Multi-Class Classification

Dan Sheldon

February 19, 2019

A Real Classification Problem

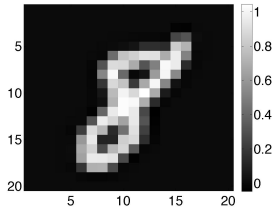
Classify handwritten digits.

$y \in \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$

We don't know how to solve this yet

Hand-written digit classification

Input: 20×20 grayscale image



Unroll the image into a vector

$$\begin{bmatrix} x_1 & x_{21} & \dots & x_{381} \\ x_2 & x_{22} & \dots & x_{382} \\ & & \vdots & \\ x_{20} & x_{40} & \dots & x_{400} \end{bmatrix}$$

Feature vector $\mathbf{x} \in \mathbb{R}^{400}$

$$\mathbf{x} = (x_1, \dots, x_{400})^T$$

Multi-class Classification

Input: $\mathbf{x} \in \mathbb{R}^m$ (continuous or discrete)
Labels: $y \in \{1, \dots, K\}$

Multi-class Classification

Input: $\mathbf{x} \in \mathbb{R}^m$ (continuous or discrete)
Labels: $y \in \{1, \dots, K\}$

Exercise: solve using logistic regression

- Use one or more binary ($y \in \{0, 1\}$) classifiers
- Hint: think about prediction first, then training.

One vs. All Classification

Learn a separate classifier for each class $c = 1, \dots, K$

One vs. All Classification

Learn a separate classifier for each class $c = 1, \ldots, K$
Labels for learning class c?

One vs. All Classification

Learn a separate classifier for each class $c = 1, \ldots, K$
Labels for learning class c?

Let $y_c^{(i)} = \begin{cases} 1 & \text{if } y^{(i)} = c \\ 0 & \text{otherwise} \end{cases}$

One vs. All Classification

Learn a separate classifier for each class $c = 1, \ldots, K$
Labels for learning class c?

Let $y_c^{(i)} = \begin{cases} 1 & \text{if } y^{(i)} = c \\ 0 & \text{otherwise} \end{cases}$

$\mathbf{x}^T$	y	y_1	y_2	y_3
$\cdots$	1	1	0	0
$\cdots$	2	0	1	0
$\cdots$	3	0	0	1

One vs. All Classification

Learn a separate classifier for each class $c = 1, \ldots, K$
Labels for learning class c?

Let $y_c^{(i)} = \begin{cases} 1 & \text{if } y^{(i)} = c \\ 0 & \text{otherwise} \end{cases}$

$\mathbf{x}^T$	y	y_1	y_2	y_3
$\cdots$	1	1	0	0
$\cdots$	2	0	1	0
$\cdots$	3	0	0	1

Training?

One vs. All Classification

Learn a separate classifier for each class $c = 1, \ldots, K$
Labels for learning class c?

Let $y_c^{(i)} = \begin{cases} 1 & \text{if } y^{(i)} = c \\ 0 & \text{otherwise} \end{cases}$

$\mathbf{x}^T$	y	y_1	y_2	y_3
$\cdots$	1	1	0	0
$\cdots$	2	0	1	0
$\cdots$	3	0	0	1

Training? for each class c, fit a binary classifier using training labels $y_c^{(i)}$ to get parameter vector $\boldsymbol{\theta}_c$

One vs. All Classification

Learn a separate classifier for each class $c = 1, \ldots, K$
Labels for learning class c?

Let $y_c^{(i)} = \begin{cases} 1 & \text{if } y^{(i)} = c \\ 0 & \text{otherwise} \end{cases}$

$\mathbf{x}^T$	y	y_1	y_2	y_3
$\cdots$	1	1	0	0
$\cdots$	2	0	1	0
$\cdots$	3	0	0	1

Training? for each class c, fit a binary classifier using training labels $y_c^{(i)}$ to get parameter vector $\boldsymbol{\theta}_c$

Prediction?

One vs. All Classification

Learn a separate classifier for each class $c = 1, \ldots, K$

Labels for learning class c?

Let $y_c^{(i)} = \begin{cases} 1 & \text{if } y^{(i)} = c \\ 0 & \text{otherwise} \end{cases}$

$\mathbf{x}^T$	y	y_1	y_2	y_3
$\cdots$	1	1	0	0
$\cdots$	2	0	1	0
$\cdots$	3	0	0	1

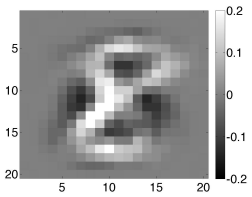
Training? for each class c, fit a binary classifier using training labels $y_c^{(i)}$ to get parameter vector $\boldsymbol{\theta}_c$

Prediction? make a prediction for each class and choose the one with *highest probability*

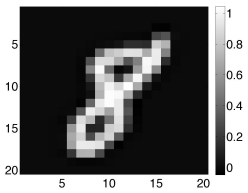
$$\text{predict } y = \text{argmax}_c \, h_{\boldsymbol{\theta}_c}(\mathbf{x})$$

Visualization

Format weight vector as an image:



$\boldsymbol{\theta}_8$



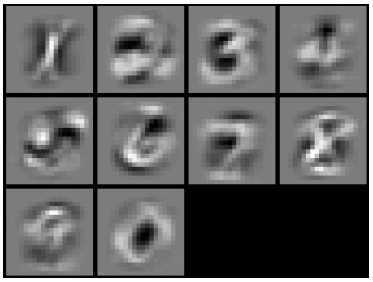
$\mathbf{x}$

Recall that

$$\text{Prediction} = \begin{cases} 1 & \boldsymbol{\theta}^T \mathbf{x} \geq 0 \\ 0 & \boldsymbol{\theta}^T \mathbf{x} < 0 \end{cases}$$

Dot product = multiply together corresponding pixels and add

Visualization: One vs. All



One vs. One

Fit a classifier for each pair of classes

One vs. One

Fit a classifier for each pair of classes

Labels for discriminating c from d?

One vs. One

Fit a classifier for each pair of classes

Labels for discriminating c from d?

Let $y_{cd}^{(i)} = \begin{cases} 1 & \text{if } y^{(i)} = c \\ 0 & \text{if } y^{(i)} = d \end{cases}$

One vs. One

Fit a classifier for each pair of classes

Labels for discriminating c from d?

Let $y_{cd}^{(i)} = \begin{cases} 1 & \text{if } y^{(i)} = c \\ 0 & \text{if } y^{(i)} = d \end{cases}$

$\mathbf{x}^T$	y	y_{12}	y_{13}	y_{23}
$\cdots$	1	1	1	-
$\cdots$	2	0	-	1
$\cdots$	3	-	0	0

One vs. One

Fit a classifier for each pair of classes

Labels for discriminating c from d?

Let $y_{cd}^{(i)} = \begin{cases} 1 & \text{if } y^{(i)} = c \\ 0 & \text{if } y^{(i)} = d \end{cases}$

$\mathbf{x}^T$	y	y_{12}	y_{13}	y_{23}
$\cdots$	1	1	1	-
$\cdots$	2	0	-	1
$\cdots$	3	-	0	0

Training? for each pair $c \neq d$, fit a binary classifier with labels $y_{cd}^{(i)}$ using **only examples from class c or d**

One vs. One

Fit a classifier for each pair of classes

Labels for discriminating c from d?

Let $y_{cd}^{(i)} = \begin{cases} 1 & \text{if } y^{(i)} = c \\ 0 & \text{if } y^{(i)} = d \end{cases}$

$\mathbf{x}^T$	y	y_{12}	y_{13}	y_{23}
$\cdots$	1	1	1	-
$\cdots$	2	0	-	1
$\cdots$	3	-	0	0

Training? for each pair $c \neq d$, fit a binary classifier with labels $y_{cd}^{(i)}$ using **only examples from class c or d**

- Result: parameter vector $\boldsymbol{\theta}_{cd}$

Prediction?

One vs. One

Fit a classifier for each pair of classes

Labels for discriminating c from d?

Let $y_{cd}^{(i)} = \begin{cases} 1 & \text{if } y^{(i)} = c \\ 0 & \text{if } y^{(i)} = d \end{cases}$

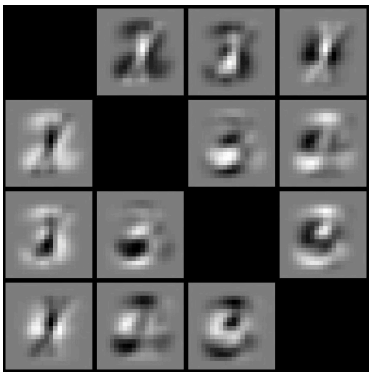
$\mathbf{x}^T$	y	y_{12}	y_{13}	y_{23}
$\cdots$	1	1	1	-
$\cdots$	2	0	-	1
$\cdots$	3	-	0	0

Training? for each pair $c \neq d$, fit a binary classifier with labels $y_{cd}^{(i)}$ using **only examples from class c or d**

- Result: parameter vector $\boldsymbol{\theta}_{cd}$

Prediction? voting scheme. Explain on board.

Visualization: One vs. One



Advanced Topic: Error-Correcting Output Codes

Learn a separate classifier for each bit of codeword

Class	Code Word														
	f_0	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}	f_{12}	f_{13}	f_{14}
0	1	1	0	0	0	0	1	0	1	0	0	1	1	0	1
1	0	0	1	1	1	1	0	1	0	1	1	0	0	1	0
2	1	0	0	1	0	0	0	1	1	1	1	0	1	0	1
3	0	0	1	1	0	1	1	1	0	0	0	0	1	0	1
4	1	1	1	0	1	0	1	1	0	0	1	0	0	0	1
5	0	1	0	0	1	1	0	1	1	1	0	0	0	0	1
6	1	0	1	1	1	0	0	0	0	1	0	1	0	0	1
7	0	0	0	1	1	1	1	0	1	0	1	1	0	0	1
8	1	1	0	1	0	1	1	0	0	1	0	0	0	1	1
9	0	1	1	1	0	0	0	0	1	0	1	0	0	1	1

Dietterich and Bakiri 1995
(Possible project idea)