Text Categorization: Methods

ChengXiang "Cheng" Zhai
Department of Computer Science
University of Illinois at Urbana-Champaign

Overview

- What is text categorization?
- Why text categorization?
- **How to do text categorization?**
  - Generative probabilistic models
  - Discriminative approaches
- How to evaluate categorization results?

Categorization Methods: Manual

哪种类别词出现的多，就将其划分哪种

- Determine the category based on rules that are carefully designed to reflect the domain knowledge about the categorization problem
- Works well when
  - The categories are very well defined
  - Categories are easily distinguished based on surface features in text (e.g., special vocabulary is known to only occur in a particular category)
  - Sufficient domain knowledge is available to suggest many effective rules
- Problems
  - Labor intensive ➔ doesn't scale up well
  - Can't handle uncertainty in rules; rules may be inconsistent ➔ not robust
- Both problems can be solved/alleviated by using machine learning

Categorization Methods: "Automatic"

- Use **human experts** to
  - Annotate data sets with **category labels** ➔ Training data
  - Provide a set of **features** to represent each text object that can potentially provide a "clue" about the category
- Use **machine learning** to learn "soft rules" for categorization from the training data
  - Figure out **which features are most useful** for separating different categories
  - **Optimally combine the features** to **minimize the errors** of categorization on the training data
  - The trained classifier can then be applied to a new text object to predict the most likely category (that a human expert would assign to it)

Machine Learning for Text Categorization

- **General setup**: Learn a classifier f: X➔Y
  - Input: X = all text objects; Output: Y = all categories
  - Learn a classifier function, f: X➔Y, such that f(x)=y ∈Y gives the correct category for x∈X ("correct" is based on the training data)
- **All methods**
  - Rely on discriminative features of text objects to distinguish categories
  - Combine multiple features in a weighted manner
  - Adjust weights on features to minimize errors on the training data
- **Different methods** tend to vary in
  - Their way of measuring the errors on the training data (may optimize a different objective/loss/cost function)
  - Their way of combining features (e.g., linear vs. non-linear)

Generative vs. Discriminative Classifiers

- **Generative** classifiers (learn **what the data “looks” like in each category**)
  - Attempt to model p(X,Y) = p(Y)p(X|Y) and compute p(Y|X) based on p(X|Y) and p(Y) by using Bayes Rule
  - Objective function is likelihood, thus indirectly measuring training errors
  - E.g., Naïve Bayes
- **Discriminative** classifiers (learn **what features separate categories**)

机器学习方法

  - Attempt to model p(Y|X) directly
  - Objective function directly measures errors of categorization on training data
  - E.g., Logistic Regression, Support Vector Machine (SVM), k-Nearest Neighbors (kNN)