MULTICLASS CONTINUED AND RANKING

David Kauchak
CS 451 – Fall 2013

Admin
Assignment 4
Course feedback
Midterm

Java tip for the day
private vs. public vs. protected

Debugging tips
Multiclass classification

Examples:

- Label: Same setup where we have a set of features for each example
- Apple: Rather than just two labels, now have 3 or more

Black box approach to multiclass

Abstraction: we have a generic binary classifier, how can we use it to solve our new problem

Can we solve our multiclass problem with this?

Approach 1: One vs. all (OVA)

Training: for each label L, pose as a binary problem
- all examples with label L are positive
- all other examples are negative

OVA: linear classifiers (e.g. perceptron)
OVA: linear classifiers (e.g. perceptron)

apple vs. not

How do we classify?

banana vs. not

pineapple vs. not

OVA: linear classifiers (e.g. perceptron)

apple vs. not

How do we classify?

banana vs. not

pineapple vs. not
OVA: linear classifiers (e.g. perceptron)

How do we classify?

banana vs. not
apple vs. not
pineapple vs. not

OVA: classify

Classify:
- If classifier doesn’t provide confidence (this is rare) and there is ambiguity, pick one of the ones in conflict
- Otherwise:
  - pick the most confident positive
  - if none vote positive, pick least confident negative

What does the decision boundary look like?
OVA: classify, perceptron

Classify:
- If classifier doesn’t provide confidence (this is rare) and there is ambiguity, pick majority in conflict
- Otherwise:
  - pick the most confident positive
  - if none vote positive, pick least confident negative

How do we calculate this for the perceptron?

Approach 2: All vs. all (AVA)

Training:
For each pair of labels, train a classifier to distinguish between them

for $i = 1$ to number of labels:
  for $k = i+1$ to number of labels:
    train a classifier to distinguish between label$_i$ and label$_k$
    - create a dataset with all examples with label$_i$ labeled positive and all examples with label$_k$ labeled negative
    - train classifier on this subset of the data

\[
prediction = b + \sum_{i=1}^{n} w_i f_i
\]

Distance from the hyperplane
AVA training visualized

AVA classify

AVA classify

AVA classify

AVA classify

In general?

What class?

To classify example \( e \), classify with each classifier \( f_{jk} \).

We have a few options to choose the final class:

- Take a majority vote
- Take a weighted vote based on confidence
  - \( y = f_k(e) \)
  - \( \text{score}_k = y \)
  - \( \text{score}_k = y \)

Here we're assuming that \( y \) encompasses both the prediction (+1,-1) and the confidence, i.e. \( y \) = prediction * confidence.
AVA classify

Take a weighted vote based on confidence

\[ y = f_j(e) \]

- \( \text{score}_j \leftarrow y \)
- \( \text{score}_k \rightarrow y \)

If \( y \) is positive, classifier thought it was of type \( j \):
- raise the score for \( j \)
- lower the score for \( k \)

if \( y \) is negative, classifier thought it was of type \( k \):
- lower the score for \( j \)
- raise the score for \( k \)

OVA vs. AVA

Train/classify runtime?

Error? Assume each binary classifier makes an error with probability \( \varepsilon \)

OVA vs. AVA

Train time:
AVA learns more classifiers, however, they’re trained on much smaller data this tends to make it faster if the labels are equally balanced

Test time:
AVA has more classifiers

Error (see the book for more justification):
- AVA trains on more balanced data sets
- AVA tests with more classifiers and therefore has more chances for errors
- Theoretically:
  - OVA: \( \varepsilon \) (number of labels -1)
  - AVA: 2 \( \varepsilon \) (number of labels -1)

Approach 3: Divide and conquer

Pros/cons vs. AVA?
Multiclass summary

If using a binary classifier, the most common thing to do is OVA

Otherwise, use a classifier that allows for multiple labels:
- DT and k-NN work reasonably well
- We’ll see a few more in the coming weeks that will often work better

Multiclass evaluation

<table>
<thead>
<tr>
<th>label</th>
<th>prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>apple</td>
<td>orange</td>
</tr>
<tr>
<td>orange</td>
<td>orange</td>
</tr>
<tr>
<td>apple</td>
<td>apple</td>
</tr>
<tr>
<td>banana</td>
<td>pineapple</td>
</tr>
<tr>
<td>banana</td>
<td>banana</td>
</tr>
<tr>
<td>pineapple</td>
<td>pineapple</td>
</tr>
</tbody>
</table>

Accuracy: 4/6

Multiclass evaluation imbalanced data

<table>
<thead>
<tr>
<th>label</th>
<th>prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>apple</td>
<td>apple</td>
</tr>
<tr>
<td>banana</td>
<td>pineapple</td>
</tr>
<tr>
<td>banana</td>
<td>banana</td>
</tr>
<tr>
<td>pineapple</td>
<td>pineapple</td>
</tr>
</tbody>
</table>

Any problems? Data imbalance!
Macroaveraging vs. microaveraging

Microaveraging:

- Average over examples (this is the "normal" way of calculating)

Microaveraging:

- Calculate evaluation score (e.g., accuracy) for each label, then average over labels

What effect does this have?
Why include it?

Macroaveraging vs. microaveraging

Macroaveraging:

- Calculate evaluation score (e.g., accuracy) for each label, then average over labels

- Puts more weight/emphasis on rarer labels
- Allows another dimension of analysis
Confusion matrix

entry \((i, j)\) represents the number of examples with label \(i\) that were predicted to have label \(j\)

another way to understand both the data and the classifier

\[
\begin{array}{cccccc}
& \text{Classic} & \text{Country} & \text{Disco} & \text{Hiphop} & \text{Jazz} & \text{Rock} \\
\text{Classic} & 86 & 2 & 0 & 4 & 18 & 1 \\
\text{Country} & 1 & 57 & 5 & 1 & 12 & 13 \\
\text{Disco} & 0 & 6 & 55 & 4 & 0 & 5 \\
\text{Hiphop} & 0 & 15 & 28 & 90 & 4 & 18 \\
\text{Jazz} & 7 & 1 & 0 & 0 & 37 & 12 \\
\text{Rock} & 6 & 19 & 11 & 0 & 27 & 48 \\
\end{array}
\]

Multilabel vs. multiclass classification

• Is it edible?
• Is it sweet?
• Is it a fruit?
• Is it a banana?

Any difference in these labels/categories?

Multilabel vs. multiclass classification

• Is it edible?
• Is it sweet?
• Is it an apple?
• Is it a banana?

Different structures

Nested/Hierarchical  Exclusive/Multiclass  General/Structured
### Multiclass vs. Multilabel

<table>
<thead>
<tr>
<th>Multiclass</th>
<th>Multilabel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Each example has one label and exactly one label.</td>
<td>Each example has zero or more labels. Also called annotation.</td>
</tr>
</tbody>
</table>

**Multilabel applications:**

- Image annotation
- Document topics
- Labeling people in a picture
- Medical diagnosis

### Which of our approaches work for multilabel?