Set-valued Classification with Confidence: Least Ambiguity with Bounded Error Levels (LABEL)

Jing Lei

Department of Statistics, Carnegie Mellon University

ICSA International Conference, Shanghai

Dec 22, 2016

Joint work with Mauricio Sadinle (Duke) and Larry Wasserman (CMU)

Research partially supported by NSF grants

www.stat.cmu.edu/~jinglei/201612_icsa.pdf

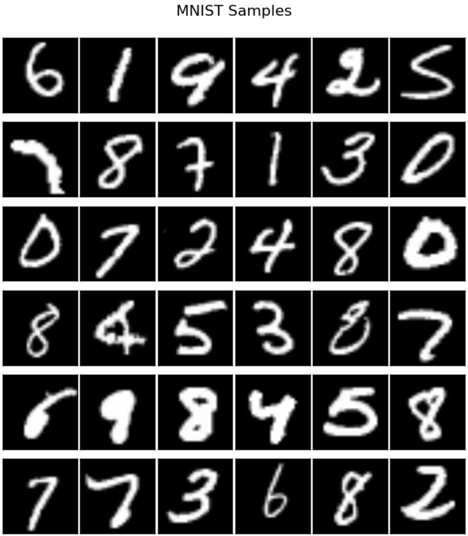
Outline

1. Set-valued multi-class classification
2. Ambiguity and coverage
3. Least ambiguity with bounded error levels
4. Examples

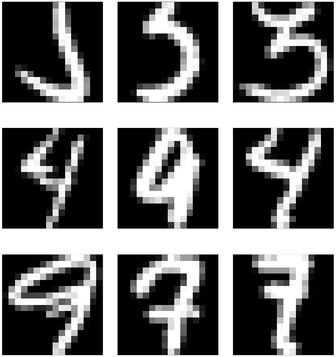
Multi-class classification

- Data: $(X_i, Y_i)_{i=1}^n \overset{iid}{\sim} P$ on $\mathscr{X} \times \mathscr{Y}$.
- Typically $\mathscr{X} \subset \mathbb{R}^d$, $\mathscr{Y} = \{1, ..., K\}$.
- Definitive classifier: for a given $x \in \mathscr{X}$, predict $\hat{y} = \hat{f}(x)$ where $\hat{f}$ is obtained from the data.
- Bayes classifier: f^* minimizes $P(f(X) \neq Y)$ over all f.

Example: MNIST Data ($n \approx 7k$, $d = 256$)



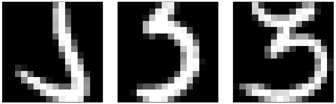
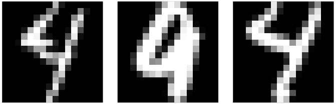
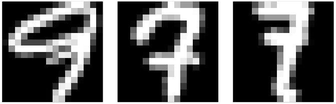
Some mis-classified samples



Method: logistic lasso with cross validation.

Observation: These samples are hard to classify, but further information can be given in addition to a single label.

Some mis-classified samples

- “3” and “5”: 
- “4” and “9”: 
- “7” and “9”: 

Method: logistic lasso with cross validation.

Observation: These samples are hard to classify, but further information can be given in addition to a single label.

Set-valued classifiers

- Idea: instead of outputting a single $\hat{y} \in \mathscr{Y}$, output a subset $H(x) \subseteq \mathscr{Y}$.
- Also known as non-deterministic classifiers (del Coz *et al*, 2009).
- A related approach is classification with rejection (Chow, 1970; Herbiri and Wegkamp 2006), where the reject option can be viewed as outputting $H(x) = \mathscr{Y}$.

Evaluating set-valued classifiers

- Existing works typically optimize some modified objective.
  1. Assign a loss between 0 and 1 to $H(x) = \mathscr{Y}$ (classification with rejection).
  2. Loss function is a combination of precision and recall (del Coz *et al*, 2009)

$$L_\beta(H(x), y) = \frac{1+\beta^2}{\beta^2 + |H(x)|} \mathbf{1}(y \in H(x))$$

Ambiguity and Coverage

- For a set-valued classifier $H : \mathscr{X} \mapsto 2^{\mathscr{Y}}$, we define two competing criteria
    - Ambiguity: $A(H) = E(|H(x)|)$
    - Coverage: $C(H) = P(Y \in H(X))$,
      or class-specific coverage: $C_y(H) = P(y \in H(X)|Y = y)$.
- A good classifier needs to cover the true class with high probability (high coverage), but outputs few classes (low ambiguity).
- Class-specific coverage is useful when classes are unbalanced.

The optimal classifier

Given $\alpha \in (0,1)$, define the *l*east *a*mbiguous classifier with *b*ounded *e*rror *l*evel (LABEL) as

$$H^* = \arg\min_H A(H) \quad \text{s.t.} \quad C(H) \geq 1 - \alpha\,.$$

Or the class-specific version: given $\alpha_y \in (0,1)$ $(y \in \mathscr{Y})$

$$H^* = \arg\min_H A(H) \quad \text{s.t} \quad C_y(H) \geq 1 - \alpha_y\,,\ \forall y \in \mathscr{Y}\,.$$

The minimization is taken over all measurable mappings from $\mathscr{X}$ to $2^{\mathscr{Y}}$.

Characterization of the optimal classifier

Let $p_x(y)$ be the conditional probability of $Y = y$ given $X = x$.

Theorem (Sadinle, L, Wasserman 2016)

For any $\alpha \in (0,1)$, the minimum of the overall LABEL problem is achieved by

$$H(x) = \{y \in \mathscr{Y} : p(y|x) \geq t\}$$

with t chosen such that $P(Y \in H(X)) = 1 - \alpha$.
For $(\alpha_y : y \in \mathscr{Y}) \in (0,1)^{\mathscr{Y}}$, the minimum of the class-specific LABEL optimization problem is achieved by

$$H(x) = \{y \in \mathscr{Y} : p(y|x) \geq t_y\}$$

with t_y chosen such that $P(y \in H(X)|Y = y) = 1 - \alpha_y$.

Connection to the Neyman-Pearson Lemma

- The characterization of optimal classifier can be viewed as a generalization to the Neyman-Pearson Lemma:

$$H_y = \{x \in \mathscr{X} : p_x(y) \geq t_y\}$$

where $H_y = \{x : y \in H(x)\}$ is the y-section of the classifier.

- A hypothesis testing perspective: Given x, we test $|\mathscr{Y}|$ hypotheses (let $p_y(\cdot)$ be the density of X given $Y = y$)

$$H_{0,y} : X \sim p_y(\cdot) \ \text{ vs } \ H_{1,y} : X \sim p_{y'}(\cdot) \ \text{ for some } \ y' \neq y\,.$$

- The optimal classifier consists of all the level $1 - \alpha_y$ non-rejection regions.

A useful lemma

Lemma (minimize incorrect labeling)

The optimal classifier also minimizes $P(y \in H(X) | Y \neq y)$ for all $y \in \mathscr{Y}$.

A plug-in estimate of optimal classifier

- Let $\hat{p}_x(y)$ be estimated conditional density of x given y.
- Let $\hat{t}_y = \hat{F}_y^{-1}(1-\alpha)$, where $\hat{F}_y$ is the empirical CDF of $\{\hat{p}_{x_i}(y) : Y_i = y\}$.
- The plug-in estimate is

$$\hat{H}_y = \{x : \hat{p}_x(y) \geq \hat{t}_y\}\,.$$

Estimation Accuracy

Let $G_y(\cdot) = P(p_X(y) \leq \cdot | Y = y)$ be the CDF of $p_X(y)$ given $Y = y$.

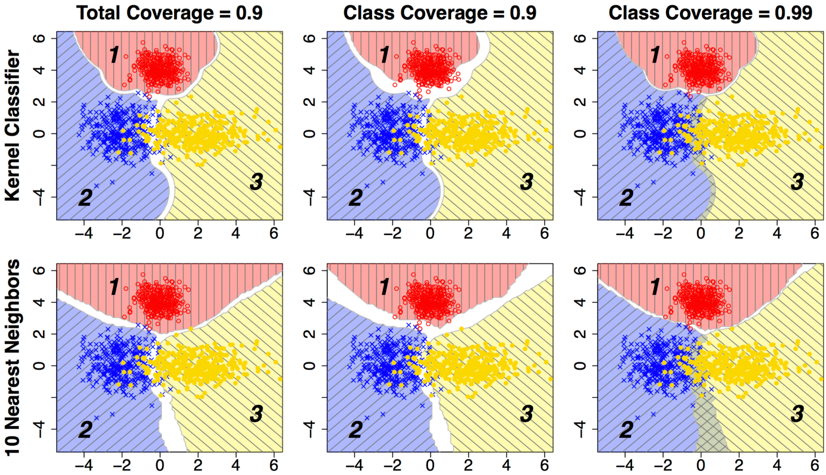
Theorem

Assume $P(\sup_x |\hat{p}_x(y) - p_x(y)| \leq \varepsilon) \leq \delta$ for all y, and $c_1|s|^\gamma \leq |G_y(t_y + s) - G_y(t_y)| \leq c_2|s|^\gamma$ for all $s \in [-s_0, s_0]$, then with probability at least $1 - |\mathscr{Y}|\delta - cn^{-1}$ we have

$$P(X \in \hat{H}_y \triangle H_y | Y = y) \leq c\left(\varepsilon^\gamma + \sqrt{\log n/n}\right), \ \ \forall\, y.$$

Remark: one can weaken the sup norm consistency condition on $\hat{p}(y|x)$ to consistency near the cut-off value.

Example: three-component Gaussian mixture



What if $H(x) = \emptyset$*?*

- If α_y's are too large, $\bigcup_y H_y$ may not equal to $\mathscr{X}$.
- If $x \in \left(\bigcup_y H_y\right)^c$, then $H(x) = \emptyset$.
- We call $N = \left(\bigcup_y H_y\right)^c$ the "null set".
- Sometimes N corresponds to points that look like outliers, but sometimes it corresponds to points that are highly ambiguous.

What if $H(x) = \emptyset$?

- If α_y's are too large, $\bigcup_y H_y$ may not equal to $\mathscr{X}$.
- If $x \in \left(\bigcup_y H_y\right)^c$, then $H(x) = \emptyset$.
- We call $N = \left(\bigcup_y H_y\right)^c$ the "null set".
- Sometimes N corresponds to points that look like outliers, but sometimes it corresponds to points that are highly ambiguous.

Lemma

If $\sum_y t_y \leq 1$, then $N = \emptyset$.
(Recall that $\hat{H}_y = \{x : \hat{p}(y|x) \geq t_y\}$.)

The Accretive Completion Algorithm

Idea: gradually reduce t with minimal incremental ambiguity

Require: $t^{(0)} = (t_1^{(0)}, \ldots, t_K^{(0)})$ from the initial estimate, step size η

  $s \leftarrow 0$

  while $\sum_y t_y^{(s)} > 1$ do

    for $y = 1, \ldots, K$ such that $t_y^{(s)} - \eta\ t_y^{(0)} > 0$ do

      $A_y \leftarrow$ empirical ambiguity using $t_1^{(s)}, \ldots, t_y^{(s)} - \eta\ t_y^{(0)}, \ldots, t_K^{(s)}$

    end for

    $y^* \leftarrow \arg\min_{y:\ t_y^{(s)} - \eta\ t_y^{(0)} > 0} A_y$

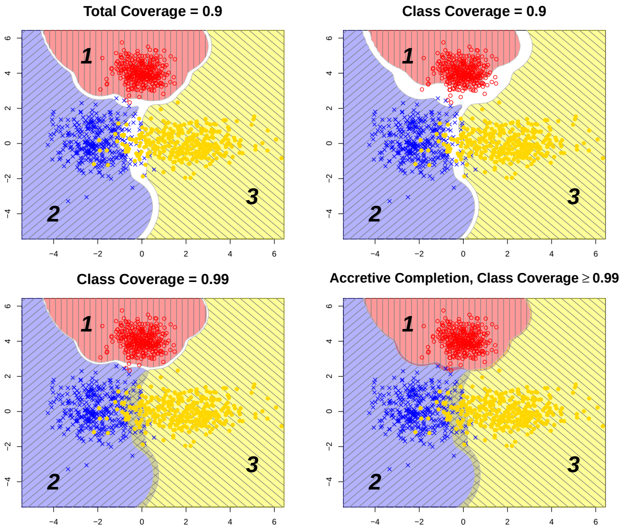
    $t^{(s+1)} = (t_1^{(s)}, \ldots, t_{y^*}^{(s)} - \eta\ t_{y^*}^{(0)}, \ldots, t_K^{(s)})$

    $s \leftarrow s + 1$

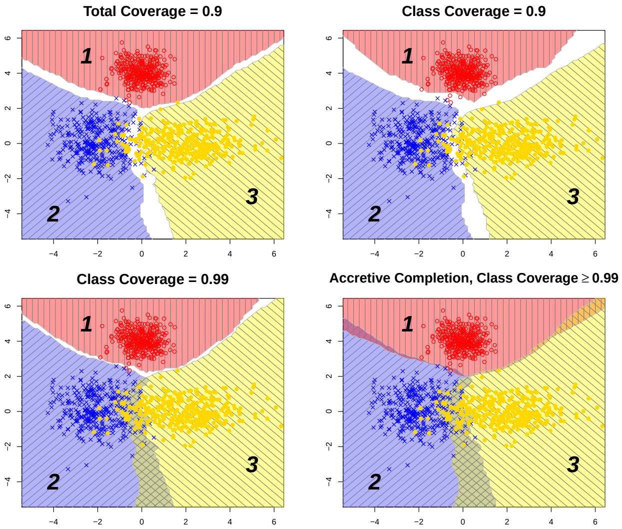
  end while

  return $t^{(s)}$

Example using kernel classification (cont'd)

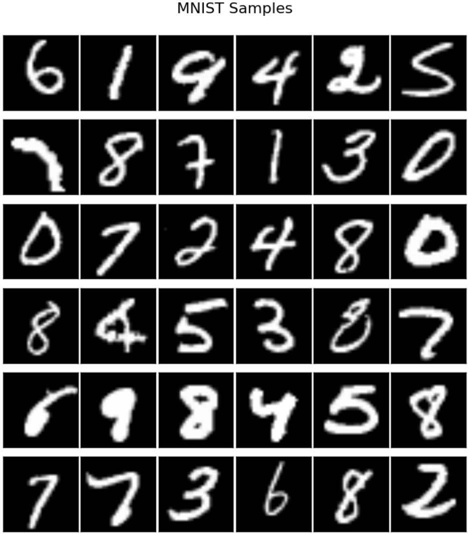


Example using 10-NN classification (cont'd)



MNIST Data

Goal: classify hand-written digits



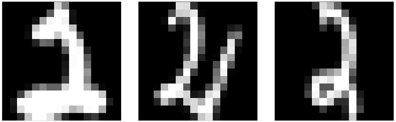
MNIST Data, cont'd

- We use kNN classifier with $k = 10$ chosen by three-fold CV on the training sample.
- Target class-specific coverage levels at $1 - 0.05$.
- 2/3 of the training sample is used to fit $\hat{p}(y|x)$, remainder for estimating the quantiles.

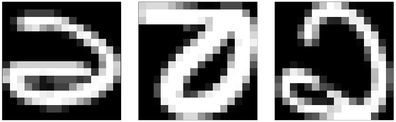
	$\lvert\hat{\mathbf{H}}(X)\rvert$				
	1	2	3	≥ 4	$\widehat{\mathbb{E}}\{\hat{\mathbf{H}}(X)\}$
Test sample frequency	1918	87	2	0	1.045

Ambiguous images reported by the algorithm

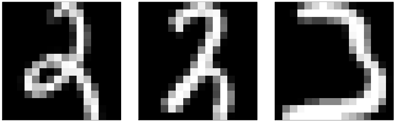
“1” or “2”



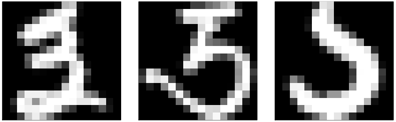
“2” or “0”



“2” or “3”



“3” or “5”



Summary

- New criteria for evaluating set-valued classifiers: coverage and ambiguity
- Optimize ambiguity subject to coverage constraints, and a generalized Neyman-Pearson Lemma.
- Accretive completion to remove null set.
- Flexible and transparent choice of parameters in the algorithm.
- Future work: distribution free, finite sample coverage; application to stomach cancer data.

References

- Lei, J., Classification with Confidence, *Biometrika*, **101**(4), 755-769.
- Sadinle, M., Lei, J., and Wasserman, L., Least Ambiguous Set-Valued Classifiers with Bounded Error Levels, *arXiv:1609.00451*.

- Slides: `www.stat.cmu.edu/~jinglei/201612_icsa.pdf`

Thanks!

Questions?