Classification

Jordan Boyd-Graber
University of Maryland
MULTICLASS

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

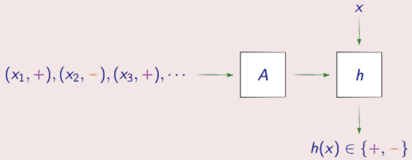
```
⟨name=Cindy , age=5 , sex=F⟩,
⟨name=Marcia, age=15, sex=F⟩,
⟨name=Bobby , age=6 , sex=M⟩,
⟨name=Jan   , age=12, sex=F⟩,
⟨name=Peter , age=13, sex=M⟩,
```

Reduction

Multiclass Data

⟨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⟩,

Binary Classifier

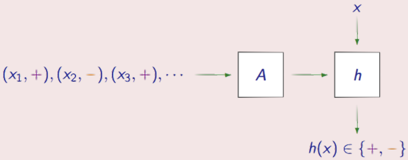


Reduction

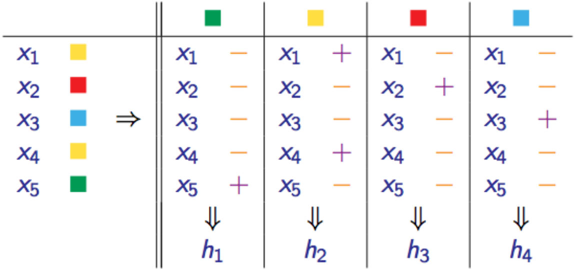
Multiclass Data

⟨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⟩,

Binary Classifier

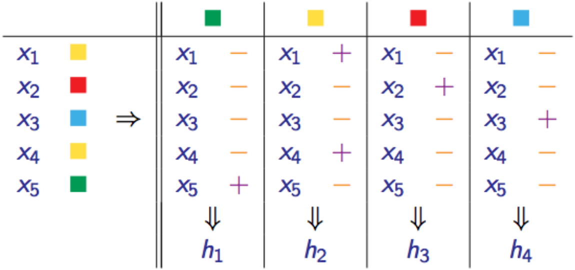


One-Against-All



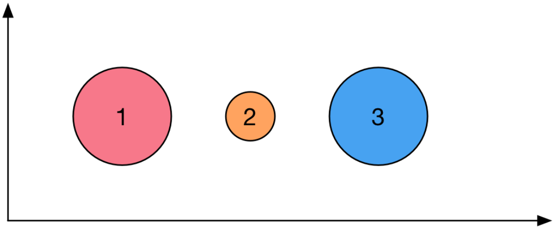
- Break k-class problem into k binary problems and solve separately
- Combine predictions: evaluate all h's, hope exactly one is $+$ (otherwise, take highest confidence)

One-Against-All

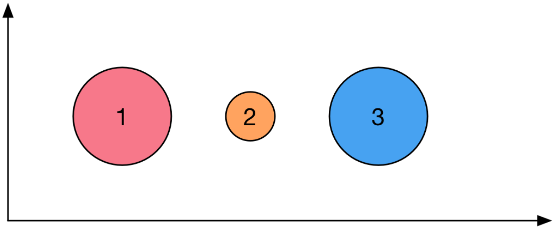


- Break k-class problem into k binary problems and solve separately
- Combine predictions: evaluate all h's, hope exactly one is $+$ (otherwise, take highest confidence)
- Incorrect prediction if only one is wrong

Does one vs. all work here?

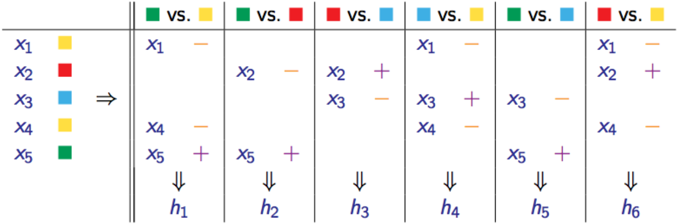


Does one vs. all work here?



Discriminating between class 2 and the rest of the classes, the optimal halfspace would be the all negative classifier

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

Time Comparison

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

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

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 $+$
	$\Downarrow$ h_1	$\Downarrow$ h_2	$\Downarrow$ h_3	$\Downarrow$ h_4	$\Downarrow$ 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

How to construct codes

- Exhaustive (if k small):length $2^{k-1}-1$
  - Row 1 has only ones
  - Row 2: 2^{k-2} zeros followed by $2^{k-2}-1$ ones
  - Row 3: 2^{k-3} zeros, 2^{k-3} ones, 2^{k-3} zeros, $2^{k-3}-1$ ones
  - ...

How to construct codes

- Exhaustive (if k small):length $2^{k-1}-1$
  - Row 1 has only ones
  - Row 2: 2^{k-2} zeros followed by $2^{k-2}-1$ ones
  - Row 3: 2^{k-3} zeros, 2^{k-3} ones, 2^{k-3} zeros, $2^{k-3}-1$ ones
  - ...
- Random codes: James and Hastie ’98 showed that this reduces variance through model averaging

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