Multiclass

Jordan Boyd-Graber
University of Colorado Boulder
LECTURE 14

Slides adapted from Rob Schapire and Fei Xia

Motivation

- Binary and Multi-class: problems and classifiers
- Solving Multi-class problems with binary classifiers
    - One-vs-all
    - All pairs
    - Error correcting codes

Classification Problems

- Natural binary
    - Spam classification (spam vs. ham)
    - Segmentation (same or different)
    - Coreference

Classification Problems

- Natural binary
  - Spam classification (spam vs. ham)
  - Segmentation (same or different)
  - Coreference
- However, many are multiclass
  - Topic classification
  - Part of speech tagging
  - Scene classification

Classifiers

- Some are directly multi-class (naïve Bayes, logistic regression, KNN)
- Other classifiers are basically binary

Classifiers

- Some are directly multi-class (naïve Bayes, logistic regression, KNN)
- Other classifiers are basically binary
  - SVM
  - Perceptron
  - Boosting

Reduction

Multiclass Data

```
⟨name=Cindy , age=5 , sex=F⟩,
⟨name=Marcia, age=15, sex=F⟩,
⟨name=Bobby , age=6 , sex=M⟩,
⟨name=Jan   , age=12, sex=F⟩,
⟨name=Peter , age=13, sex=M⟩,
```

Reduction

Multiclass Data

$\langle$name=Cindy , age=5 , sex=F$\rangle$,

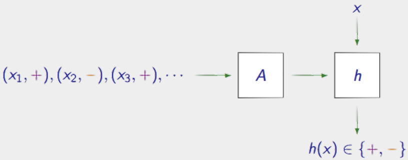
$\langle$name=Marcia, age=15, sex=F$\rangle$,

$\langle$name=Bobby , age=6 , sex=M$\rangle$,

$\langle$name=Jan , age=12, sex=F$\rangle$,

$\langle$name=Peter , age=13, sex=M$\rangle$,

Binary Classifier

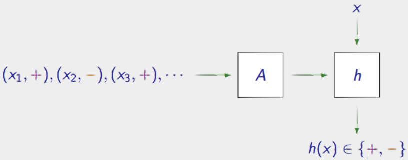


Reduction

Multiclass Data

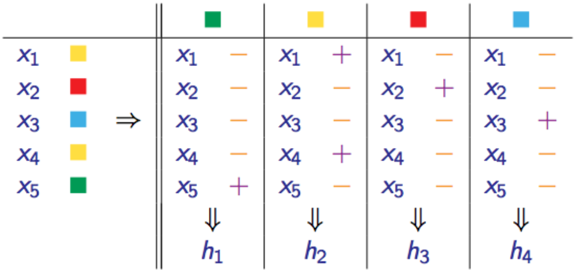
$\langle$name=Cindy , age=5 , sex=F$\rangle$,
$\langle$name=Marcia, age=15, sex=F$\rangle$,
$\langle$name=Bobby , age=6 , sex=M$\rangle$,
$\langle$name=Jan , age=12, sex=F$\rangle$,
$\langle$name=Peter , age=13, sex=M$\rangle$,

Binary Classifier



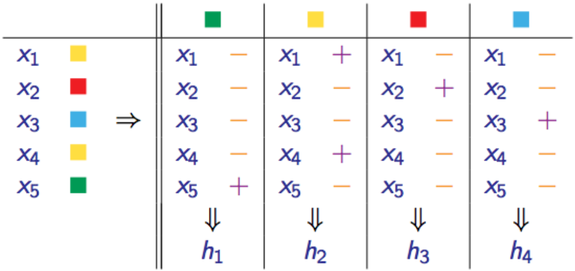
Goal: Multiclass Classifier

One-Against-All



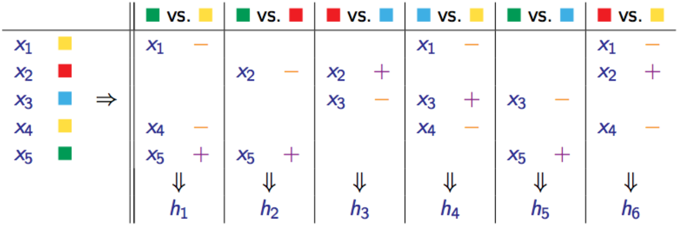
- Break k-class problem into k binary problems and solve separately
- Combine predictions: evaluate all h's, hope exactly one is $+$ (otherwise, take highest confidence)

One-Against-All



- Break k-class problem into k binary problems and solve separately
- Combine predictions: evaluate all h's, hope exactly one is $+$ (otherwise, take highest confidence)
- Incorrect prediction if only one is wrong

All-Pairs (Friedman; Hastie & Tibshirani)



- One binary problem for each pair of classes
- Take class with most positives and least negatives
- Faster and more accurate than one-against-all

Time Comparison

Assume training time is $\mathcal{O}(m^\alpha)$ and test time is $\mathcal{O}(c_t)$

	Training	Testing
OVA	$\mathcal{O}(km^\alpha)$	$\mathcal{O}(kc_t)$
All-pairs	$\mathcal{O}\left(k^2\left(\frac{m}{k}\right)^\alpha\right)$	$\mathcal{O}\left(k^2c_t\right)$

Time Comparison

Assume training time is $\mathcal{O}\left(m^\alpha\right)$ and test time is $\mathcal{O}\left(c_t\right)$

	Training	Testing
OVA	$\mathcal{O}\left(km^\alpha\right)$	$\mathcal{O}\left(kc_t\right)$
All-pairs	$\mathcal{O}\left(k^2\left(\frac{m}{k}\right)^\alpha\right)$	$\mathcal{O}\left(k^2c_t\right)$

OVA better for testing time, all-pairs better for training. (All-pairs usually better for performance.)

Error Correcting Output Codes (Dietterich & Bakiri)

- Reuce to binary using "coding" matrix

M	1	2	3	4	5
	+	−	+	−	+
	−	−	+	+	+
	+	+	−	−	−
	+	+	+	+	−

Error Correcting Output Codes (Dietterich & Bakiri)

- Reuce to binary using “coding” matrix
- Train classifier for each bit

			1		2		3		4		5	
x_1			x_1	$-$	x_1	$-$	x_1	$+$	x_1	$+$	x_1	$+$
x_2			x_2	$+$	x_2	$+$	x_2	$-$	x_2	$-$	x_2	$-$
x_3		$\Rightarrow$	x_3	$+$	x_3	$+$	x_3	$+$	x_3	$+$	x_3	$-$
x_4			x_4	$-$	x_4	$-$	x_4	$+$	x_4	$+$	x_4	$+$
x_5			x_5	$+$	x_5	$-$	x_5	$+$	x_5	$-$	x_5	$+$
			$\Downarrow$		$\Downarrow$		$\Downarrow$		$\Downarrow$		$\Downarrow$	
			h_1		h_2		h_3		h_4		h_5	

Error Correcting Output Codes (Dietterich & Bakiri)

- Reuce to binary using "coding" matrix
- Train classifier for each bit

			1	2	3	4	5
x_1			x_1 $-$	x_1 $-$	x_1 $+$	x_1 $+$	x_1 $+$
x_2			x_2 $+$	x_2 $+$	x_2 $-$	x_2 $-$	x_2 $-$
x_3		$\Rightarrow$	x_3 $+$	x_3 $+$	x_3 $+$	x_3 $+$	x_3 $-$
x_4			x_4 $-$	x_4 $-$	x_4 $+$	x_4 $+$	x_4 $+$
x_5			x_5 $+$	x_5 $-$	x_5 $+$	x_5 $-$	x_5 $+$
			$\Downarrow$	$\Downarrow$	$\Downarrow$	$\Downarrow$	$\Downarrow$
			h_1	h_2	h_3	h_4	h_5

- Choose closest row of coding matrix to predict

ECOC

- If rows of M are far apart, will be robust to error
- Much faster if k is large
- Disadvantage: binary problems may be unnatural

That's it for classification!

- You can implement multiple forms of classification
- Derive theoretical bounds for many classification tasks
- Today is bridge to the future: classification foundation of other ML tasks