Advanced Machine Learning

Imbalanced Learning: Cost-Sensitive Learning Part 3

Confusion matrix

	True class	
	$y = 1$	$y = -1$
Pred. $\hat{y}$ = 1	TP	FP
class $\hat{y}$ = -1	FN	TN

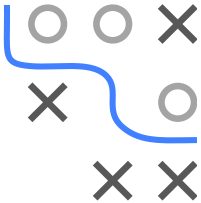
Cost matrix

	True class	
	$y = 1$	$y = -1$
Pred. $\hat{y}$ = 1	$C(1, 1)$	$C(1, -1)$
class $\hat{y}$ = -1	$C(-1, 1)$	$C(-1, -1)$

Learning goals

- Instance specific costs
- Cost-Sensitive OVO

BINARY INSTANCE-SPECIFIC COST LEARNING



- Assumes instance-specific costs for every observation: $\mathcal{D}^{(n)} = \{(\mathbf{x}^{(i)}, \mathbf{c}^{(i)})\}_{i=1}^{n}$, where $(\mathbf{x}^{(i)}, \mathbf{c}^{(i)}) \in \mathbb{R}^p \times \mathbb{R}^2$.
- Define "true class" as cost minimal class
- Define observation weights: $|\mathbf{c}^{(i)}[1] - \mathbf{c}^{(i)}[0]|$

	$\mathbf{c}^{(i)}[0]$	$\mathbf{c}^{(i)}[1]$	$y^{(i)}$	$w^{(i)}$
$\mathbf{x}^{(1)}$	1	1	0	0
$\mathbf{x}^{(2)}$	1	2	0	1
$\mathbf{x}^{(3)}$	7	3	1	4

- Now solve weighted ERM:

$$\mathcal{R}_{emp}(\boldsymbol{\theta}) = \sum_{i=1}^{n} w^{(i)} L\left(y^{(i)}, f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right)$$

- NB: Instances with equal costs are effectively ignored.

MULTICLASS COSTS

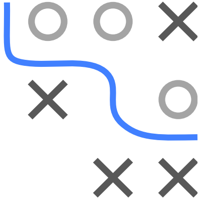
- Consider $g > 2$. Vanilla CSL is special case of instance specific, use $\mathbf{c}^{(i)}$ same for all $\mathbf{x}^{(i)}$ of the same class

		True class		
		$y = 1$	$y = 2$	$y = 3$
Pred. class	$\hat{y} = 1$	0	1	3
	$\hat{y} = 2$	1	0	1
	$\hat{y} = 3$	7	1	0

- For two $\mathbf{x}^{(i)}$ with $y = 2$ and $y = 3$:

	$\mathbf{c}^{(i)}[1]$	$\mathbf{c}^{(i)}[2]$	$\mathbf{c}^{(i)}[3]$	$y^{(i)}$
$\mathbf{x}^{(1)}$	1	0	1	2
$\mathbf{x}^{(2)}$	3	1	0	3
$\mathbf{x}^{(3)}$	1	0	1	2

- Set $\mathbf{c}^{(i)}[y^{(i)}] = 0$, i.e. zero-cost for correct prediction.

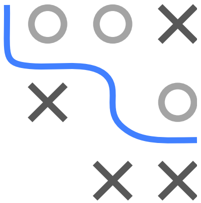


CSOVO ▸ LIN ET AL. 2014

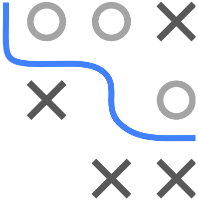
- Let $\mathcal{D}^{(n)} = \{(\mathbf{x}^{(i)}, \mathbf{c}^{(i)})\}_{i=1}^{n}$, $(\mathbf{x}^{(i)}, \mathbf{c}^{(i)}) \in \mathbb{R}^p \times \mathbb{R}^g$.
- Example:

	$\mathbf{c}^{(i)}[1]$	$\mathbf{c}^{(i)}[2]$	$\mathbf{c}^{(i)}[3]$
$\mathbf{x}^{(1)}$	0	2	3
$\mathbf{x}^{(2)}$	1	0	1
$\mathbf{x}^{(3)}$	2	0	3

- Idea: Reduction principle to binary case (weighted fit) by one-versus-one (OVO).
- For class j vs. k:
  - How to deal with the label $y^{(i)}$? $y^{(i)}$ can be neither j nor k.
  - How to deal with the costs $\mathbf{c}^{(i)}[j]$ and $\mathbf{c}^{(i)}[k]$?



CSOVO

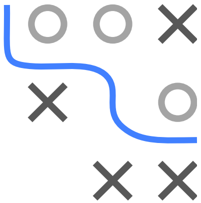


- When training a binary classifier $f^{(j,k)}$ for class j vs. k,
  - Choose cost min class from pair $\arg\min_{l \in \{j,k\}} \mathbf{c}^{(i)}[l]$ as ground truth
  - Sample weight is simply diff between the 2 costs $|\mathbf{c}^{(i)}[j] - \mathbf{c}^{(i)}[k]|$
- Example continued:

	$\mathbf{c}^{(i)}[1]$	$\mathbf{c}^{(i)}[2]$	$\mathbf{c}^{(i)}[3]$	$\mathbf{c}^{(i)}[1 \text{ vs } 2]$	$\tilde{y}^{(i)}[1 \text{ vs } 2]$	$w^{(i)}[1 \text{ vs } 2]$
$\mathbf{x}^{(1)}$	0	2	3	0/2	1	2
$\mathbf{x}^{(2)}$	1	0	1	1/0	2	1
$\mathbf{x}^{(3)}$	2	0	3	2/0	2	2

	$\mathbf{c}^{(i)}[1]$	$\mathbf{c}^{(i)}[2]$	$\mathbf{c}^{(i)}[3]$	$\mathbf{c}^{(i)}[2 \text{ vs } 3]$	$\tilde{y}^{(i)}[2 \text{ vs } 3]$	$w^{(i)}[2 \text{ vs } 3]$
$\mathbf{x}^{(1)}$	0	2	3	2/3	2	1
$\mathbf{x}^{(2)}$	1	0	1	0/1	2	1
$\mathbf{x}^{(3)}$	2	0	3	0/3	2	3

CSOVO



- Example continued

	$\mathbf{c}^{(i)}[1]$	$\mathbf{c}^{(i)}[2]$	$\mathbf{c}^{(i)}[3]$	$\mathbf{c}^{(i)}[1 \text{ vs } 3]$	$\tilde{y}^{(i)}[1 \text{ vs } 3]$	$w^{(i)}[1 \text{ vs } 3]$
$\mathbf{x}^{(1)}$	0	2	3	0/3	1	3
$\mathbf{x}^{(2)}$	1	0	1	-/-	-	0
$\mathbf{x}^{(3)}$	2	0	3	2/3	1	1

- Wrap everything up:
  1. For class j vs. k, transform all $(\mathbf{x}^{(i)}, \mathbf{c}^{(i)})$ to $(\mathbf{x}^{(i)}, \arg\min_{l \in \{j,k\}} \mathbf{c}^{(i)}[l])$ with sample-wise weight $|\mathbf{c}^{(i)}[j] - \mathbf{c}^{(i)}[k]|$.
  2. Train a weighted binary classifier $f^{(j,k)}$ using the above
  3. Repeat step 1 and 2 for different (j, k).
  4. Predict using the votes from all $f^{(j,k)}$.
- Theoretical guarantee:
  test costs of final classifier $\leq 2 \sum_{j<k}$ test cost of $f^{(j,k)}$.