KM 15(1984)
pag 45 - 52


# EMPIRISCHE BAYES REGELS IN EEN EENVOUDIG CLASSIFICATIE-PROBLEEM

Hans van Houwelingen

Samenvatting: In een versimpeld classificatie-probleem wordt nagegaan hoe informatie over de onbekende a priori kans kan worden gebruikt om optimale beslissingen te nemen. De link wordt gelegd met het Poscon-project dat betrouwbaarheidsintervallen oplevert voor de a posteriori kansen.

## 1. Het classificatie-probleem

Wij beschouwen de situatie van twee populaties individuen, aangeduid met I en II. Aan elk individu kan een stochastische variabele X worden waargenomen. De kansverdeling van X voor beide populaties is volledig bekend. Het probleem is te beslissen tot welke populatie een individu behoort als alleen de waarde van X voor dat individu is waargenomen. Als de a priori kans dat een individu tot populatie I (resp. II) behoort, bekend is, kan de a posteriori kans gegeven X = x worden berekend en op basis daarvan kan worden beslist om het individu te classificeren als behorende tot populatie I of II. Schematisch kan dat als volgt worden weergegeven.

| | populatie I | populatie II |
|---|---|---|
| a priori kans | $1-p$ | $p$ |
| dichtheid van X | $f_0(x)$ | $f_1(x)$ |
| a posteriori kans | $1-\pi_p(x)$ | $\pi_p(x)$ |

waarbij $\pi_p(x) = \Pr(II|X=x) = pf_1(x)/(pf_1(x)+(1-p)f_0(x))$.
Een natuurlijke classificatieregel is:
Classificeer als I (resp. II) wanneer $\pi_p(x) \leqslant$ (resp. $>$) k, waarbij k een constante is die nog gekozen moet worden.
Om de gedachte te bepalen, bekijken wij de situatie waarin $f_0(x) = n(x+\Delta/2)$, $f_1(x) = n(x-\Delta/2)$, met $\Delta > 0$ bekend en $n(x)$ de standaardnormale dichtheid. In dat geval is $\pi_p(x) = pe^{\Delta x}/(pe^{\Delta x}+1-p)$.
De keuze van de constante k wordt bepaald door de verliestabel. Wij bekijken de volgende situatie:

---

Voordracht gehouden op 18 mei 1984 in Groningen op de miniconferentie "De verborgen wereld van de statistiek"
Adres: Instituut voor mathematische statistiek, Budapestlaan 6,
3508 TA UTRECHT, tel. 030-531459.

VERLIESTABEL

| | classificeer als I | classificeer als II |
|---|---|---|
| populatie I | 0 | b |
| populatie II | a | 0 |
| verwacht verlies gegeven X = x | $a\pi_p(x)$ | $b(1-\pi_p(x))$ |

Uit de laatste regel is het duidelijk dat het verlies geminimaliseerd wordt door de regel:
Classificeer als I (II) wanneer $a\pi_p(x) \leqslant (>)\ b(1-\pi_p(x))$.
Dit is equivalent met de keuze $k = b/(a+b)$. Voor het geval van de twee normale verdelingen wordt dit:
Classificeer als I (II) wanneer $x \leqslant (>)\ c(p) = [\ln(b(1-p)/ap)]/\Delta$.
Het totaal verwachte verlies wordt dan

$$R(p) = apN(c(p) - \Delta/2) + b(1-p)(1 - N(c(p) + \Delta/2)).$$

(N(x) is de cumulatieve standaardnormale verdeling).
Hiermee is het probleem opgelost in het geval dat p bekend is. Wij beschouwen nu het geval dat $f_0$ en $f_1$ volledig bekend zijn, maar p onbekend is. Een dergelijke situatie is voorstelbaar in de medische diagnostiek waar uit het verleden de dichtheden $f_0$ en $f_1$ voor de populaties "gezond" en "ziek" wel bekend zijn, maar in een nieuw geopende praktijk het nog onbekend is wat de a priori kans op "ziek" is.
Gezien het voorgaande is het redelijk om te gaan kijken naar classificatieregels van het type:
Classificeer als I (II) wanneer $x \leqslant (>)\ c$,
waarbij de constante c nog gekozen moet worden. (Het kan eenvoudig worden aangetoond dat dit inderdaad een essentieel volledige klasse is in de zin van Ferguson (1967).) Duiden wij met het risico R(p,c) het verwachte verlies aan bij hantering van bovenstaande regel als p de echte a priori kans, dan zien wij dat

$$R(p,c) = apN(c - \Delta/2) + b(1-p)(1 - N(c + \Delta/2)).$$

Om de "beste" c te kiezen kunnen wij uitgaan van verschillende principes.
1) Minimax pricipe.
Het is eenvoudig na te gaan dat $\sup_p R(p,c)$ minimaal is voor $c = c^*$ de oplossing van $aN(c^* - \Delta/2) = b(1 - N(c^* + \Delta/2))$.

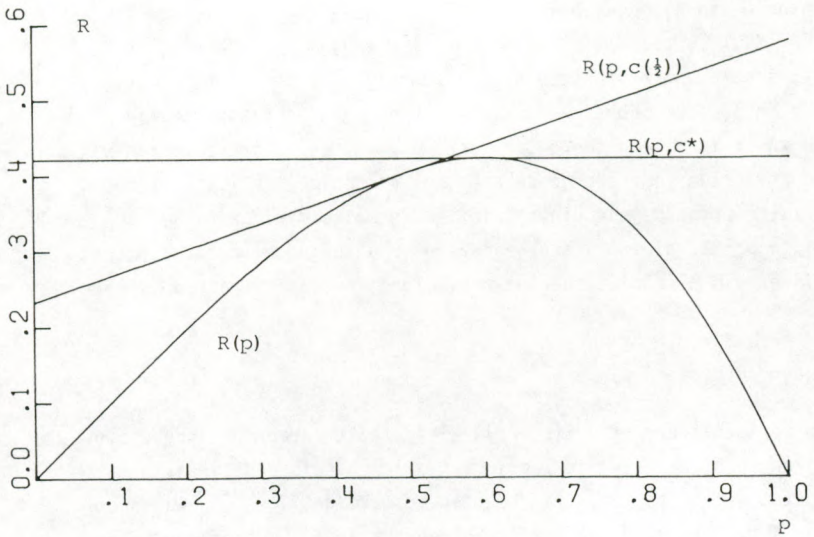
2) Bayes principe

Beschouwen wij p als onbekende parameter en veronderstellen wij dat p een verdeling G heeft, dan wordt het gemiddelde of Bayes risico

$$\int_0^1 R(p,c)dG(p) = a\bar{p}N(c - \tfrac{1}{2}\Delta) + b(1-\bar{p})(1 - N(c + \tfrac{1}{2}\Delta))$$

met $\bar{p} = \int_0^1 pdG(p)$. Deze uitdrukking is minimaal voor $c = c(\bar{p}) = \{\ln(b(1-\bar{p})/a\bar{p})\}/\Delta$. De minimale waarde van het Bayes risico is precies $R(\bar{p})$. De keuze $\bar{p} = 0.5$ levert de zogeheten minimum-oppervlakte regel op, omdat deze oppervlakte onder de functie R(p,c) minimaliseert. De ongunstigste a priori heeft een $\bar{p}^*$ die voldoet aan $c(\bar{p}^*) = c^*$. Het maximum risico $R^* = R(\bar{p}^*) = R(p,c^*)$. Wij demonstreren een en ander voor het geval a = 1, b = 2, $\Delta$ = 1 in onderstaand plaatje. Het blijkt dat $c^* = .303$, $\bar{p}^* = .596$ en $R^* = .422$. Verder is $c(0.5) = .693$.

Fig. 1: Risicofunctie bij verschillende keuzes van c.




2. Empirische Bayes aanpak

Wij bekijken nu het geval waarin het individu waarin wij geinteresseerd zijn, is voorafgegeaan door een rij vergelijkbare gevallen. (In ons gedachte-voorbeeld van de medische praktijk, veronderstellen wij nu dat de praktijk al een tijdje draait.( Schematisch kan de situatie als volgt

worden weergegeven.

| | Individu | X | Populatie |
|---|---|---|---|
| voorafgaand | 1 | $X_1$ | $T_1$ |
| | 2 | $X_2$ | $T_2$ |
| | . | . | . |
| | . | . | . |
| | . | . | . |
| | n | X | T |
| nieuw individu | | X | ? |

$T_i$ = I of II

Dat de voorafgaande gevallen vergelijkbaar zijn, preciseren wij door het kansmodel

- $P(II) = p$, $P(I) = 1 - p$
- de dichtheid van $X_i$ gegeven $T_i$ = I is $f_0$, gegeven $T_i$ = II is $f_1$
- de paren $(X_1,T_1),\ldots,(X_n,T_n)$ zijn onderling en van X onafhankelijk.

De aanpak is nu om p te schatten uit de voorafgaande gegevens, d.m.v. de schatter $\hat{p}$ en $c(\hat{p})$ te gebruiken als kritiek punt in de classificatie-regel. De kunst is om een goede schatter $\hat{p}$ te vonden. Daartoe moeten wij eerst een criterium voor een goede schatter vinden. Het blijkt dat het minimax-risico-principe tot niets leidt en de enige minimax-risico schatter $\hat{p} \equiv \bar{p}^*$ is, die de voorafgaande informatie geheel niet gebruikt. Daarom kijken wij hier niet naar de risicofunctie R, maar naar de spijt-functie

$$S(p,c) = R(p,c) - R(p)$$

en zoeken een schatter $\hat{p}$ die $S(p,c(\hat{p}))$ klein maakt in een of andere zin. Wij bekijken eerst de situatie waarin van alle voorafgaande individuen bekend is tot welke populatie zij behoren, d.w.z. dat $T_1,\ldots T_n$ waarneembaar zijn. $X_1,\ldots,X_n$ zijn dan irrelevant en Y = aantal voorafgaande individuen in populatie II is een voldoende steekproeffunctie. Y heeft de bin(n,p)-verdeling en is onafhankelijk van X. Wij bekijken drie mogelijke schatters voor p.

MA. De meest aannemelijke schatter is $\hat{p} = Y/n$.

MO. De schatter met minimum oppervlakte onder de spijtfunctie. Omdat p lineair in de spijtfunctie voorkomt is het eenvoudig na te gaan dat MO

geleverd wordt door de Bayes schatter $\hat{p} = (Y+1)/(n+2)$ t.o.v. de uniforme verdeling op (0,1).

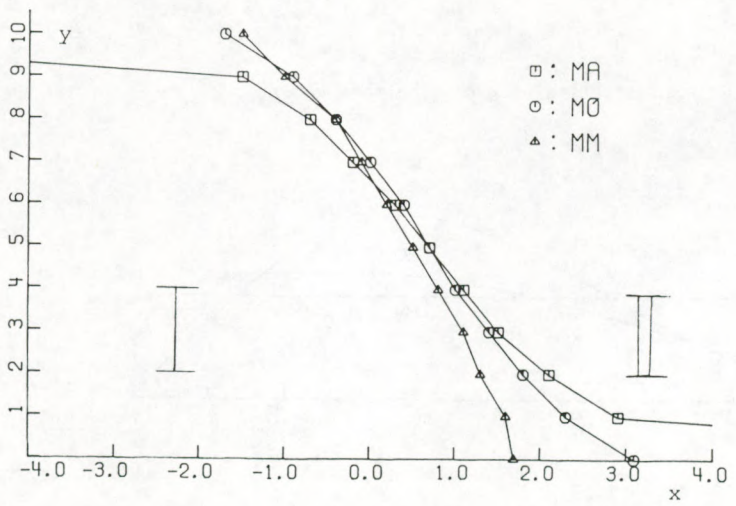
MM. De schatter die het maximum van de spijtfunctie minimaliseert. Voor deze schatter is geen expliciete uitdrukking te geven, maar het is wel mogelijk deze numerieke te construeren m.b.v. het algorithme dat de ongunstigste a priori verdeling construeert door steeds meer massa te stoppen in de punten waar de spijtfunctie maximaal is.

Wij bekijken deze schatters in detail voor het geval $a = 1$, $b = 2$, $\Delta = 1$, $n = 10$.

| | y | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MA | $\hat{p}$ | .0 | .1 | .2 | .3 | .4 | .5 | .6 | .7 | .8 | .9 | .10 |
| | $c(\hat{p})$ | $\infty$ | 2.9 | 2.1 | 1.5 | 1.1 | 0.7 | 0.3 | -0.2 | -0.7 | -1.5 | $-\infty$ |
| MO | $\hat{p}$ | .083 | .167 | .250 | .333 | .417 | .500 | .583 | .667 | .750 | .833 | .917 |
| | $d(\hat{p})$ | 3.1 | 2.3 | 1.8 | 1.4 | 1.0 | 0.7 | 0.4 | 0 | -0.4 | -0.9 | -1.7 |
| MM | $\hat{p}$ | .261 | .297 | .347 | .405 | .467 | .537 | .610 | .680 | .756 | .841 | .897 |
| | $c(\hat{p})$ | 1.7 | 1.6 | 1.3 | 1.1 | 0.8 | 0.5 | 0.2 | -0.1 | -0.4 | -1.0 | -1.5 |

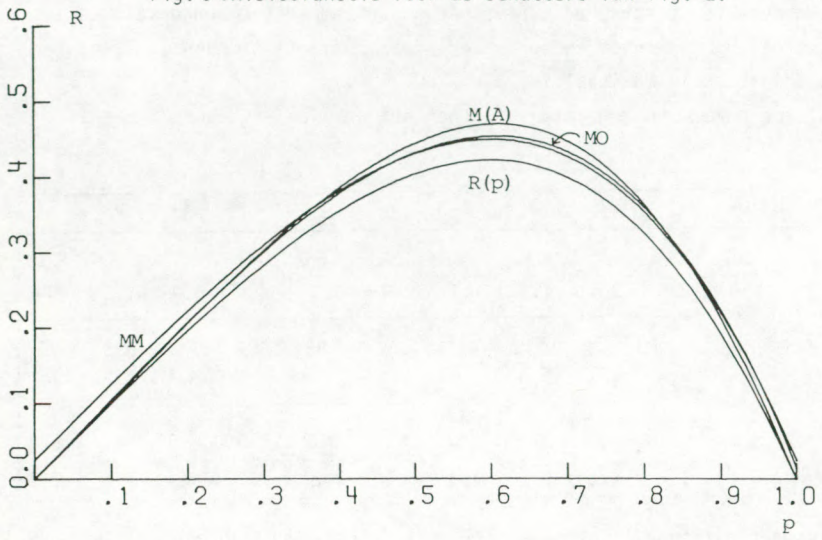
In onderstaande figuur zijn in het X-Y vlak de gebieden aangegeven behorend bij classificatie als I en classificatie als II. Wij zien dat MM meer door X en MA meer door Y wordt beinvloed.

Fig. 2: Classificatiegebieden bij verschillende schatters.



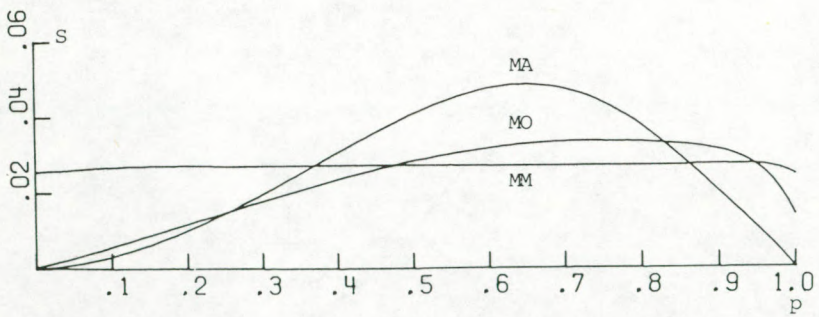
Hieronder is de grafiek weergegeven van de risicofunctie van alle drie regels en het minimale risico R. Wij zien dat de optimale strategie bij bekende p al aardig benaderd wordt.

Fig. 3 :Risicofunctie voor de schatters van fig. 2.




In de figuur hieronder is uitvergroot de spijtfunctie weergegeven. MA lijkt niet verkieslijk, maar over de keuze tussen MO en MM valt te twisten.

Fig. 4: Spijtfunctie voor de schatters van fig. 2.



In het klassieke empirische Bayes probleem zijn $T_1,\ldots,T_n$ niet waarneembaar, maar alleen $X_1,\ldots,X_n$. In ons voorbeeld komt de dokter niet achteraf te weten of de patiënten nu "ziek" of "gezond" waren, maar heeft hij alleen de X-scores geregistreerd.

In dit geval zijn $X_1,\ldots,X_n$ onderling onafhankelijk met dichtheid $pf_1(x) + (1-p)f_0(x)$. Hieruit is p wel te schatten, maar het is niet onmiddellijk duidelijk wat een goede schatter is. De moeilijkheid is dat er geen voldoende steekproeffunctie voor p is en dat zowel de meest aannemelijke als Bayesschatters bij toenemende n steeds moeilijker uit te rekenen zijn.

Tsao (1980) geeft een algorithme om Bayesschatters te bepalen, maar dat blijft rekenintensief. Van Houwelingen (1976) gebruikt een schatter van het type $\hat{p} = \sum_1^n s(X_i)/n$ en bepaalt s zodanig dat $\hat{p}$ zuiver is met minimax variantie.

In een recent ontwikkeld computerprogramma wordt als criterium genomen het minimaliseren van het maximum van de benaderde spijtfunctie. Men kan m.b.v. Taylorontwikkeling aantonen dat $S(p,c(\hat{p})) \approx h(p)(p-\hat{p})^2$. M.b.v. een algorithme dat de Bayes-keuze voor s bij de ongunstigste verdeling van p berekent, kan de minimax-spijt regel bepaald worden. Deze berekeningen zijn ook tijdrovend, maar de rekentijd is onafhankelijk van n. Voor het geval a = 1, b = 2, $\Delta$ = 1, n = 10 vond het programma als optimale schatter $\hat{p} = \sum_1^n s(X_i)/n$ met $s(x) = 0.73 + 0.43\ (e^x-1)/(.73e^x+0.27)$.

(Als $\hat{p}$ buiten het interval [0,1] valt, wordt $\hat{p}$ afgekapt op 0 of 1.) Onderstaande grafieken geven de benaderde risico- en spijtfunctie voor de corresponderende classificatieregel. Ook in deze situatie, waar minder informatie over p beschikbaar is, wordt de optimale regel bij bekende p al aardig benaderd.

Fig. 5: Risicofunctie behorend bij $\hat{p}$

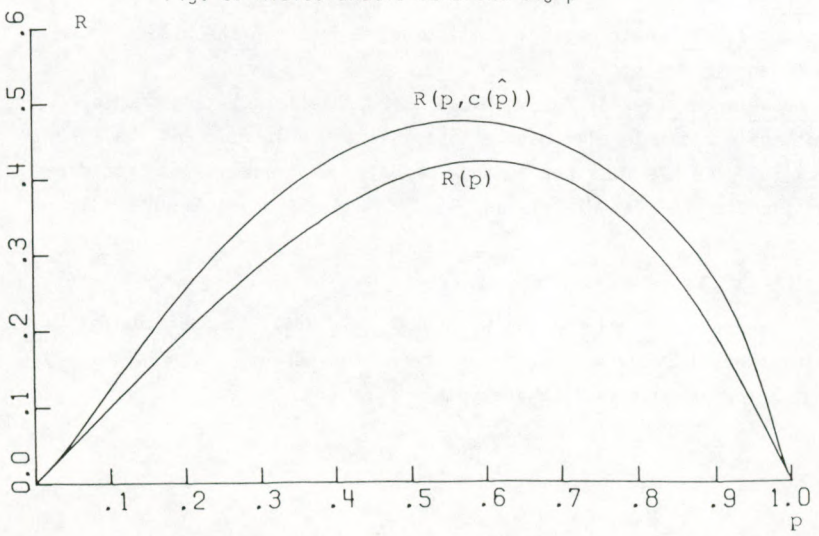
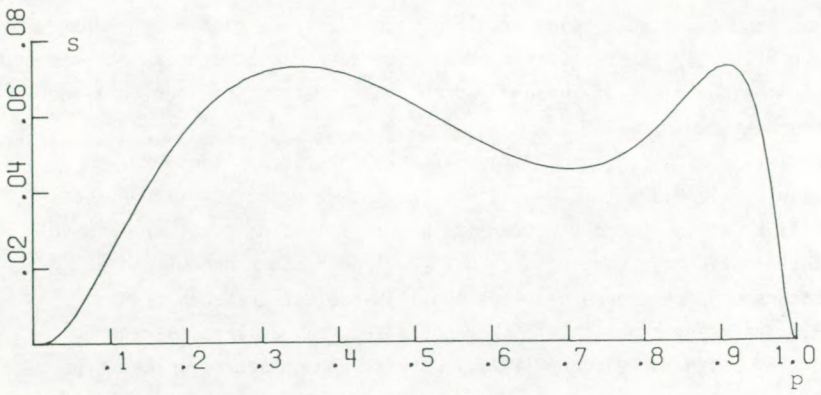


Fig. 6: Spijtfunctie behorend bij $\hat{p}$




3. Relatie met POSCON-project

Zoals al vermeld is dit artikel een weergave van een lezing op een mini-conferentie in Groningen waar het POSCON-project*een belangrijk onderwerp van gesprek was. De POSCON-aanpak zou in boven besproken problemen een betrouwbaarheidsinterval voor de a posteriori kans $\pi_p(x)$ geven. De gebruiker moet op basis van zijn eigen verliesfunctie dan een beslissing nemen. Alle beschikbare informatie wordt dus gereduceerd tot een schatter plus foutmarge van $\pi_p(x)$. De vraag is in hoeverre bij deze reductie relevante informatie verloren gaat en of dit leidt tot classificatieregels met duidelijk groter risico. Boven geschetste eenvoudige modellen kunnen van belang zijn om het effect van de reductie te analyseren en na te gaan of de reductie geoorloofd is. Dit is nog een onderwerp van nadere studie.

Referenties

T.S. Ferguson (1967). Mathematical statistics, a decision theoretic approach. Academic Press.

J.C. van Houwelingen (1974). An empirical Bayes rule for testing simple hypothesis against simple alternatice. Statistica Neerlandia, 28, 209.221.

H.J. Tsao (1980). On the risk performance of Bayes empirical Bayes procedures for classification between N(-1,1) and N(1,1). Statistica Neerlandica 34, 197-208.

*) (Noot van de redactie) Zie bijv.: W. Schaafsma (1985). Standard errors of posterior probabilities and how to use them. Te verschijnen in P.R. Krishnaiah (ed.), Multivariate Analysis - VI.

*Ontvangen: 18-6-84*