Statistische Methoden der Datenanalyse

Wintersemester 2012/2013

Albert-Ludwigs-Universität Freiburg

Prof. Markus Schumacher, Dr. Stan Lai

Physikalisches Institut Westbau 2 OG

UNI FREIBURG

E-Mail: Markus.Schumacher@physik.uni-freiburg.de

stan.lai@cern.ch

Kapitel IX

Ereignisklassifizierung

Ereignisklassifizierung: Zielsetzung

Ziel: Trennung von mehrereren (hier meist 2) Klassen von Ereignissen an Hand von n Observablen ($x_1 \ldots . x_n$) die für jedes Ereignis gemessen werden

Fragen: wie erhält man optimale Trennung zwischen Klassen?
wie klassifiziert man unbekanntes Ereignis?

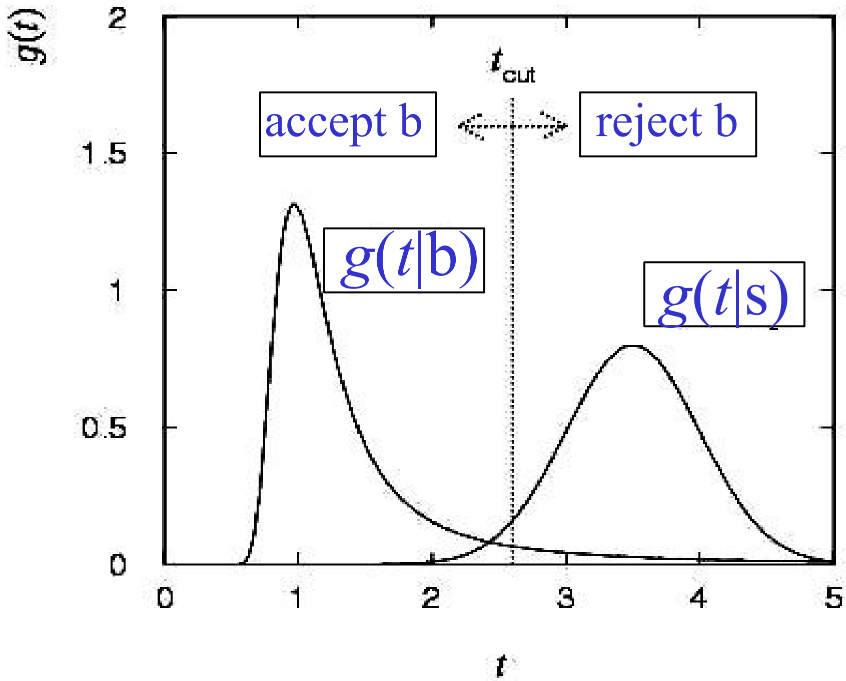
H_0: Ereignis gehört zur Untergundklasse b, WDF g(t|b)
H_1: Ereignis gehört zur Signalklasse s, WDF g(t|s)

Untergrund-nachweiswahrscheinlichkeit:

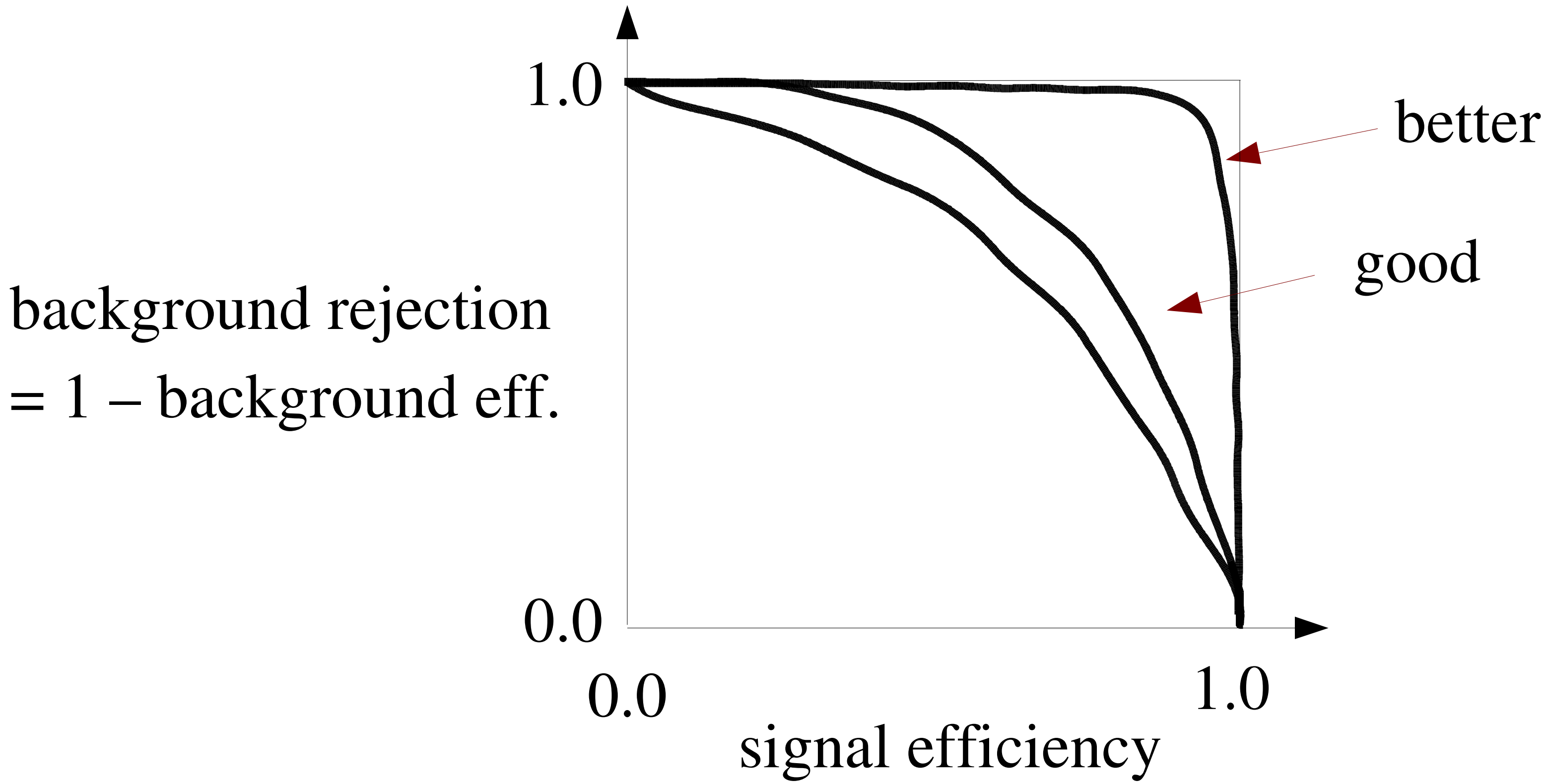
$$\varepsilon_{\mathrm{b}} = \int_{t_{\mathrm{cut}}}^{\infty} g(t|\mathrm{b})\, dt = \alpha$$

Signal-Nachweiswahrscheinlichkeit:

$$\varepsilon_{\mathrm{s}} = \int_{t_{\mathrm{cut}}}^{\infty} g(t|\mathrm{s})\, dt = 1 - \beta$$



Ereignisklassifizierung: Zielsetzung



"Receiver Operation Characteristics" ROC-Kurve

Ziel: maximiere Signaleffizienz und Untergrundunterdrückung ($1-\varepsilon_b$)

Reinheit der Ereignisselektion

Annahme: wir haben nur eine Untergrundklasse;
Anteil an Signal und Untergrundereignissen sind
π_s and π_b (A-Priori Wahrscheinlichkeiten).

Annahme: wir selektieren Signalereignisse mit der Bedingung $t > t_{cut}$.
Was is die 'Reinheit' unseres selektierten Samples?

Reinheit bedeutet hier die Wahrscheinlichkeit, dass ein akzeptiertes/ selektiertes Ereignis der Signalklasse entstammt.

Unter der Verwendung des Theorem von Bayes finden wir:

$$P(\mathrm{s}|t > t_{\mathrm{cut}}) = \frac{P(t > t_{\mathrm{cut}}|\mathrm{s})\pi_{\mathrm{s}}}{P(t > t_{\mathrm{cut}}|\mathrm{s})\pi_{\mathrm{s}} + P(t > t_{\mathrm{cut}}|\mathrm{b})\pi_{\mathrm{b}}}$$
$$= \frac{\varepsilon_{\mathrm{s}}\pi_{\mathrm{s}}}{\varepsilon_{\mathrm{s}}\pi_{\mathrm{s}} + \varepsilon_{\mathrm{b}}\pi_{\mathrm{b}}}$$

Die Reinheit hängt sowohl von den A-Priori Wahrscheinlichkeiten als auch von den Nachweiswahrscheinlichkeiten für Signal und Untergrund ab.

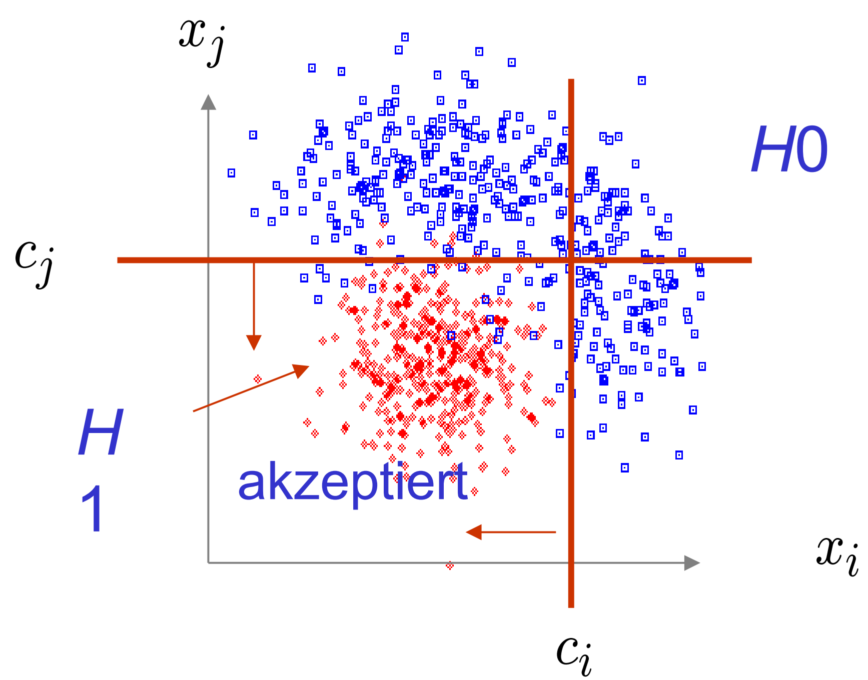
Ereignisklassifizierung

Jedes Ereignis ist ein Punkt im n-dimensionalen $\vec{x}$ Raum.
Wie sollten wir die Entscheidungsgrenze im n-dimensionalen Raum der Observablen wählen umd Ereignisse zu akzeptieren/verwerfen vom Ereignistyp H_0 oder H_1?

Eine Möglichkeit:
Selektion der Ereignisse mit Hilfe von konsekutiven Schnitten

$$x_i < c_i$$

$$x_j < c_j$$



Akzeptanzregion ist ein n-dimensionaler Quader

Optimale Trennkraft durch Neyman-Person-Lemma

für gegebene Signaleffizienz εs maximale Untegrudnunterdrückung $1-\varepsilon_b$

Fragen: welche Teststatistik t? welche Wahl der kritischen Region S_{krit}?

Antwort durch Neyman-Pearson-Lemma:
Ein Test der einfachen Nullhypothesen H_0 bzgl der einfachen Alternativhypothese H_1 ist ein bester Test, wenn die kritische Region S_{krit} im Stichprobenraum E so gewählt wird, dass gilt

$$\frac{P(\mathbf{x}|H_1)}{P(\mathbf{x}|H_0)} > c$$

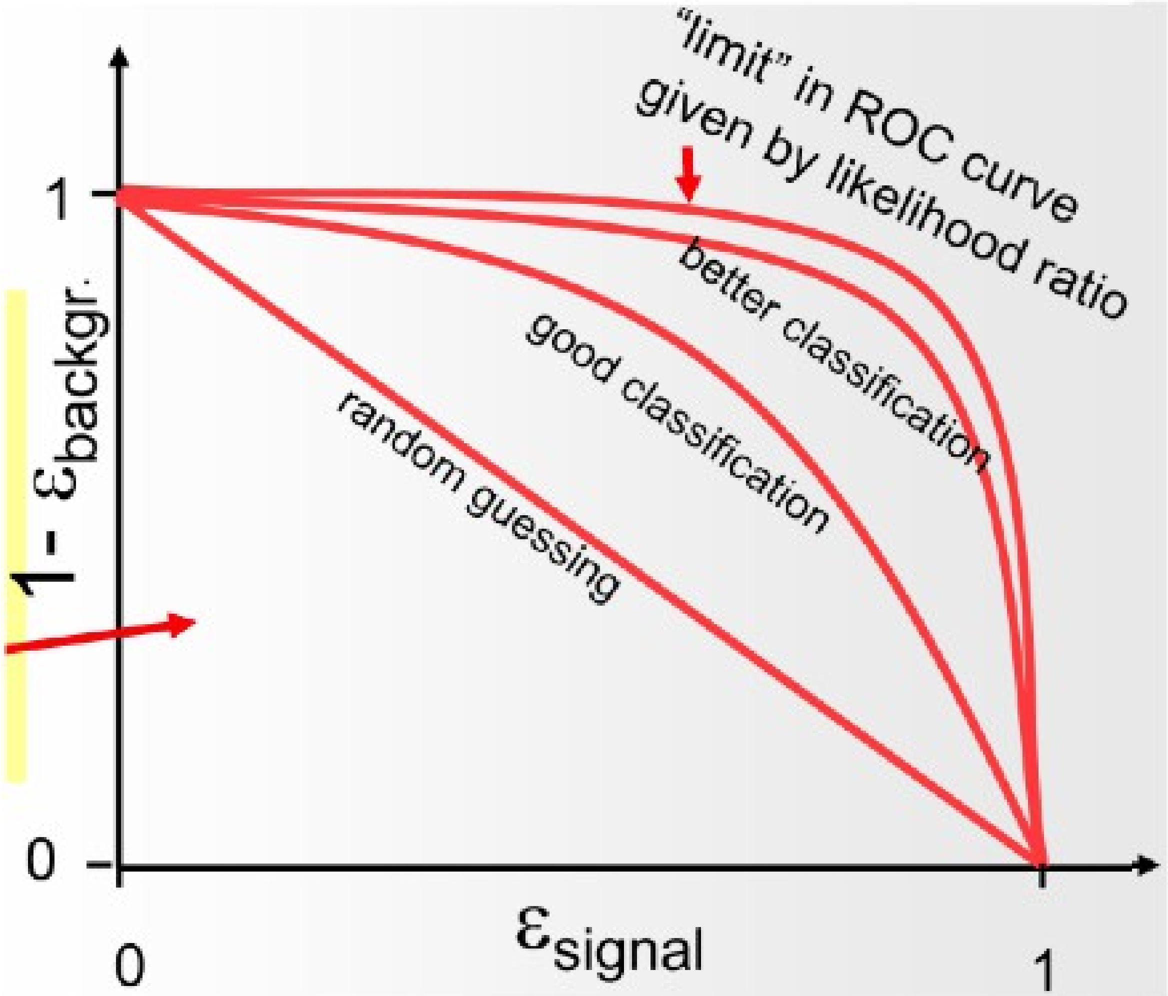
(≤ c außerhalb kritischer Region)
c ist Konstante die von Signifkanzniveau abhängt.

äquivalente Aussage: die optimale Teststatistik ist

$$\boxed{t(\mathbf{x}) = \frac{P(\mathbf{x}|H_1)}{P(\mathbf{x}|H_0)}}$$

Achtung: oft wird auch der Kehrwert des Likelihoodverhältnisses verwendet.
Jede monotone Funktion von t(x) ist ebenso optimal wie t selbst z.B. t/(1+t)

Optimale Trennkraft durch Neyman-Person-Lemma

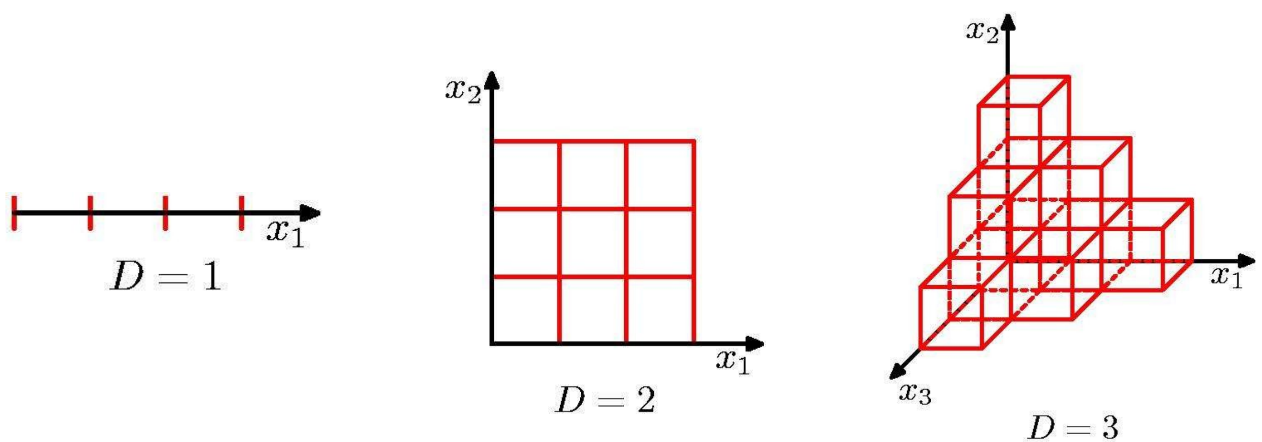


Receiver Operation Characteristics ROC-Kurve

Warum Neyman-Pearson-Lemma nicht immer hilft?

Das Problem besteht häufig darin, dass wir kein analytischen Ausrücke für die n-dimenionalenn WDFs $P(\boldsymbol{x}|H_0)$, $P(\boldsymbol{x}|H_1)$ haben.

Im Prinzip Erstellung aus simulierten Trainigsereignissen für Signal und Untergrund möglich. Fülle simulierte Messergebnisse in ein n-dimensionales Histogramm (für n Observablen). Verwende *M* Bins für jede der *n* Dimensionen. Sehr viele Ereignisse für kleine Fehler in jedem Bin nötig. In der Praxis nicht genug simulierte Ereignisse verfügbar. “Fluch der Dimensionalität FdD”



M^n Zellen

“curse of dimensionality” (Bellman, 1961)