Albert-Ludwigs-Universität Freiburg Wintersemester 2009-2010

Statistische Methoden der Datenanalyse

Markus Schumacher

Übung XIII

Matthew Beckingham und Markus Warsinsky

4.2.2010

Computerübung

Aufgabe 49 *Fisher-Diskriminante*

Im folgenden soll die sogenannte Fisher-Diskriminante zur Trennung zwischen zwei Ereignisklassen benutzt werden. Dabei werden i.A. n trennende Variablen $x_1, \ldots x_n$ benutzt. Die Fisher-Diskriminante ist dann gegeben durch

$$t = \sum_{i=1}^{n} a_i x_i, \tag{1}$$

wobei $\bar{x}_i^{(1,2)}$ die Mittelwerte der Messwerte der Klassen 1 bzw. 2 sind. Vorige Woche wurde gezeigt dass der maximale Wert der Trenngröße

$$J(\vec{a}) = \frac{(\tau_0 - \tau_1)^2}{\Sigma_0^2 + \Sigma_1^2} \tag{2}$$

gegeben ist für

$$a_i \propto W_{ij}^{-1}(\vec{\mu_0} - \vec{\mu_1})_j, \tag{3}$$

wobei W die Summe der Kovarianzmatrizen der zwei Ereignisklassen (V_0 und V_1) ist:

$$W_{ij} = (V_0 + V_1)_{ij} \tag{4}$$

Im folgenden wollen wir die folgende Situation betrachten: Eine Signalklasse C_S und eine Untergrundklasse C_U sollen durch eine Fisher-Diskrimante in den Variablen (x,y) optimal getrennt werden. Die Signalereignisse sollen gemäß einer zweidimensionalen Gaußverteilung mit Mittelwerten ($\bar{x}$,$\bar{y}$) = (0. 4,0. 4) und $\sigma_S = 0.15$, und die Untergrundereignisse ebenfalls nach einer zweidimensionalen Gaußverteilung mit $\sigma_U = 0.15$ und Mittelwerten ($\bar{x}$,$\bar{y}$) = (0. 7,0. 7) verteilt sein.

(i) Verwenden Sie das Makro `fisher_generateTree_i.C` um einen TTree mit 10000 Paare (x,y) zu generieren und in einem Datei zu speichern. Ergänzen Sie das Makro und lassen Sie des Makro zwei Mal laufen mit unterschiedliche Werten für ($\bar{x}$,$\bar{y}$) um die Dateien "signal.root" und "untergrund.root" zu erzeugen.

(ii) Mit dem nächsten Makro, `fisher_determineCoeffs_i.C`, sollen Sie die Werte a_0 und a_1, die $J(\bar{a})$ maximieren, bestimmt werden mittels Gleichung 3. Dafür müssen die Durchschnittswerte $\bar{x}$ und $\bar{y}$ geschätzt werden. Zusätzlich müssen die Kovarianzmatrizen der zwei Ereignisklassen aus den jeweils 10000 (x,y)-Paare aus i) geschätzt werden mittels

$$\hat{V}_{xy} = \frac{1}{n-1} \sum_{i=i}^{n} (x_i - \bar{x})(y_i - \bar{y}). \tag{5}$$

Die Matrix W in Gleichung 4 wird mittels der Klasse "TMatrixD" definiert und invertiert.

(iii) Im letzten Makro, `fisher_plotFisherT_i.C`, werden die in ii) bestimmten Werte für a_0 und a_1 verwendet um die Fischer Diskriminante (t) für jedes der 10000 Signal- und Untergrundsereignisse zu bestimmen. Für ein Schnittwert t_{cut} ist die Effizienz (ϵ) und Reinheit (π) der selektierten Ereignismenge ist definiert als

$$\epsilon = \frac{N_S(t > t_{\text{cut}})}{N_S} \qquad \pi = \frac{N_S(t > t_{\text{cut}})}{N_S(t > t_{\text{cut}}) + N_U(t > t_{\text{cut}})}, \tag{6}$$

wobei $N_S = 10000$ und $N_{S/U}(t > t_{\text{cut}})$ die Anzahl der Signalereignisse(/Untergrundereignisse) die einen Wert t größer als t_{cut} haben, ist. Die Effizienz und Reinheit der selektierten Ereignismenge sollen berechnet werden für 100 verschiedene Werte von t_{cut} zwischen -5 und 15 und in einem TGraph aufgetragen werden. Zum Vergleich sollen Sie auch eine Ereignismenge definieren durch einen einfachen, rechteckigen Schnitt $x > c$ und $y > c$, wobei c eine Konstante ist. Die Effizienz und Reinheit der dadurch selektieren Ereignismenge werde für 100 verschiedene Werte für c zwischen -1 und 3 in einen TGraph eingetragen. Ist bei der Untersuchung des Signals eine durch einen rechteckigen Schnitt in (x,y) oder eine mittels Fisher-Diskriminante ausgewählte Ereignismenge vorzuziehen?

(iv) Wiederholen Sie die Schritte i)-iii) mit $\sigma_S = 0.3$ und $\sigma_B = 0.15$, und danach mit $\sigma_B = 0.3$ und $\sigma_S = 0.15$. Wie ändern sich $\pi(\epsilon)$ graphisch?

Aufgabe 50 *P-Wert*

Zwei Zufallszahlen x und y werden aus einer WDF gezogen. Die Nullhypotese ist die WDF einer Normalverteilung (G(0,1)). Die Differenz $|x - y|$ ist gleich 3.7. Was ist der P-Wert des Ergebnis? Hinweise: Simulien Sie viele Paare x, y aus einer Normalverteilung mittels TRandom3 und in einen Histogramme eintragen. Das Integral eines Histogrammes von Bin nr n bis zum letzten Bin ist hist−>Integral(n, hist−>GetNbinsX()). Um die Bin-Nummer zu finden wo 3.7 liegt, verwenden Sie hist−>GetXaxis()−>FindBin(3.7).