Übung zur Vorlesung

Maschinelles Lernen 2

Sommersemester 2013

Abteilung Maschinelles Lernen
Institut für Softwaretechnik und theoretische Informatik
Fakultät IV, Technische Universität Berlin
Prof. Dr. Klaus-Robert Müller
Email: klaus-robert.mueller@tu-berlin.de

Exercise Sheet 6

Due **May 27**, 9am local time on ISIS

1. **String Kernels (40+30 points)**
   Fix a set of characters $\mathcal{A}$. The set of strings with characters in $\mathcal{A}$ is denoted by $\mathcal{A}^*$; for a string $x \in \mathcal{A}^*$, we denote its length by $|x|$. If $s, x \in \mathcal{A}^*$, one denotes the number of occurrences of s as a substring of x by
   $$\#(s \sqsubseteq x).$$
   For example, $\#(\text{na} \sqsubseteq \text{ananasanna}) = 3$. In general, a string kernel on $\mathcal{A}^*$ is defined as follows:
   $$k(.,.) : \mathcal{A}^* \times \mathcal{A}^* \longrightarrow \mathbb{R}$$
   $$(x, y) \mapsto \sum_{s \in \mathcal{A}^*} \#(s \sqsubseteq x) \cdot \#(s \sqsubseteq y) \cdot w(s),$$
   where $w(.) : \mathcal{A}^* \to \mathbb{R}$ is some kernel and substring specific weighting function.

   (a) A string kernel is called bag-of-words-kernel when $w(s) \neq 0$ if and only if s is a word in a natural language, e.g., German. That is, the sum in the kernel ranges over the words of a natural language only (e.g., ziegenpeter or ananasanna). One possible assignment is $w(s) = 1$ if s is a word, the *unweighted bag-of-words*; a different weighting is given by the so-called inverse document frequency (IDF), which gives rise to the IDF-kernel. That is, given N documents $D_1, \ldots, D_N$ (in the same language), one considers for each word s the number of documents
   $$\mathrm{DF}(s) = \#\{k \; ; \; 1 \leq k \leq N, s \sqsubseteq D_k\}$$
   in which the word s occurs - the so-called document frequency, and then sets
   $$w(s) = \mathrm{IDF}(s) = \log N - \log \mathrm{DF}(s)$$
   (de facto, that's semantically the log of the inverse document frequency).

   Prove that for both assignments of $w(s)$ (i.e., unweighted bag-of-words, and IDF), the so-defined functions $k(.,.)$ are in fact kernels. Recall: a function $k(.,.)$ is a kernel function if and only if all possible kernel matrices are symmetric positive semi-definite.

   (b) A string kernel is called n-spectrum kernel when $w(s) = 1$ if $|s| = n$ and $w(s) = 0$ else. A string kernel is called blended n-spectrum kernel, or mixed spectrum kernel when $w(s) = 1$ if $|s| \leq n$ and $w(s) = 0$ else. *Prove* that these are indeed kernels (you can use (a) for that).

   Calculate the kernel matrices for the n-spectrum kernel and the blended n-spectrum kernel over the latin alphabet for the special case of $n = 3$ and the data set
   $$\text{ananas, anna, natter, otter, otto}$$

2. **Tree kernels (30 points)**
   In the lecture, the subsequence kernel was presented which uses suffix trees for efficient representation and calculation of the kernel matrix. *Prove* the following basic results about suffix trees:
   (a) A string x contains $\mathcal{O}(|x|^2)$ substrings.
   (b) A substring w of x can be reached in $\mathcal{O}(|w|)$ in the suffix tree of x.
   (c) The suffix tree of x can be stored in $\mathcal{O}(|x|)$ space.

Please ask questions in the ISIS discussion forums for Machine Learning 2:
`https://www.isis.tu-berlin.de/course/view.php?id=8005`