CS769 Advanced NLP Homework 4

Due 3/10/2010

1 SVM-light Software

In `http://pages.cs.wisc.edu/~jerryzhu/cs769/dataset/jokes/`, there are two files. Each line is a document. All documents in `positive.txt` have label $y = 1$, and in `negative.txt` have label $y = -1$. Do not further tokenize the files – treat each string separated by whitespace as a word. Punctuations are words too.

For this question we will use SVM-light, available at `http://svmlight.joachims.org/`. Download the code and study the manual.

Q1.[5pt] Use bag-of-word count vector as the feature representation for documents. Convert positive.txt and negative.txt into SVM-light format (feature:value pairs). Show the first document in positive.txt, in SVM-light format.

Q2.[5pt] We will perform 10-fold cross validation on the dataset. *Briefly* describe how you generate the 10 random folds (i.e., deciding which document goes to which fold).

Q3.[5pt] Use all default parameters (and linear kernel) in SVM-light. Perform 10-fold CV. List the 10 accuracies, one for each test fold.

Q4.[5pt] Repeat Q3 (on the same folds), except this time use a polynomial kernel $(s\ a^\top b + c)^d$ with $s = 1, c = 1, d = 2$. List the 10 accuracies.

Q5.[5pt] Perform paired t-test between the linear and polynomial kernel results. Is the difference between them statistically significant at 5% level?

Q6.[5pt] Design at least one new feature for the documents. Describe your new feature, including why you think it might improve classification, and how to compute it (if non-trivial).

Q7.[5pt] Add (append) your new feature(s) to the existing bag-of-word feature vectors. Use all default parameters (and linear kernel) in SVM-light. Perform 10-fold CV. List the 10 accuracies. Perform paired t-test against Q3. Does your new feature(s) improve classification?

2 SVM with Must-Link and Cannot-Link

You are given a training set $(x_1, y_1) \dots (x_n, y_n)$, where each document $x_i \in \mathbb{R}^d$, and the label $y_i \in \{-1, 1\}$. In addition, there are 4 more documents: $x_{n+1}, x_{n+2}, x_{n+3}, x_{n+4}$. You do not know the label for these 4 documents. However, an expert has told you that x_{n+1} and x_{n+2} *must be in the same class* (either all positive or all negative, which is known as a "must-link"), and that x_{n+3} and x_{n+4} cannot be in the same class (a "cannot-link").

> **Q8.[15pt]** Formulate the primal SVM problem *with the must-link and cannot-link*. Use the primal SVM with slack variables (equations (24)-(26) in the SVM lecture notes) as your starting point. Briefly explain how your formulation incorporates the must-link and cannot-link.

3 The EM Algorithm for Gaussian Mixture Models

Q9.[15pt] Write down the EM algorithm for Gaussian Mixture Models with two components, where each component is a Gaussian distribution. Assume the (unlabeled) data is $x_1, \dots, x_n \in \mathbb{R}$. The parameters are π_1, π_2 the component weights (which sum to 1), μ_1, σ_1^2 the mean and variance for the first component, and μ_2, σ_2^2 the mean and variance for the second component. Be sure to clearly mark the E-step and the M-step.

Q10.[20pt] We have provided a small Matlab dataset $x_1, \dots, x_n$ in `http://pages.cs.wisc.edu/~jerryzhu/cs769/dataset/x.mat`. Starting from the following initial parameters: $\pi_1 = 0.5, \mu_1 = -1.4781, \sigma_1 = 1, \mu_2 = 0.8800, \sigma_2 = 1, \pi_2 = 0.5$, implement the EM algorithm. What to hand in: Print out the log likelihood and the parameters in each iteration.

4 Logistic Regression

You are given a training set $(x_1, y_1) \dots (x_n, y_n)$, where each document $x_i \in \mathbb{R}^d$, and the label $y_i \in \{-1, 1\}$. Consider the *negative logistic loss* function

$$c(\theta) = \sum_{i=1}^{n} \log_2(1 + \exp(-y_i \theta^\top x_i)), \tag{1}$$

where $\theta \in \mathbb{R}$ is an arbitrary weight vector (not necessarily the solution to logistic regression).

> **Q11.[15pt]** Prove that $c(\theta)$ upper bounds the number of classification mistakes θ makes on the training set. Hint: θ makes a mistake on x_i if $\text{sgn}(\theta^\top x_i) \neq y_i$.