Übung zur Vorlesung

Maschinelles Lernen 2

Sommersemester 2007

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 6

Abgabe 19. April 2007 in der Vorlesung, oder bis 23:59:59 Uhr bei `mikio@cs.tu-berlin.de`

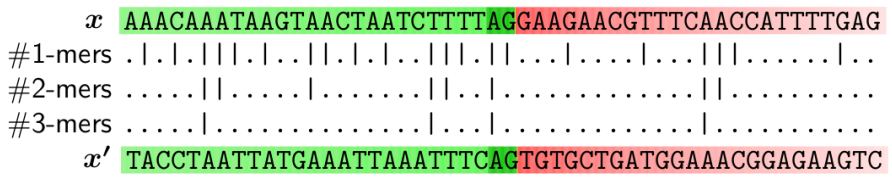
Aufgaben

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. Der Weighted Degree Kern ist wie folgt definiert (Beispiel siehe Bild):

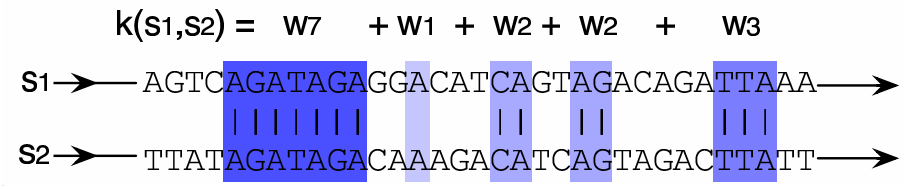
$$\mathrm{k}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{k=1}^{K} \beta_k \sum_{l=1}^{L-k+1} \mathbf{I}(\boldsymbol{u}_{k,l}(\boldsymbol{x}_i) = \boldsymbol{u}_{k,l}(\boldsymbol{x}_j)). \tag{1}$$

   Dabei ist $\boldsymbol{u}_{k,l}(\boldsymbol{x})$ ein String der Länge k welcher an Position l der Sequenz $\boldsymbol{x}$ anfängt und $I(.)$ 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 Mercer Kerne handelt. **(10 Punkte)**

2. Berechne die Gewichte in der "Block"-Formulierung des WD Kerns (im Bild dargestellt), so dass dieser äquivalent ist. **(10 Punkte)**



3. Lade die SHOGUN-toolbox für Bioinformatik/Maschinelles Lernen herunter, `http://www.shogun-toolbox.org` und trainiere mit Hilfe des WD-Kerns einen Spleissstellenerkenner. Führe dabei eine Kreuzvalidierung für verschiedene Regularisierungsparameter C und WD-Kern Grade K durch. Die zugehörigen Daten findest Du auf der Vorlesungsseite. Gib die ermittelten besten Parameter für C und K, Deine Performance auf den Daten, sowie das Modell ab. **(20 Punkte)**