Bayessche Statistik

Christian Meisel

19.07.2007

Gliederung

Wahrscheinlichkeitsbegriff

- In Bayesscher Statistik wird Wahrscheinlichkeit p im Sinne von unvollständigem Wissen über ein Ereignis verwendet.
- Bei gleichem Vorwissen über unterschiedliche Ereignisse werden diesen dementsprechend auch gleiche Wahrscheinlichkeiten zugeordnet.

Wahrscheinlichkeitsbegriff

Im Gegensatz dazu wird Frequenz f abgegrenzt, welche die Häufigkeit des Auftretens eines Ereignisses bei repetitiven Zufallsexperimenten beschreibt.

Wahrscheinlichkeitsbegriff

Beispiel Münzwurf: beide Ergebnisse haben p = 1/2 und f = 1/2 (Anzahl der Versuche gegen unendlich)
p und f dürfen jedoch nicht verwechselt oder gleich gesetzt werden. Statt dessen soll die Wahrscheinlichkeit p(f)df, dass die Frequenz im Intervall df liegt, berechnet werden.

Bedingte Wahrscheinlichkeiten

- Bayessche Statistik arbeitet mit bedingten Wahrscheinlichkeiten
- A ist wahr unter der Bedingung B: A|B
- deren Wahrscheinlichkeit p(A|B)

Bayessches Theorem

- aus klassischer Statistik gilt:

$$p(A|BC)p(B|C) = \frac{p(A \bigcap B \bigcap C)}{p(B \bigcap C)} * \frac{p(B \bigcap C)}{p(C)} = \frac{p(A \bigcap B \bigcap C)}{p(C)}$$

- somit ist p(AB|C) = p(A|BC)p(B|C)
- da AB und BA den gleichen Zustand beschreiben, lässt sich A und B auf rechter Seite austauschen, also
- $p(A|BC)p(B|C) = p(B|AC)p(A|C)$

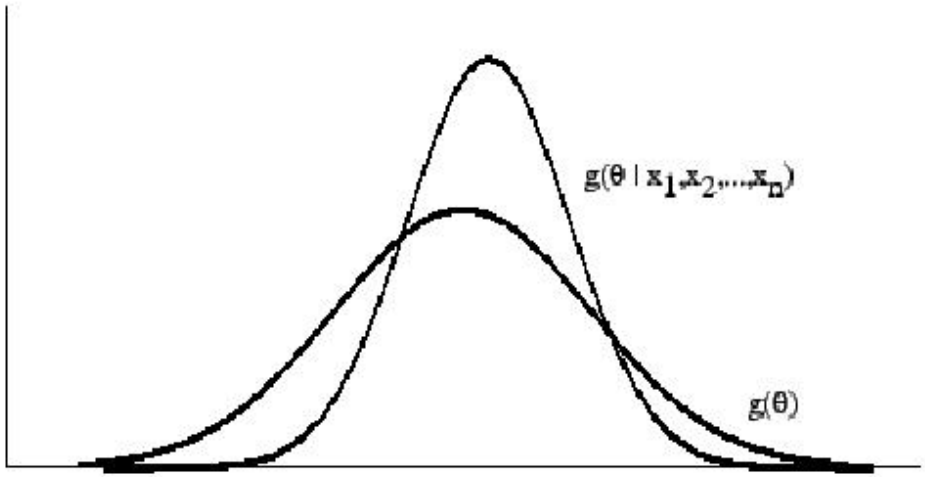
Durch umstellen folgt das Bayessche Theorem:

$$p(A|BC) = p(A|C)\frac{p(B|AC)}{p(B|C)}$$

Bayessches Theorem

$$p(A|BC) = p(A|C)\frac{p(B|AC)}{p(B|C)}$$

- p(A|C) ist a priori-Wahrscheinlichkeit, abhängig nur von A
- p(A|BC) ist a posteriori-Wahrscheinlichkeit, beeinflusst durch die zusätzliche Information B
- p(B|AC) wird als Likelihood bezeichnet



A-priori- und A-posteriori-Dichte

Bayessches Theorem

- Bayessches Theorem als Lernprozess
- erstem Wissen, welches nur C beinhaltet, werden die veränderten Bedingungen unter B hinzugefuegt
- neue Informationen $B_1, B_2, \ldots$ können eingeschlossen werden.

Umgekehrt wichtig zu fragen, was wussten wir über A bevor B gesehen zu haben.

Anwendungsbeispiel zum Bayesschen Theorem

Berühmtes Beispiel von Laplace

- A: Masse von Saturn M_S liegt in bestimmtem Intervall
- B: Daten von Observatorien ueber gegenseitige Störungen von Jupiter und Saturn
- C: 'common sense', dass M_S weder so klein sein kann, dass er seine Ringe verliert, noch so gross, dass er das Sonnensystem zerstoert

Daten aus dem 18. JH liessen mit BT eine Schätzung von M_S und Vorhersage auf 1 Prozent Genauigkeit zu, bis heute Wert nur um 0,63 Prozent korrigiert.

Anwendungsbeispiel zum Bayesschen Theorem

- zwei Freunde in Pub, einer muss zahlen
- wie gross ist Wahrscheinlichkeit des Betrugs?
- $p(C|W_n) = p(C) * \frac{p(W_n|C)}{p(W_n)}$
- $p(W_n) = p(W_n|C) * p(C) + p(W_n|H) * p(H)$
- ferner seien $p(W_n|C) = 1$; $p(W_n|H) = 2^{-n}$
- unter freundschaftlicher Annahme p(C) = 0,05:

n	$P(C \mid W_n)$ (%)	$P(H \mid W_n)$ (%)
0	5.0	95.0
1	9.5	90.5
2	17.4	82.6
3	29.4	70.6
4	45.7	54.3
5	62.7	37.3
6	77.1	22.9
...	...	...

Problematik der objektiven a priori-Verteilung

Subjektivität

- verschiedene Physiker
- verschiedene Ansichten
- verschiedene Resultate

Objektivierbarkeit durch

- minimale Information
- maximale Ignoranz
- Skalenargumente

Minimale Information

- wichtig ist, nicht zuviel 'Wissen' in a priori Verteilung zu integrieren
- von allen Verteilungen ist die zu nehmen, in welcher das a priori Wissen minimal ist
- anders formuliert, muss das Mass an Unsicherheit maximal sein

Shannon Entropie

$$S_I = -\sum p_i \lg p_i$$

- ist in Mass für Unsicherheit
- $S_I = \textit{maximal}$ wenn alle $p(x_i)$ identisch
- $S_I = \textit{minimal}$ wenn $p(x_0) = 1$ oder 0

Maximale Entropie

- Ergebnisraum X bestehend aus Elementarereignissen $(x_1 ... x_N)$
- aufgrund von physikalischen Restriktionen nicht alle in gleicher Weise realisierbar
- daher bestehen zusätzliche Einschränkungen auf $p_i = p(x_i)$
- Problem: Suche p_i auf erlaubten Konfigurationen, so dass möglichst viele Freiheitsgrade erhalten bleiben

Maximale Entropie

- **Problem**: $S(p) \doteq max$ unter NB $\sum f_k(x_i)p(x_i) = F_k$
- **Lösung** mit `Lagrangesche Multiplikatoren`:

$$L(p;\lambda) = -\sum p_i \lg p_i + \sum \lambda_k (\sum f_k(x_i) - F_k) = max$$

- $\delta_{p_j} = -\lg p(x_j) - 1 + \sum \lambda_k f_k(x_j) = 0$
- $p_j = \frac{1}{Z(\lambda)} \exp(\lambda_1 f_1(x_i) + ... + \lambda_m f_m(x_i))$

  mit $Z(\lambda_1 ... \lambda_m) = \sum \exp(\lambda_1 f_1(x_i) + ... + \lambda_m f_m(x_i))$
- $F_k = \delta_{\lambda_k} \lg Z(\lambda_1 ... \lambda_m)$

Transinformation

- seien x,y Parameter mit gemeinsamer Dichte p(x,y)
- $S(p) = -\sum p(x,y) \lg p(x,y)$
- Relative Entropie/Transinformation

$$-\sum p(x,y) \lg \frac{p(x,y)}{p(x)p(y)} =$$

$$= -\sum p(x,y) \lg p(x,y) + \sum p(x,y) \lg p(x) + \sum p(x,y) \lg p(y)$$
$$= S(x,y) - S(x) - S(y)$$
$$= I(x,y) = \text{Transinformation}$$

- I(x,y) = 0 wenn x,y unabhängig

Reference Prior

- I(x,y) entspricht Verminderung (im Mittel) der Unsicherheit eines Parameters durch Beobachten des anderen
- 'Lerneffekt'
- Reference Prior: wähle jene a priori Verteilung, welche Lerneffekt I maximiert

Skaleninvarianz

- Zuordnung der a priori Verteilung muss skaleninvariant sein
- gegeben sei p(x|C)dx = p(y|C)dy
- unter Ortstransformation
  - $y = x + b \Rightarrow p(x|C)dx = p(x + b|C)dx$
  - p(x|C) = const., ergibt also gleichen Prior
- unter Skalentransformation
  - $y = \alpha x, dy = \alpha dx \Rightarrow p(x|C)dx = \alpha p(\alpha x|C)dx$
  - p(x|C) = 1/x ist Jeffreys Prior

Skaleninvarianz

- beispielhaft zwei Beobachter die mit verschiedenen Uhren Experiment beobachten
- haben gleiches Vorwissen und sollten deshalb auch gleichen Prior verwenden
- Wahl eines anderen Priors würde ein unterschiedliches Vorwissen implizieren

Bezug zur Maximum Likelihood Methode

- sei $\underline{x} = x_1, ..., x_n$ N-Tupel von Observablen
- Likelihood unter Parameter μ $\underline{x}$ zu finden ist in diesem Fall $f(\underline{x}|\mu)$
- BT $f(\mu|\underline{x}) = f_0(\mu)\frac{f(\underline{x}|\mu)}{f(\underline{x})}$
- es kann gezeigt werden, dass der Einfluss des Priors auf die a posterior Verteilung für zunehmende Information (z.B. durch Iteration) asymptotisch abnimmt
- $f(\mu|\underline{x}) \approx \frac{f(\underline{x}|\mu)}{f(\underline{x})}$
- $f(\mu|\underline{x}) \propto f(\underline{x}|\mu) \equiv L(\mu; \underline{x})$
- asymptotisch nähert sich die a posteriori Wskt der normalisierten Likelihood

Beispiel Binomialverteilung

- $p(x|\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$
- BT $f(\theta|x) = \frac{p(x|\theta)f(\theta)}{\int p(x|t)f(t)dt}$
- Mittelwert mit MaxLike Schätzer $\hat{p}$ = arg max $p(x|\theta) = k/n$
- betrachten Standardisierte Gauss-Verteilung
  $f(\theta) = 1/\sqrt{2\pi}\exp -\theta^2/2$