Übungen Algorithmen der Bioinformatik II
Wintersemester 2006/07
Dr. Ivo Große, Dipl.-Bioinf. Jan Grau

Institut für Informatik
Universität Halle

Blatt 11

Aufgabe 11.1
Zeigen Sie, dass für beliebige Zufallsvariablen $X_1, X_2, \ldots, X_n$ mit $P(X_1, X_2, \ldots, X_{n-1}) > 0$ gilt:

$$P(X_1, X_2, \ldots, X_n) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2) \cdots P(X_n|X_1, X_2, \ldots, X_{n-1})\,.$$

Aufgabe 11.2 Zeigen Sie:

(a) wenn $\vec{p}$ ein Zufallsvektor und P eine stochastische Matrix ist, so ist auch $\vec{q} = \vec{p}P$ ein Zufallsvektor.

(b) wenn P und Q stochastische Matrizen sind, so sind auch PQ und QP stochastische Matrizen.

Bemerkungen:

(a) $\vec{x}$ ist ein D-dimensionaler Zufallsvektor, gdw. $x_d \geq 0 \;\; \forall d = 1, ..., D$ und $\sum_{d=1}^{D} x_d = 1$

(b) X ist eine stochastische $D{\times}D$ Matrix, gdw. $x_{ij} \geq 0 \;\; \forall i, j = 1, ..., D$ und $\sum_{j=1}^{D} x_{ij} = 1 \;\; \forall i = 1, ..., D$

Aufgabe 11.3 Zeigen Sie:

(a) wenn eine stochastische $D \times D$ Matrix X nur identische Zeilen $\vec{x} = (x_{i1}, ..., x_{iD}),\; i = 1, ..., D$ besitzt, so ist die stationäre Verteilung dieser Matrix gleich $\vec{x}$.

(b) jede Anfangsverteilung $\vec{y}$ mit positiven Elementen $y_i > 0 \;\forall i = 1, ..., D$ und $\sum_{d=1}^{D} y_d = 1$ konvergiert durch einmaliges Anwenden von X zur stationären Verteilung $\vec{x}$.

Aufgabe 11.4 Betrachten Sie folgendes Glücksspiel, welches in Ihrer Lieblingsspielhölle angeboten wird: Sie starten mit einem Kapital von $i, 0 < i < S$ Euro, der Spielhöllenbesitzer mit $S - i$ Euro. Nun wird eine Münze geworfen, die mit der Wahrscheinlichkeit p Kopf, mit $1 - p$ Zahl liefert. In ersterem Fall erhalten Sie vom Spielhöllenbesitzer einen Euro aus dessem Stapel, andernfalls müssen Sie einen rüberwandern lassen. Das Spiel endet, sobald einer von Ihnen keinen Euro mehr hat (und konsequenterweise der andere sich im Besitz von S Euro befindet).

Abgabe: 16.01.07

(a) Modellieren Sie dieses Spiel durch ein geeignetes Markov Modell.

(b) Gibt es absorbierende Zustände? Wenn ja, welche?

(c) Gibt es Garten-Eden Zustände? Wenn ja, welche?

(d) Betrachten Sie nun den Fall $S = 5, p = 0.2, i = 2$. Wie groß ist die Wahrscheinlichkeit, dass Sie das Spiel nach $n = 3, n = 5$ bzw. $n = 10$ Münzwürfen gewinnen?

Aufgabe 11.5 Auf der Internetseite zur Vorlesung finden Sie in der Datei `arabido.txt` einen Datensatz mit Macroarray Expressionsdaten. Von *Arabidopsis thaliana*. Die Daten wurden bereits normiert und logarithmiert, und die daraus resultierenden Werte nennen wir im folgenden kurz Log-Intensitäten. Für jedes Gen i erhalten wir 2 Log-Intensitäten, einmal bei Tag (x_i), einmal bei Nacht (y_i).

(a) Stellen Sie das Histogramm der x-Werte und das Histogramm der y-Werte grafisch dar und erstellen Sie einen Scatterplot von x versus y.

Wir wollen im folgenden annehmen, dass die Daten durch Ziehen aus einer bivariaten Gaussverteilung (mit unbekannten Parametern μ_1, μ_2, σ_{11}, σ_{22}, ρ) generiert wurden. Statt der Parameter σ_{11} und σ_{22} schreiben wir im folgenden C_{11} und C_{22}, und statt des Parameters ρ schreiben wir im folgenden $C_{12} = \rho\sqrt{\sigma_{11}\sigma_{22}}$.

(b) Leiten Sie die erzeugende Funktion, den Erwartungswertvektor sowie die Kovarianzmatrix einer bivariaten Gaussverteilung her.

(c) Leiten Sie den Maximum Likelihood Schätzer von μ_1, μ_2, C_{11}, C_{22} und C_{12} her und berechnen sie die entsprechenden Schätzwerte für den gegebenen Datensatz.

(d) Berechnen Sie die beiden Randdichten, stellen Sie sie grafisch dar, und vergleichen Sie sie mit den beiden Histogrammen aus Aufgabe (a). Berechnen Sie weiterhin die Erwartungswerte und Varianzen der beiden Randdichten, $E(X)$ und $E(Y)$ sowie $Var(X)$ und $Var(Y)$.

(e) Berechnen Sie die beiden bedingten Dichten $p(x|y)$ und $p(y|x)$ sowie die dazugehörigen bedingten Erwartungswerte und Varianzen, $E(X|y)$ und $E(Y|x)$ sowie $Var(X|y)$ und $Var(Y|x)$. Stellen Sie $E(X|y)$ und $Var(X|y)$ als Funktion von y sowie $E(Y|x)$ und $Var(Y|x)$ als Funktion von x grafisch dar. Vergleichen Sie die bedingten Momente mit den (nicht bedingten) Momenten aus Aufgabe (d). Was lässt sich aus diesem Vergleich über die statistische Abhängigkeit zwischen X und Y schlussfolgern?

(f) Wie groß ist der Korrelationskoeffizient zwischen X und Y, und was bedeutet dies im Kontext der Genexpression von *Arabidopsis thaliana* bei Tag und bei Nacht? Welche Gene würden Sie als differentiell exprimiert bezeichnen und Ihrem Biologen, der Ihnen die Daten zur Analyse bereitstellte, zur näheren Untersuchung empfehlen?

Abgabe: 16.01.07