NN graphs," in *Adv. Neural Inf. Proc. Syst.*, 2011.
- V. Berisha, A. Wisler, AO Hero, and A Spanias, "Empirically estimable classification bounds based on a nonparametric divergence measure," *IEEE Trans. Signal Processing*, 2016.