Übung zur Vorlesung

Maschinelles Lernen 2

Sommersemester 2009

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

Blatt 9

Abgabe 21. Juni 2009 bis 13 Uhr, Ausarbeitung im Sekretariat FR6052, oder bei Mikio Braun, FR6058.

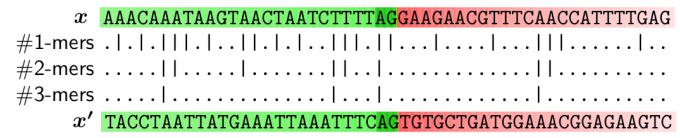
Aufgaben

In der Vorlesung wurden maschinelle Lernmethoden in der Bioinformatik anhand des Beispiels *mGene* besprochen. Die Gensuchmaschine *mGene* funktioniert nach einem 2-Schichten Prinzip. In der ersten Schicht werden biologische Signale durch Support-Vector-Maschinen (SVM) gelernt und in der zweiten Schicht werden die Ausgaben der SVMs als Eingaben zum Strukturlernen verwendet, um eine genaue Segmentierung in "nicht-Gen", "Gen" (Intron,Exon, usw.) zu erhalten. In der Übung sollen der Spektrum und der Weighted Degree Kernel—zwei Kerne aus der Gruppe der String-Kerne—weiter vertieft werden, da sie eine entscheidende Rolle beim Lernen der ersten Schicht in *mGene* spielen.

1. **(10 Punkte)** Der Weighted Degree Kern ist wie folgt definiert (Beispiel siehe Bild):

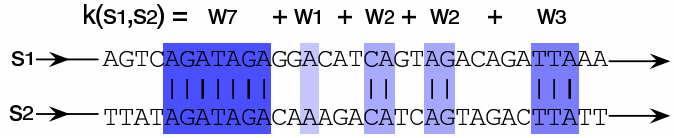
$$\mathrm{k}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{k=1}^{K} \beta_k \sum_{l=1}^{L-k+1} \mathbf{I}(\boldsymbol{u}_{k,l}(\boldsymbol{x}_i) = \boldsymbol{u}_{k,l}(\boldsymbol{x}_j)). \tag{1}$$

   Dabei ist $\boldsymbol{u}_{k,l}(\boldsymbol{x})$ ein String der Länge k welcher an Position l der Sequenz $\boldsymbol{x}$ anfängt und $\mathbf{I}(.)$ die Indikatorfunktion, welche 1 ist wenn ihr Argument wahr ist, sonst 0. Als Gewichtung wird $\beta_k = 2\frac{K-k+1}{K(K+1)}$ gewählt.



   Zeige für den Spektrum und den Weighted Degree (WD) Kern, dass es sich um positiv-definite Mercerkerne handelt.

2. **(20 Punkte)** Alternativ kann man den WDK auch in der "Block"-Formulierung betrachten. Hierbei werden zunächst längste gemeinsame Teilstrings identifiziert, und dann entsprechende Gewichte für die länge dieser Teilstrings aufaddiert (im Bild dargestellt).



   Berechne die Gewichte w_L, die den in Aufgabe 1 definierten WDK mit Gewichten β_k ergeben. Was ist zu beachten, wenn die Teilstrings länger werden als K?