Multiclass Classification

COMP 135 Intro to Machine Learning
Liping Liu

slides modified from Roni Khardon’s with permission

Multiple classes

• Each label \( y_i \in \{1, ..., L\} \)
• Need classifiers \( f: R^d \to \{1, ..., L\} \)
• Methods
  - Extending binary classifiers
  - Combining binary classifiers

Extend existing classifiers

• Method 1: extend the predictive probability \( p(y|x) \) to a categorical distribution

• Method 2: output \( L \) scores, \( \eta_1, \eta_2, ..., \eta_L \), from an instance \( x_i \), and then choose the class with the largest score

Extending predictive probability

• KNN
  - Estimate \( p(y|x) \) from neighbors by counting
• Trees
  - Calculate information gain from multiple classes (need to calculate entropy of classes within each node)
  - Estimate \( p(y|x) \) from each leaf by counting

Extending predictive probability

• Logistic regression
  - Use a vector \( w_\ell \) for each class
  - Extend the predictive probability as
    \[
    p(y = \ell|x) = \frac{\exp(\eta_\ell)}{\sum_{\ell'=1}^{L} \exp(\eta_{\ell'})}
    \]
    \[
    = \frac{\exp(w_\ell x_i + w_0)}{\sum_{\ell'=1}^{L} \exp(w_{\ell'} x_i + w_0)}
    \]

Multiple class scores

• SVM
  - Use a vector \( w_\ell \) for each class
  - Soft loss \( \xi_i \) is defined as below when the true label is \( y_i = \ell \)
    \[
    w_\ell x_i - w_{\ell'} x_i \geq 1 - \xi_i, \forall \ell' \neq \ell
    \]
  - Equivalently
    \[
    w_\ell x_i - \max_{\ell' \neq \ell} w_{\ell'} x_i \geq 1 - \xi_i
    \]
  - Principle: the most confusing label should get a score with margin 1, or suffer loss

Ensemble classifiers

- Random forests & BAgging
  - Vote to decide the class label
- Gradient boosting
  - Construct \( L \) sequences of trees, each one output a score \( \eta^\ell(x) \) for a label \( \ell \) from instance \( x \)
- AdaBoost
  - Need to consider how to update weights of data points (out of the scope of this course)

One-vs-All classification

- For each class, train a classifier
  - Positives: instances from this class
  - Negatives: instances from all other classes
- Choose the class with the largest predictive score
- Need to train \( L \) binary classifiers

One-vs-One classification

- For each pair of classes, train a binary classifier to classify the two classes
  - Need to train \( \frac{L(L-1)}{2} \) classifiers
- Classifiers vote to decide final class label

Evaluation and diagnosis

- Confusion matrix
- Cost matrix: different cost for different types misclassifications

One-vs-One classification

- For each pair of classes, train a binary classifier to classify the two classes
  - Need to train \( \frac{L(L-1)}{2} \) classifiers
- Classifiers vote to decide final class label

Error-Correcting Output Coding

- Convert a multiclass classification problem into binary classification problems

Other types of classification problems

- Each instance can take multiple labels
  - Multilabel learning
  - Labels of single instances have structures