Statistische Methoden der Datenanalyse

Markus Schumacher

Übung IX

Matthew Beckingham und Markus Warsinsky

1.7.2009

Anwesenheitsaufgaben

Aufgabe 39 *Fisherdiskriminante und Likelihoodverhältnis*

Betrachten Sie die Fisherdiskriminante für den Fall zweier Wahrscheinlichkeitsdichtefunktionen $f(\vec{x}|H_0)$ und $f(\vec{x}|H_1)$, die jeweils multidimensionale Gaussverteilungen mit identischen Kovarianzmatrizen $V_0 = V_1 = V$ sein sollen. Die Wahrscheinlichkeitsdichtefunktion ist also gegeben als:

$$f(\vec{x}|H_k) = \frac{1}{(2\pi)^{n/2}|V|^{1/2}} \exp\left[-\frac{1}{2}(\vec{x}-\vec{\mu}_k)^T V^{-1}(\vec{x}-\vec{\mu}_k)\right] \qquad k \in \{0,1\} \tag{1}$$

(i) Zeigen Sie, dass das Likelihoodverhältnis gegeben ist durch

$$r = \frac{f(\vec{x}|H_0)}{f(\vec{x}|H_1)} = \exp(t), \tag{2}$$

wobei t die Fischerdiskrimante

$$t(\vec{x}) = a_0 + (\vec{\mu}_0 - \vec{\mu}_1)^T V^{-1}\vec{x} \tag{3}$$

mit einem beliebigem Schwellenwert a_0 ist. Dementsprechend ist eine Optimierung des Likelihoodverhältnisses äquivalent zu einer Optimierung der Fisherdiskrimante.

(ii) Benutzen Sie das Bayes-Theorem mit Aprioriwahrscheinlichkeiten π_0 und π_1 für H_0 bzw. H_1, um zu zeigen, dass die bedingte Wahrscheinlichkeit für H_0 bei gegebenen Daten $\vec{x}$ gegeben ist durch

$$P(H_0|\vec{x}) = \frac{1}{1+e^{-t}} = s(t), \tag{4}$$

wobei $s(t)$ die Logistische Funktion ist. Betrachten Sie dazu eine Redefinition des Schwellenwertes von der Form $a_0' = a_0 + \ln\frac{\pi_0}{\pi_1}$.

Aufgabe 40 *Zählexperiment für eine Signal- und Untergrundmessung*

Betrachtet wird ein Experiment, dessen Ziel die Entdeckung eines neuen Teilchens oberhalb eines von momentanen Theorien vorhergesagten Untergrundes ist. Dabei könnte es sich beispielsweise um das Higgs-Boson oder auch um supersymmetrische Teilchen handeln, wobei die Untergrundvorhersage durch das Standardmodell der Teilchenphysik erfolgt.

Die zu betrachtenden Hypothesen sind also die Nullhypothese H_0, dass nur Ereignisse aus Untergrundprozessen gemessen wurden, sowie die Alternativhypothese H_1, dass sowohl Signal- als auch Untergrundereignisse beobachtet wurden.

Im Experiment wurde eine Gesamtanzahl x von Ereignissen aufgezeichnet. Weiterhin wurde eine andere, signalfreie kinematische Region definiert, aus der man die Normierung des Untergrundes bestimmen kann. In dieser Region wurden y Ereignisse gefunden. Das Verhältnis der Untergrundereignisse in der signalfreien Region zu denjenigen in der Signalregion sei τ. Die mittlere Anzahl der Untergrundereignisse in der Signalregion sei b und die mittlere Anzahl der Signalereignisse im Falle von Hypothese H_1 sei s.

(i) Stellen Sie die Likelihoodfunktionen für die Hypothesen H_0 und H_1 auf. Nehmen Sie dabei an, dass die Gesamtanzahlen von Ereignissen in Signal- und Kontrollregion jeweils Poissonverteilt sind.

(ii) Betrachten Sie jetzt die Schätzer für s und b unter der Hypothese H_1 ($\hat{s}$ bzw. $\hat{b}$) sowie den Schätzer für b unter der Nullhypothese, $\hat{\hat{b}}$.

a) Stellen Sie die Profile Likelihood λ auf.

b) Bestimmen Sie Ausdrücke für $\hat{s}$, $\hat{b}$ und $\hat{\hat{b}}$. Betrachten Sie dazu

$$\left.\frac{\partial L(H_0)}{\partial b}\right|_{\hat{\hat{b}}}, \tag{5}$$

sowie die die beiden gleichzeitigen Einschränkungen

$$\left.\frac{\partial L(H_1)}{\partial s}\right|_{\hat{s}} \quad \text{und} \quad \left.\frac{\partial L(H_1)}{\partial b}\right|_{\hat{b}}. \tag{6}$$

c) Berechnen Sie $q = -2\ln\lambda$.

Hausaufgaben

Aufgabe 41 *Zählexperiment für eine Signal- und Untergrundmessung 2* **7 Punkte**

In dieser Aufgabe sollen Sie die Ergebnisse von Aufgabe 40 benutzen, um die erwartete Sensitivität des ATLAS-Experimentes auf eine Entdeckung des Higgs-Bosons im Zerfallskanal $H \to \gamma\gamma$ zu ermitteln.

Monte-Carlo-Studien von Proton-Proton-Kollisionen im ATLAS-Detektor haben gezeigt, dass der Wirkungsquerschnitt für $pp \to H+X \to \gamma\gamma+X$-Ereignisse, die die Ereignisselektion passieren, gegeben ist durch $\sigma_S = 25{,}4\,\text{fb}$. Der Wirkungsquerschnitt für Untergrundereignisse, die dieselbe Ereignisselektion passieren, beträgt $\sigma_B = 947\,\text{fb}$. In einer weiteren Analyse kann ein reiner Untergrunddatensatz mit einem Wirkungsquerschnitt von $\sigma_T = 10300\,\text{fb}$ ausgewählt werden.

(i) Benutzen Sie die Relation

$$N = \mathcal{L}\sigma, \tag{7}$$

um die Anzahlen von Signal- (x), Untergrundereignissen (y) sowie die Anzahl von Ereignissen in der Seitenbandregion (τb) für eine integrierte Luminosität von $\mathcal{L} = 10\,\text{fb}^{-1}$ auszurechnen.

(ii) Berechnen Sie die Schätzer für die Anzahl der Signalereignisse $\hat{s}$, der Untergrundereignisse $\hat{b}$ unter der Hypothese von Signal plus Untergrund , sowie den Schätzer $\hat{\hat{b}}$ auf die Anzahl der Untergrundereignisse in der Nur-Untergrund Hypothese.

(iii) Berechnen Sie die Größe

$$q = -2\ln\lambda, \tag{8}$$

wobei λ die in Aufgabe 40 berechnete Profile Likelihood ist.

(iv) Berechnen Sie daraus die Signifikanz des vorhergesagten Signals.

(v) Nehmen Sie nun an, dass die Untergrundrate in der Signalregion mit einer relativen Genauigkeit von $\Delta b/b = 5\%$ abgeschätzt werden kann. Wenn man annimmt, dass es sich dabei um einen Poissonfehler handelt, kann man eine effektive Seitenbandregion konstruieren mit Ereignisanzahl $y' = \tau' b$. Dazu setzt man die relative statistische Unsicherheit in der hypothetischen Seitenbandregion (Poissonfehler) gleich der relativen Genauigkeit der Untergrundvorhersage:

$$\frac{\sqrt{\tau' b}}{\tau' b} = \frac{\Delta b}{b} \Leftrightarrow \tau' = \frac{b}{(\Delta b)^2} \tag{9}$$

Berechnen Sie die Werte für τ' und y' für die effektive Seitenbandregion.

(vi) Benutzen Sie die Werte für τ' und y' sowie die ursprüngliche Anzahl von Ereignissen in der Signalregion x, um den neuen Wert für q und daher der Signifikanz für eine Messung mit einem Fehler von 5% auf die Untergrundvorhersage zu bekommen.

Aufgabe 42 *Maximale Separation der Fisherdiskriminante* **6 Punkte**

Ein Maß für die Separation zweier Hypothesen H_0 und H_1 mittels einer Fisherdiskriminanten ist gegeben durch

$$J(\vec{a}) = \frac{\vec{a}^T B \vec{a}}{\vec{a}^T W \vec{a}}, \tag{10}$$

woei die Fisherdiskrimante definiert ist als

$$t(\vec{x}) = \vec{a}^T \vec{x}. \tag{11}$$

Die Separation zwischen H_0 und H_1 wird repräsentiert durch

$$B_{ij} = (\mu_0 - \mu_1)_i (\mu_0 - \mu_1)_j, \tag{12}$$

und die Summe der Kovarianzen zwischen den zwei Klassen ist gegeben durch

$$W_{ij} = (V_0 + V_1)_{ij}. \tag{13}$$

(i) Bilden Sie die Ableitung $\partial J(\vec{a})/\partial \vec{a}$ von $J(\vec{a})$ nach $\vec{a}$ und zeigen Sie, dass das Maximum von $J(\vec{a})$ durch die Eigenwertgleichungen

$$W^{-1} B \vec{a} = \lambda \vec{a} \tag{14}$$

gegeben ist.

(ii) Zeigen Sie, dass für einen beliebigen Vektor $\vec{a}$ der Vektor $B\vec{a}$ parallel zu $(\vec{\mu}_0 - \vec{\mu}_1)$ ist.

(iii) Zeigen Sie damit, dass

$$\vec{y} \propto W^{-1}(\vec{\mu}_0 - \vec{\mu}_1) \tag{15}$$

eine Lösung der Eigenwertgleichungen aus (i) ist und daher $J(\vec{a})$ maximiert.

Aufgabe 43 *Likelihoodverhältnis für die Ξ^0 Lebensdauer* **7 Punkte**

Gegeben sei ein Datensatz von n gemessenen Zerfallszeiten $t_1, t_2, \ldots, t_n$ eines Ξ^0-Teilchens. Betrachten Sie einen Test der Nullhypothese $H_0 : \tau = 1$, wobei τ die wahre Lebensdauer der Ξ^0-Teilchens ist, gegenüber der Alternativhypothese $H_1 : \tau \neq 1$.

(i) Wie lauten die zwei Likelihoodfunktionen $\alpha(\vec{x}, \vec{\lambda}(\omega))$ für H_0 und für den vollen Parameterraum $\alpha(\vec{x}, \vec{\lambda}(\Omega))$?

(ii) Zeigen Sie, dass das Likelihoodverhältnis gegeben ist durch

$$T = \frac{\alpha(\vec{x}, \vec{\lambda}(\omega))}{\alpha(\vec{x}, \vec{\lambda}(\Omega))} = \bar{t}^n \exp\left(-n(\bar{t} - 1)\right), \tag{16}$$

wobei $\bar{t} = \frac{1}{n} \sum_{i=1}^{n} t_i$ der arithmetische Mittelwert der gemessenen Zerfallszeiten ist.

(iii) Für große n ist $\bar{t}$ nach $N(1, \frac{1}{n})$ verteilt. Berechnen Sie damit den kritischen Wert von $\bar{t}$ für eine Signifikanz $\alpha = 0.05$.

(iv) Berechnen Sie den kritischen Wert T_{cr} des Likelihoodverhältnisses und zeigen Sie, dass $T_{\mathrm{cr}} \to 0$ für $n \to \infty$.