Übungen Sequenzanalyse I
Sommersemester 2007

Dr. Ivo Große

Institut für Informatik
Universität Halle

Blatt 5

Aufgabe 5.1 (8 Punkte)

Benutzen Sie den Datensatz `seq_1` zum Trainieren eines PWM-Mischmodells mit K Klassen und eines WAM-Mischmodells mit K Klassen, und benutzen Sie `seq_3` zum Trainieren eines PWM-Mischmodells mit L Klassen und eines WAM-Mischmodells mit L Klassen. Berechnen Sie für gegebenes K und L die Fläche unter der ROC-Kurve, $F(K, L)$, für jede der vier Modellkombinationen (PWM-Mischmodell für Donor und PWM-Mischmodell für Nicht-Donor, PWM-Mischmodell für Donor und WAM-Mischmodell für Nicht-Donor, WAM-Mischmodell für Donor und PWM-Mischmodell für Nicht-Donor, WAM-Mischmodell für Donor und WAM-Mischmodell für Nicht-Donor) für die Datensätze `seq_2` und `seq_4`. Schreiben Sie für $K = 1, 2, ..., 5$ und $L = 1, 2, ..., 5$ die Werte $F(K, L)$ in eine 5×5 Matrix. Vergleichen Sie die vier 5×5 Matrizen – eine Matrix füer jede der vier Modellkombinationen – miteinander. Welche Modellkombination liefert die genaueste Klassifizierung von Donorstellen und Nicht-Donorstellen? Welche Schlussfolgerungen können Sie aus diesen Ergebnissen ziehen? D. h. welche statistischen Abhängigkeiten sind modellierungswürdig, welche Modelle neigen zum Übertraining, etc.?

Vertauschen Sie die Datensätze `seq_1` und `seq_3` mit den Datensätzen `seq_2` und `seq_4`, und wiederholen Sie die Analyse. Wie robust sind die vier 5×5 Matrizen?

Aufgabe 5.2 (4 Punkte)

Berechnen Sie – für jede der vier Datensätze `seq_1`, `seq_2`, `seq_3` und `seq_4` – $\tilde{Y}_1(i, j)$, $\tilde{Y}_2(i, j)$ und $\tilde{Y}_3(i, j)$ für alle Positionen $i, j = 1, 2, ..., L$, und stellen Sie die zwölf $L \times L$ Matrizen $\tilde{Y}_1$, $\tilde{Y}_2$ und $\tilde{Y}_3$ grafisch dar. Unter der Annahme der Nullhypothese, daß es keine statistischen Abhängigkeiten zwischen X_i und X_j gibt, sind $\tilde{Y}_1$, $\tilde{Y}_2$ und $\tilde{Y}_3$ χ^2-verteilt mit 3 Freiheitsgraden. Für welche Paare (i, j) finden Sie *statistisch signifikante* Abhängigkeiten, wenn Sie einen P-Wert kleiner als 0.01 als signifikant betrachten? Vergleichen Sie die zwölf $L \times L$ Matrizen $\tilde{Y}_1$, $\tilde{Y}_2$ und $\tilde{Y}_3$ mit den zwölf $L \times L$ Matrizen Y_1, Y_2 und Y_3 aus Aufgabe 3.2. Welche Gemeinsamkeiten und welche Unterschiede können Sie beobachten, und wie würden Sie diese interpretieren?

Aufgabe 5.3 (8 Punkte)

Benutzen Sie die Datensätze `seq_1` und `seq_3` zum Trainieren von MDD-Modellen. Setzen Sie im MDD-Algorithmus die Schwelle T so, dass die Wahrscheinlichkeit für

$S_i > T$ gleich 0.01 ist, und setzen Sie die Schwelle $U = 100$. Drucken Sie den Entscheidungsbaum sowie die PWMs an dessen Blättern aus. Plotten Sie die ROC-Kurve für die Datensätze `seq_2` und `seq_4`, und vergleichen Sie die ROC-Kurve des MDD-Modells mit den ROC-Kurven der PWM-Mischmodelle aus Aufgaben 4.4 und 5.1. Welche Modellkombination liefert die genaueste Klassifizierung von Donorstellen und Nicht-Donorstellen? Sei K die Anzahl der Blätter des Entscheidungsbaums für Datensatz `seq_1`, und sei L die Anzahl der Blätter des Entscheidungsbaums für Datensatz `seq_3`. Benutzen Sie nun `seq_1` zum Trainieren eines PWM-Mischmodells mit K Klassen, benutzen Sie `seq_3` zum Trainieren eines PWM-Mischmodells mit L Klassen, und plotten Sie die ROC-Kurve für `seq_2` und `seq_4`. Vergleichen Sie diese ROC-Kurve mit der ROC-Kurve für das MDD-Modell.

Vertauschen Sie die Datensätze `seq_1` und `seq_3` mit den Datensätzen `seq_2` und `seq_4`, und wiederholen Sie die Analyse. Wie robust sind die ROC-Kurven des MDD-Modells und des dazugehörigen PWM-Mischmodells?

Aufgabe 5.4 (8 Punkte)

Benutzen Sie die Datensätze `seq_1` und `seq_3` zum Trainieren von MDD-Modellen mit Schwellen T und U, und berechnen Sie für gegebenes T und U die Fläche unter der ROC-Kurve, $F(T, U)$, für die Datensätze `seq_2` und `seq_4`. Schreiben Sie für $T = 5, 10, 15, ..., 100$ und $U = 50, 100, 150, 200, 250$ die Werte $F(T, U)$ in eine 20×5 Matrix. Welche MDD-Modelle neigen zum Übertraining? Welches MDD-Modell liefert die genaueste Klassifizierung von Donorstellen und Nicht-Donorstellen? Drucken Sie – für das optimale MDD-Modell – die Entscheidungsbäume für `seq_1` und `seq_3`. Sei K die Anzahl der Blätter des Entscheidungsbaums für Datensatz `seq_1`, und sei L die Anzahl der Blätter des Entscheidungsbaums für Datensatz `seq_3`. Benutzen Sie nun `seq_1` zum Trainieren eines PWM-Mischmodells mit K Klassen, benutzen Sie `seq_3` zum Trainieren eines PWM-Mischmodells mit L Klassen, und plotten Sie die ROC-Kurve für `seq_2` und `seq_4`. Vergleichen Sie diese ROC-Kurve mit der ROC-Kurve für das MDD-Modell.

Vertauschen Sie die Datensätze `seq_1` und `seq_3` mit den Datensätzen `seq_2` und `seq_4`, und wiederholen Sie die Analyse. Wie robust ist die 10×5 Matrix? Wie robust ist der Entscheidungsbaum des optimalen MDD-Modells? Wie robust ist die ROC-Kurve der zum optimalen MDD-Modell gehörigigen PWM-Mischmodelle?

Abgabetermin: 20. Juni