Übung zur Vorlesung

Maschinelles Lernen 2

Sommersemester 2011

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

Blatt 9

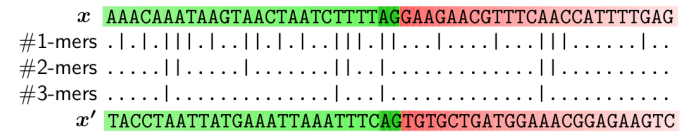
Abgabe bis Montag, 27. Juni 2011, 12:00 Uhr, Briefkasten bei Raum FR6061

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.

1. **(10 Punkte)** In der Aufgabe 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. 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)** In dieser Aufgabe soll der WDK zur Detektierung von Splicestellen verwendet werden. Wir verwenden hierzu die SVM Implementierung SVM^{light} (http://svmlight.joachims.org/).

   SVM^{light} unterstützt von Hause aus nicht den WDK, daher werde wir explizit die Merkmale

$$\phi_{w,k,l} = \sqrt{\beta_k}\mathbf{I}(\boldsymbol{u}_{k,l}(\boldsymbol{x}) = w).$$

   für alle $w \in \{A, C, G, T\}^k$, $k \in \{1, \ldots, K\}$, $l \in \{1, \ldots |\boldsymbol{x}| - k\}$ berechnen und dann einen linearen Kern verwenden.

   a) **10 Punkte** Vervollständige im Programmskelett die Funktion `write_wdk_features`, die die WDK Merkmale wie oben beschrieben in eine SVM^{light}-Datei schreibt.

   b) **10 Punkte** Trainiere SVM^{light} auf dem Splice-Datensatz für $K = \{1, 2, 3\}$ mit Regularisierungskonstante $C = \{0.001, 0.01, 0.1, 1, 10\}$ und messe die Vorhersagegenauigkeit auf dem Trainings- und Testdatensatz. Erzeuge für jedes k einen Plot, der diese Genauigkeiten für Training und Test gegen C plotte.

      Gib die Ergebnisse als Tabelle zusammen mit diesen drei Plots ab.

```
function sheet09

X = textread('splice-train-data.txt', '%s');
Y = load('splice-train-label.txt');
XE = textread('splice-test-data.txt', '%s');
YE = load('splice-test-label.txt');

tic
fprintf('Writing training features for k = 1\n');
write_wdk_features('wdk1-train.txt', 1, X, Y);
fprintf('Writing test features for k = 1\n');
write_wdk_features('wdk1-test.txt', 1, XE, YE);
toc

tic
fprintf('Writing training features for k = 2\n');
write_wdk_features('wdk2-train.txt', 2, X, Y);
fprintf('Writing test features for k = 2\n');
write_wdk_features('wdk2-test.txt', 2, XE, YE);
toc

tic
fprintf('Writing training features for k = 3\n');
write_wdk_features('wdk3-train.txt', 3, X, Y);
fprintf('Writing test features for k = 3\n');
write_wdk_features('wdk3-test.txt', 3, XE, YE);
toc

function beta = beta(K, k)
beta = 2 * (K - k + 1) / K / (K + 1);

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%
% Your solution below
%

% 1. write out weighted degree kernel features
% out into file FN up to degree K. (You should accept values of K = 1, 2,
% 3
function write_wdk_features(FN, K, X, Y)
% ...
```

Für Fragen zum Übungsblatt bitte in der Google Group http://groups.google.com/group/ml-tu registrieren und die Fragen an die Mailingliste stellen.