# Naiv Bayes klassifier

## Bayes klassifikation

For at introducere teorien om Bayes naive klassifikation, vil vi starte med at se på et eksempel for at få en idé om, hvad Bayes klassifickation går ud på.

Vi vil se på en person, og vi ønsker at give et bud på om vedkommende stemmer på rød eller blå blok. Vi har på forhånd oplysninger om en del andre personer og ønsker at bruge den viden til at give det bedste bud på, om personen stemmer på rød eller blå blok.

Her har vi følgende data, der viser, hvem der stemmer på rød og blå blok.

| | Alle | Mænd | Kvinder | Ung | Ældre | Sjælland | Jylland | Anden bopæl |
|---|---|---|---|---|---|---|---|---|
| Rød | 51.85% | 48.00% | 55.00% | 65.00% | 47.47% | 54.00% | 49.90% | 52.78% |
| Blå | 48.15% | 52.00% | 45.00% | 35.00% | 52.53% | 46.00% | 50.10% | 47.22% |
| Antal | 10000 | 4500 | 5500 | 2500 | 7500 | 3000 | 4500 | 2500 |

**Opgave**

Brug tabellen ovenfor og giv det bedste bud på hvilken blok en person stemmer på:

a) Hvis det er en tilfældig person.

b) Hvis det er en mand.

Fra skemaet med oplysninger kan det være svære at give et bud på, hvad en ældre kvinde fra Sjælland vil stemme på, da oplysningen om køn tyder på personen vil stemme på rød, mens information om, at det er en ældre person, tyder på personen vil stemme på blå. Endelig vil oplysningen om, at kvinden bor på Sjælland igen få os til at tænke, at hun stemmer på rød blok.

Her kunne vi selvfølgelig løse problemet ved at få information for hver kombination af køn, aldersgruppe og bopæl. Men det viser sig ikke at være en helt gangbar fremgangsmåde. Forklaringen følger her: Hvis vi ser på kombinationer af køn, aldersgruppe og bopæl vil det i dette eksempel give $2 \cdot 2 \cdot 3 = 12$ kombinationer, og hvis vi i stedet havde set på, om man svarer ja eller nej til 50 spørgsmål, vil man kunne få $2^{50}$ forskellige kombinationer af svar. Hvis man ser på en person, der har svaret på de 50 spørgsmål, kan man her forvente, at man i ens data kun har ganske få eller måske slet ingen personer, der har svaret på fuldstændig samme måde, og der vil ikke være meget at basere ens bud på.

Derfor ønsker vi en metode, hvor vores bud på hvad en ny person vil stemme på udelukkende baseres på information svarende til det fra skemaet ovenfor, hvor vi ikke ser på alle de forskellige kombinationer. Det er *det* Naive Bayes klassifikation kan.

## Bayes klassifier

I det følgende indfører vi det nødvendige matematik og notation til Naive Bayes klassifikation. Først og fremmest indfører vi en stokastisk variabel $Y$, som kan antage de værdier, der svarer til vores forskellige forudsigelser/bud. I vores eksempel vil

$$Y \in \{bl, rd\}.$$

Lidt mere generelt siger man, at $Y$ skal være en diskret stokastisk variabel med et bestemt antal mulige udfald, og der behøver altså ikke nødvendigvis kun at være to udfald.

Derudover indfører vi en stokastisk variabel $\mathbf{X}$, hvor de mulige udfald er alle kombinationer af informationer. Her kan vi tænke $\mathbf{X}$ som en stokastisk vektor $\mathbf{X} = (X_1, X_2, ..., X_q)$, hvor man ved

eksemplet kunne sige $X_1$:køn, $X_2$:aldersgruppe og $X_3$:bopæl, og et udfald kunne være $\mathbf{x} = (kvinde, ldre, Sjlland)$.

For hvert udfald af $Y$ ønsker vi at bestemme sandsynligheden for at værdien $y$ antages, når vi allerede har observeret, at $\mathbf{X} = \mathbf{x}$.

Sandsynligheden vil vi skrive som

$$P(Y = y \mid \mathbf{X} = \mathbf{x})$$

Denne notation og betydningen deraf ser vi snart på.

Vi kalder $P(Y = y \mid \mathbf{X} = \mathbf{x})$ en **posterior sandsynlighed**, fordi den udtrykker sandsynligheden for $Y$ **efter** (post), vi har informationen $\mathbf{x}$.

Det mest sandsynlige udfald for $Y$, når vi har informationen $\mathbf{x}$, betegnes $C(\mathbf{x})$ og kaldes **Bayes klassifikation**.

## Betinget sandsynlighed og uafhængighed

Først vender vi dog lige tilbage til notationen $P(Y = y \mid \mathbf{X} = \mathbf{x})$, som vi kaldte en posterior sandsynlighed. I sandsynlighedsregningen kalder vi det også for en **betinget sandsynlighed**, hvilket er grunden til notationen $P(Y = y \mid \mathbf{X} = \mathbf{x})$.

Givet to hændelser $A$ og $B$ så benyttes notationen $P(A \mid B)$ som sandsynligheden for, at $A$ sker, når det er givet, at $B$ er sket. Det læses derfor også som sandsynligheden for $A$ givet $B$.

Så $P(Y = y \mid \mathbf{X} = \mathbf{x})$ er derved sandsynligheden for $Y = y$, når det er givet, at $\mathbf{X} = \mathbf{x}$.

Et banalt eksempel kunne være at $Y$ angiver antal ben på et givent dyr, mens $\mathbf{X}$ angiver dyrearten. Her er det oplagt, at sandsynligheden for fire eller to ben afhænger af hvilken dyreart, der er tale om.

Formelt defineres betinget sandsynlighed for to *hændelser* $A$ og $B$ som:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \tag{1}$$

Udtrykket $P(A \cap B)$ i tælleren er sandsynligheden for *fælleshændelsen* mellem $A$ og $B$ – det vil sige hændelsen at både $A$ og $B$ omtræffer – og i nævneren sørger vi for, at man kun ser på de udfald, hvor $B$ er givet[1].

[1] Man siger også, at nævneren *normaliserer* sandsynligheden i forhold til sandsynligheden for hændelsen $B$.

### Eksempel med betinget sandsynlighed

Lad os fokusere på en almindelig terning med seks sider. Lad $B$ være hændelsen at antal øjne er mindre eller lig med 3. Det vil sige, at hændelsen $B$ består af udfaldene: $B = \{\ ,\ ,\ \}$. Lad hændelsen $A$ være udfald med ulige antal øjne: $A = \{\ ,\ ,\ \}$.

Da kan vi nemt indse, at

$$P(A) = 3/6 = 1/2$$

samt ligeledes at

$$P(B) = 1/2$$

på grund af det symmetriske udfaldsrum.

Ser vi imidlertid på den betingede sandsynlighed for at $A$ indtræffer givet, at $B$ allerede er indtruffet, får vi $P(A \mid B)$. Det svarer til sandsynligheden for at slå et ulige antal øjne, hvis vi allerede ved at antallet af øjne er mindre end eller lig med 3.

Først ser vi, at $A \cap B = \{\ ,\ ,\ \} \cap \{\ ,\ ,\ \} = \{\ ,\ \}$, hvilket igen på grund af det symmetriske sandsynlighedsfelt betyder, at

$$P(A \cap B) = 2/6 = 1/3$$

Efter at vi normaliserer sandsynligheden ud fra betingelsen om at $B$ **er** indtruffet får vi

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{1/3}{1/2} = \frac{2}{3}.$$

At betinge med hændelsen $B$ svarer i dette simple eksempel til at *indskrænke* udfaldet for $A$ fra alle ulige øjne til dem, som er mindre end eller lig med 3. Der er således tre mulige udfald i vores "$B$-verden", hvoraf to er ulige.

## Stokastisk uafhængighed

Man siger, at to hændelser $A$ og $B$ er uafhængige af hinanden, hvis

$$P(A \cap B) = P(A) \cdot P(B)$$

Hvis vi ser på udtrykket for $P(A \mid B)$ i (1) og antager, at $A$ og $B$ er uafhængige, ser vi at

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \overset{\text{uafh.}}{=} \frac{P(A) \cdot P(B)}{P(B)} = P(A)$$

Med andre ord betyder det, at sandsynligheden for $A$ givet $B$ er den samme som sandsynligheden for $A$. Det vil sige, at oplysningen om, at $B$ allerede er indtruffet, ikke ændrer på sandsynligheden for $A$. Information om $B$ tilfører altså ikke noget nyt i forhold til information om $A$, og det giver derfor mening af sige, at $A$ og $B$ er uafhængige af hinanden.

## Bayes' sætning

En meget vigtig matematisk egenskab ved betinget sandsynlighed er muligheden for at ombytte rollerne i formlen, således vi kan udtrykke $P(B \mid A)$ ud fra vores viden om $P(A \mid B)$. Sætningen kaldes Bayes' sætning (eller formel) og kan let vises ved først at bestemme $P(A \cap B)$ ved at isolere denne sandsynlighed. Fra (1) får vi

$$P(A \cap B) = P(A \mid B) \cdot P(B)$$

På helt tilsvarende vis må der også gælde, at

$$P(B \cap A) = P(B \mid A) \cdot P(A)$$

Og da $A \cap B = B \cap A$ må også $P(A \cap B) = P(B \cap A)$. De to ovenfor udledte sandsynligheder, må derfor være ens:

$$P(A \mid B) \cdot P(B) = P(B \mid A) \cdot P(A)$$

Her Kan $P(A \mid B)$ isoleres

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

Dette resultat er netop **Bayes' sætning**:

> **Sætning 0.1** (Bayes' sætning)**.** *Lad $A$ og $B$ være hændelser, hvor $P(B) \neq 0$. Da gælder, at*
>
> $$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

Vi kan altså ved at kende $P(B \mid A)$, $P(B)$ og $P(A)$ udtrykke den betingede sandsynlighed $P(A \mid B)$. Vi vender lige om lidt tilbage til, hvad vi kan bruge det til.

Som sidste bemærkning er det væsenligt at understrege at $P(A \mid B) \neq P(B \mid A)$ med mindre $P(A) = P(B)$ jævnfør (1) ovenfor. F.eks. er sandsynligheden for et tilfældigt dyr er en elefant, givet dyret har fire ben ikke den samme som sandsynligheden for, at dyret har fire ben givet, at dyret er en elefant!

## Binær Bayes klassifier

Antag nu at $Y$ kun kan antage to tilstande som ved eksemplet med *rød* eller *blå*. I dette tilfælde oversætter man ofte de to udfald til henholdsvis 0 og 1, eller i visse sammenhænge til $-1$ og $+1$. Husk på at **Bayes klassifikationen** $C(\mathbf{x})$ er det mest sandsynlige udfald for $Y$, når vi har informationen $\mathbf{x}$. I det tilfælde hvor $Y$ kun kan antage to tilstande, får vi derfor

$$C(\mathbf{x}) = \begin{cases} 0 & \text{hvis } P(Y = 0 \mid \mathbf{X} = \mathbf{x}) > P(Y = 1 \mid \mathbf{X} = \mathbf{x}) \\ 1 & \text{ellers} \end{cases} \tag{2}$$

Dette kan også udtrykkes på anden vis:

$$P(Y = 0 \mid \mathbf{X} = \mathbf{x}) > P(Y = 1 \mid \mathbf{X} = \mathbf{x}) \Leftrightarrow \frac{P(Y = 0 \mid \mathbf{X} = \mathbf{x})}{P(Y = 1 \mid \mathbf{X} = \mathbf{x})} > 1.$$

Her er vi dog nødt til at antage, at

$$P(Y = 1 \mid \mathbf{X} = \mathbf{x}) \neq 0,$$

så vi ikke kommer til at dividere med 0.

Bruger vi denne omskrivning, kan vi udtrykke den binære Bayes klassifikation i (2) på denne måde:

$$C(\mathbf{x}) = \begin{cases} 0 & \text{hvis } \frac{P(Y=0|\mathbf{X}=\mathbf{x})}{P(Y=1|\mathbf{X}=\mathbf{x})} > 1 \\ 1 & \text{ellers} \end{cases} \tag{3}$$

I det følgende vil vi benytte os af Bayes' sætning til at se på, hvordan ovenstående brøk kan beregnes.

## Bayes' sætning i anvendelse

Vi bruger først Bayes' sætning til at udtrykke $P(A \mid C)$ og $P(B \mid C)$.

$$P(A \mid C) = \frac{P(C \mid A)P(A)}{P(C)} \quad \text{og} \quad P(B \mid C) = \frac{P(C \mid B)P(B)}{P(C)},$$

Når vi bestemmer forholdet mellem $P(A \mid C)$ og $P(B \mid C)$ vil vi kunne slippe af med nævneren, som de har til fælles:

$$\begin{aligned} \frac{P(A \mid C)}{P(B \mid C)} &= \frac{\frac{P(C|A)P(A)}{P(C)}}{\frac{P(C|B)P(B)}{P(C)}} = \frac{P(C \mid A)P(A)}{P(C)} \cdot \frac{P(C)}{P(C \mid B)P(B)} \\ &= \frac{P(C \mid A)P(A)}{P(C \mid B)P(B)}, \end{aligned}$$

hvor vi har udnyttet, at man dividerer med en brøk ved at gange med den omvendte brøk.

I den binære Bayes klassifikation i (3) indgår brøken $\frac{P(Y=0|\mathbf{X}=\mathbf{x})}{P(Y=1|\mathbf{X}=\mathbf{x})}$, som vi ved at benytte ovenstående kan omskrive til:

$$\frac{P(Y = 0 \mid \mathbf{X} = \mathbf{x})}{P(Y = 1 \mid \mathbf{X} = \mathbf{x})} = \frac{P(\mathbf{X} = \mathbf{x} \mid Y = 0)P(Y = 0)}{P(\mathbf{X} = \mathbf{x} \mid Y = 1)P(Y = 1)}. \tag{4}$$

## Naiv Bayes klassifier

Ud fra udtrykket for forholdet mellem de to posterior sandsynligheder $P(Y = 0 \mid \mathbf{X} = \mathbf{x})$ og $P(Y = 1 \mid \mathbf{X} = \mathbf{x})$ som indgår i (4) kan vi se, at der indgår to typer af sandsynligheder:

$$P(\mathbf{X} = \mathbf{x} \mid Y = y) \quad \text{og} \quad P(Y = y)$$

Disse benævnes henholdsvis *likelihood* og *prior sandsynlighed*, idet $P(\mathbf{X} = \mathbf{x} \mid Y = y)$ udtrykker *likelihooden* (troligheden) for at observere $\mathbf{X} = \mathbf{x}$ givet $Y = y$. Omvendt er *prior sandsynligheden* $P(Y = y)$ et udtryk for forhåndsandsynligheden for at $Y = y$. Altså bruger vi disse betegnelser:

$$\begin{aligned} &\text{Likelihood:} && P(\mathbf{X} = \mathbf{x} \mid Y = y) \\ &\text{Prior sandsynlighed:} && P(Y = y) \end{aligned}$$

Vi kan sammenfatte udtrykket i (4) til det såkaldte *posterior forhold*:

$$\underbrace{\frac{P(Y = 0 \mid \mathbf{X} = \mathbf{x})}{P(Y = 1 \mid \mathbf{X} = \mathbf{x})}}_{\text{Posterior forhold}} = \underbrace{\frac{P(\mathbf{X} = \mathbf{x} \mid Y = 0)}{P(\mathbf{X} = \mathbf{x} \mid Y = 1)}}_{\text{Likelihood forhold}} \cdot \underbrace{\frac{P(Y = 0)}{P(Y = 1)}}_{\text{Prior forhold}}$$

Hvis vi vender tilbage til vores spørgsmål om at stemme på blå eller rød blok og så kan vi sige, at

$$Y = \begin{cases} 0 & \text{hvis der stemmes på rød blok} \\ 1 & \text{hvis der stemmes på blå blok} \end{cases}$$

Hvis $x = (kvinde, ung, Sjlland)$, så udtrykker $P(\mathbf{X} = \mathbf{x} \mid Y = 0)$ således sandsynligheden for, at en person er en ung kvinde fra Sjælland givet, at personen stemmer på *rød blok*. Når man skal bestemme den sandsynlighed skal vi huske, at det er sandsynligheden for det *samlede* udsagn med køn, alder og bopæl.

For at kunne beregne ovenstående sandsynligheder bliver vi nødt til at antage et eller andet, der gør det muligt. Man siger, at vi opstiller en *model*.

Én af de simpleste modeller er at antage at køn, alder og bopæl er uafhængige af hinanden givet $Y = y$. Denne forsimplende

antagelse har medvirket til metodens navn: *Naiv Bayes* eller *Uafhængig Bayes klassifikation.*

Det betyder ifølge vores tidligere definition af uafhængighed at

$$\begin{aligned} P(\mathbf{X} = \mathbf{x} \mid Y = y) &= P(X_1 = x_1, X_2 = x_2, \ldots, X_q = x_q \mid Y = y) \\ &= P(X_1 = x_1 \mid Y = y) P(X_2 = x_2 \mid Y = y) \cdots P(X_q = x_q \mid Y = y) \\ &= \prod_{i=1}^{q} P(X_i = x_i \mid Y = y), \end{aligned}$$

hvor $\prod$-symbolet i sidste linje betyder, at vi tager produktet af alle faktorerne på formen $P(X_i = x_i \mid Y = y)$ fra $i = 1$ op til $q$ – altså præcist det som står i linjen over. Det minder således om sum-tegnet

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n,$$

men blot for multiplikation i stedet for addition.

## Posterior forholdet, score og vægte

Samler vi nu udtrykkene, som indgår i vores posterior forhold i (4), samtidig med at vi antager, at $X_1, X_2, \cdots, X_q$ er uafhængige af hinanden givet $Y$, får vi nedenstående:

$$\begin{aligned} \frac{P(Y = 0 \mid \mathbf{X} = \mathbf{x})}{P(Y = 1 \mid \mathbf{X} = \mathbf{x})} &= \frac{P(Y = 0)}{P(Y = 1)} \cdot \frac{P(\mathbf{X} = \mathbf{x} \mid Y = 0)}{P(\mathbf{X} = \mathbf{x} \mid Y = 1)} \\ &= \frac{P(Y = 0)}{P(Y = 1)} \prod_{i=1}^{q} \frac{P(X_i = x_i \mid Y = 0)}{P(X_i = x_i \mid Y = 1)}, \end{aligned}$$

hvor hver faktor på højre siden bidrager ligeligt til, om observationen $\mathbf{x}$ skal klassificeres som $Y = 0$ eller $Y = 1$.

Når vi skal lave beregninger på computer baseret på data, er det ofte væsentligt at tage højde for numerisk præcision. Alle tal på en computer skal *repræsenteres* af et endelig antal bits. Det betyder, at visse tal (f.eks. $1/3$) bliver afrundet efter et

vist antal decimaler. Derfor kan der opstå problemer, når man enten ganger eller adderer meget små (eller store) tal sammen. For at undgå dette i udtrykket ovenfor, er det derfor tit en god idé at benytte sig af (den naturlige) logaritme på begge sider af lighedstegnet:

$$\ln\left(\frac{P(Y=0 \mid \mathbf{X}=\mathbf{x})}{P(Y=1 \mid \mathbf{X}=\mathbf{x})}\right) = \ln\left(\frac{P(Y=0)}{P(Y=1)}\right) + \sum_{i=1}^{q} \ln\left(\frac{P(X_i=x_i \mid Y=0)}{P(X_i=x_i \mid Y=1)}\right), \tag{5}$$

hvor vi har brugt logaritmeregnereglen

$$\ln(a \cdot b) = \ln(a) + \ln(b)$$

gentagende gange.

Vi minder om, at forholdet mellem $P(Y=0 \mid \mathbf{X}=\mathbf{x})$ og $P(Y=1 \mid \mathbf{X}=\mathbf{x})$ er af særlig interesse omkring værdien 1 – se (3). Når forholdet er 1 betyder det, at de to klasser er lige sandsynlige givet $\mathbf{x}$. Endvidere, når forholdet er over 1, er $Y=0$ mere sandsynlig end $Y=1$, og når det er under 1, er det omvendte tilfældet.

Lad os nu se på hvilken effekt det får, at vi ikke længere ser direkte på det posterior forhold

$$\frac{P(Y=0 \mid \mathbf{X}=\mathbf{x})}{P(Y=1 \mid \mathbf{X}=\mathbf{x})}$$

men i stedet på logaritmen af det posterior forhold

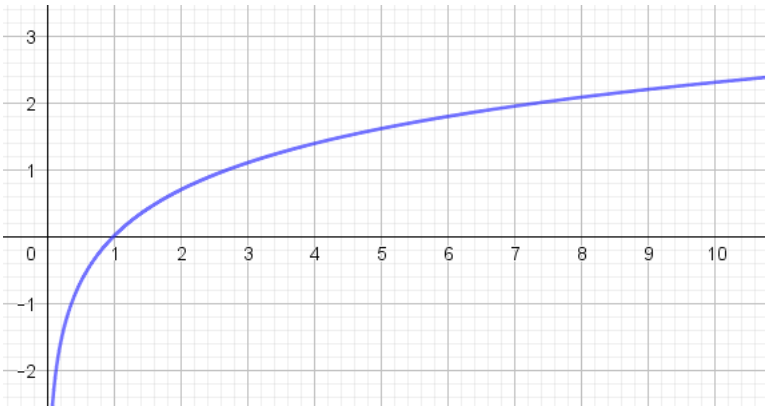
$$\ln\left(\frac{P(Y=0 \mid \mathbf{X}=\mathbf{x})}{P(Y=1 \mid \mathbf{X}=\mathbf{x})}\right)$$

I figur 1 nedenfor er grafen for logaritmefunktionen tegnet for $x \in ]0, 10]$.

Vi ved, at $\ln(1) = 0$ (hvilket også kan ses på grafen i figur 1), samt at for $x < 1$ er $\ln(x) < 0$, mens for $x > 1$ er $\ln(x) > 0$.

Så når vi ser på

$$\ln\left(\frac{P(Y=0 \mid \mathbf{X}=\mathbf{x})}{P(Y=1 \mid \mathbf{X}=\mathbf{x})}\right)$$




Figur 1: Grafen for $f(x) = \ln(x)$.

bliver det vigtige nu, om denne størrelse er positiv eller negativ.

Lad os derfor indføre $S$ som en *score*, der er lig med logaritmen til posterior forholdet:

$$S = \ln\left(\frac{P(Y = 0 \mid \mathbf{X} = \mathbf{x})}{P(Y = 1 \mid \mathbf{X} = \mathbf{x})}\right) \tag{6}$$

Den binære Bayes klassifikation i (3) kan derfor i stedet skrives ved hjælp af scoren $S$ på denne måde:

$$C(\mathbf{x}) = \begin{cases} 0 & \text{hvis} \quad S > 0 \\ 1 & \text{ellers} \end{cases} \tag{7}$$

Vi ved således, at hvis $S > 0$, så klassificerer vi $\mathbf{x}$, som $C(x) = 0$ og ellers $C(x) = 1$.

Sammenholder vi definitionen af $S$ i (6) med udtrykket i (5), ser vi, at $S$ også kan skrives som

$$\begin{aligned} S &= \ln\left(\frac{P(Y = 0 \mid \mathbf{X} = \mathbf{x})}{P(Y = 1 \mid \mathbf{X} = \mathbf{x})}\right) \\ &= \ln\left(\frac{P(Y = 0)}{P(Y = 1)}\right) + \sum_{i=1}^{q} \ln\left(\frac{P(X_i = x_i \mid Y = 0)}{P(X_i = x_i \mid Y = 1)}\right), \end{aligned} \tag{8}$$

Indfører vi nu bidragene til $S$ som $w_0$ og $w_i(x_i)$ således

$$w_0 = \ln\left(\frac{P(Y=0)}{P(Y=1)}\right) \quad \text{og} \quad w_i(x_i) = \ln\left(\frac{P(X_i = x_i \mid Y=0)}{P(X_i = x_i \mid Y=1)}\right)$$

kan vi skrive udtrykket for $S$ i (8) som

$$S = w_0 + \sum_{i=1}^{q} w_i(x_i), \tag{9}$$

hvor det tydeliggøres, at hvis $w_i(x_i) > 0$, så understøtter bidraget fra den $i$'te oplysning $x_i$, at $Y = 0$ og ellers hvis $w_i(x_i) < 0$ at $Y = 1$. Denne egenskab gør, at man også kan omtale $w_i(x_i)$ som en slags "bevismæssig" vægt.

### Vægten $w_0$

Vi har altså set i (5) og (9), hvordan vi kan omskrive forholdet mellem posterior sandsynlighederne for de to klasser $Y = 0$ og $Y = 1$ til en sum af bidrag.

Det første led $w_0$ afhænger ikke af nogen information $x_i$, og vi har tidligere omtalt disse sandsynligheder som *prior* sandsynligheder. Man kan sige, at det er vores umiddelbare bud på hvad $Y$ er, uden at vi kender noget som helst til informationerne i $\mathbf{x}$.

Når vi går fra vores *model*, som vi har udledt i det foregående, til at skulle implementere den i en specifik anvendelse, bliver vi derfor nødt til at estimere de parametre, som indgår i modellen. For $w_0$ betyder det, at vi skal estimere både $P(Y = 0)$ og $P(Y = 1)$. Her vil det være oplagt blot at estimere $P(Y = 0)$ og $P(Y = 1)$ ud fra træningsdata ved helt simpelt at bestemme andelen, som stemmer på henholdsvis rød og blå, hvorefter vi kan beregne

$$w_0 = \ln\left(\frac{P(Y=0)}{P(Y=1)}\right)$$

Hvis du vil se en mere teoretisk begrundelse for dette valg, kan du folde boksen nedenfor ud.

**Teoretisk begrundelse for hvordan $P(Y = 0)$ kan estimeres**

Vi ønsker at bestemme det bedst mulige estimat for $p = P(Y = 0)$ ud fra vores træningsdata. Vi vil her tænke på resultaterne fra datasættet som kommende fra et binomialforsøg med sandsynligheds-parameter $p$ og antalsparameter $n$. I eksemplet kan vi derfor lade $Z$ være en stokastisk variabel, der betegner antallet, som stemmer på rød blok. Det vil sige, at

$$Z \sim bin(n, p)$$

og fra binomialfordelingen ved vi, at

$$P(Z = r) = \binom{n}{r} p^r (1-p)^{n-r} = \frac{n!}{(n-r)!r!} p^r (1-p)^{n-r}$$

Når vi skal estimere $p$ (altså sandsynligheden for at stemme på rød blok – det vil sige $P(Y = 0)$) ud fra data benyttes en metode, som kaldes for maksimum likelihood estimation. Den går i al sin enkelhed ud på at bestemme den værdi af $p$, som gør de data, vi har set, mest sandsynlige. Altså vil vi maksimere udtrykket ovenfor med hensyn til $p$. Dette kan vi gøre ved at differentiere udtrykket og sætte det lig med 0. I stedet for at arbejde direkte med $P(Z = r)$ er det nemmere at arbejde med udtrykket for $\ln(P(Z = r))$, idet der gælder, at $f(p) = P(Z = r)$ og $\ln(f(p))$ har maksimum ved samme $p$ (fordi logaritmefunktionen er voksende).
Vi finder derfor først et udtryk for $\ln(P(Z = r))$:

$$\ln(P(Z = r)) = \ln \binom{n}{r} + r \cdot \ln(p) + (n - r) \cdot \ln(1 - p)$$

hvor vi har brugt logaritmeregnereglerne

$$\ln(a \cdot b) = \ln(a) + \ln(b) \quad \text{og} \quad \ln(a^x) = x \cdot \ln(a)$$

Derfor bliver den afledede med hensyn til $p$

$$\frac{d}{dp} \left( \ln(P(Z = r)) \right) = \frac{r}{p} - \frac{n - r}{1 - p}$$

Sætter vi ovenstående lig 0 og isolerer for $p$, får vi, at

$$\hat{p} = \frac{r}{n}$$

hvilket svarer til den andel af de $n$ observationer, som har $Y = 0$. Vi sætter en "hat" på $p$ for at tydeliggøre, at det er et estimat af $p$ – og altså ikke den ukendte, *sande* værdi af $p$.
Bemærk, at man også kan vise, at denne værdi af $p$ rent faktisk svarer til et maksimumssted.

### Vægtene $w_i$

De øvrige bidrag til $S$ afhænger af den specifikke værdi af $x_i$. Det er altså her data for observationen, vi ønsker at klassificere, kommer ind i billedet. Her vil vægten $w_i(x_i)$, som bidrager til den samlede score $S$, afhænge af informationen[2] $x_i$.

[2] Hvis $w_i(x_i)$ er mere eller mindre konstant for forskellige værdier af $X_i$, betyder det, at den $i$'te information ikke er særlig informativ (og måske bør udelades fra modellen).

Ved hver information $X_i$ estimeres $P(X_i = x_i \mid Y = y)$ ved at se på andelen af $x_i$ blandt alle træningsdata med $Y = y$. Vægtene $w_i$ estimeres således på tilsvarende måde som for $w_0$.

Når disse estimater er fundet, kan man bestemme

$$w_i(x_i) = \ln\left(\frac{P(X_i = x_i \mid Y = 0)}{P(X_i = x_i \mid Y = 1)}\right)$$

### Eksempel med rød/blå blok

Lad os se på eksemplet fra tidligere med at stemme på rød eller blå blok, hvor vi tænker på $Y = 0$ som en stemme på rød blok og $Y = 1$ som en stemme på blå blok.

Vi havde allerede følgende information fra træningsdata:

| | Alle | Mænd | Kvinder | Ung | Ældre | Sjælland | Jylland | Anden bopæl |
|---|---|---|---|---|---|---|---|---|
| Rød | 51.85% | 48.00% | 55.00% | 65.00% | 47.47% | 54.00% | 49.90% | 52.78% |
| Blå | 48.15% | 52.00% | 45.00% | 35.00% | 52.53% | 46.00% | 50.10% | 47.22% |
| Antal | 10000 | 4500 | 5500 | 2500 | 7500 | 3000 | 4500 | 2500 |

Fra dette bestemmes først vægten $w_0$ ved

$$w_0 = \ln\left(\frac{P(Y=0)}{P(Y=1)}\right) = \ln\left(\frac{51.85\%}{48.15\%}\right) = 0.074$$

For at kunne beregne vores vægte

$$w_i(x_i) = \ln\left(\frac{P(X_i = x_i \mid Y=0)}{P(X_i = x_i \mid Y=1)}\right)$$

skal vi for at beregne tælleren i ovenstående brøk have fat på hvor stor en andel af de stemmer, der går til rød blok, som kommer fra henholdsvis mænd, kvinder, unge, ældre og så videre.

Vi tager beregningen for kvinder og starter med at finde $P(X_1 = kvinde \mid Y = 0)$. Her ved vi at 55% af de i alt 5500 kvinder stemte på rød blok. Samtidig ved vi, at der var 51.85% ud af i alt 10000 adspurgte (det vil sige 5185 pesroner), som stemte på rød blok. Derfor får vi

$$P(X_1 = kvinde \mid Y = 0) = \frac{55\% \cdot 5500}{5185} = 58.34\%$$

Da $X_1$ kun kan antage værdierne kvinde og mand, ved vi også, at

$$P(X_1 = mand \mid Y = 0) = 100\% - 58.34\% = 41.66\%.$$

På tilsvarende måde kan vi finde $P(X_1 = kvinde \mid Y = 1)$. Her ved vi, at 45% af de 5500 kvinder stemte på blå blok, og samtidig ved vi, at der var 4815 stemmer på blå blok i alt (igen 48.15% af 10000). Derfor får vi

$$P(X_1 = kvinde \mid Y = 1) = \frac{45\% \cdot 5500}{4815} = 51.40\%$$

og

$$P(X_1 = mand \mid Y = 1) = 100\% - 51.40\% = 48.60\%$$

Alle tilsvarende sandsynligheder kan beregnes, så man kan se hvor stor en andel af stemmerne på de to blokke, der komme fra hver gruppe. Resultatet ses i nedenstående tabel:

| | Mænd | Kvinder | Ung | Ældre | Sjælland | Jylland | Anden bopæl |
|---|---|---|---|---|---|---|---|
| Rød | 41.66% | 58.34% | 31.34% | 68.66% | 31.24% | 43.31% | 25.45% |
| Blå | 48.60% | 51.40% | 18.17% | 81.83% | 28.66% | 46.82% | 24.52% |

Nu kan vi bestemme vægtene $w_i(x_i)$. Her findes $w_1(kvinde)$ ved

$$w_1(kvinde) = \ln\left(\frac{P(X_1 = kvinde \mid Y = 0)}{P(X_1 = kvinde \mid Y = 1)}\right) = \ln\left(\frac{58.34\%}{51.40\%}\right) = 0.1267$$

Herunder ses vægtene for alle grupper:

| $w_0$ | $w_1$ | | $w_2$ | | $w_3$ | | |
|---|---|---|---|---|---|---|---|
| | Mænd | Kvinder | Ung | Ældre | Sjælland | Jylland | Anden bopæl |
| 0.074 | −0.154 | 0.127 | 0.545 | −0.175 | 0.086 | −0.078 | 0.037 |

Tidligere havde vi indset, at når $w_i(x_i) > 0$, så understøtter bidraget fra den $i$'te oplysning $x_i$, at $Y = 0$ (altså at stemme på rød blok). I tabellen ovenfor ses det derfor, at oplysningerne kvinde, ung, Sjælland og anden bopæl gør det mere sandsynligt med en stemme på rød blok, mens oplysningerne mand, ældre og Jylland gør det mere sandsynligt med en stemme på blå blok.

Vi kan nu beregne scoren $S$ for en kvinde, som er ældre og fra Sjælland, altså hvor $x = (kvinde, ldre, Sjlland)$.

$$\begin{aligned} S &= w_0 + \sum_{i=1}^{q} w_i(x_i) = w_0 + w_1(kvinde) + w_2(ldre) + w_3(Sjlland) \\ &= 0.074 + 0.127 + (-0.175) + 0.086 = 0.11 \end{aligned} \tag{10}$$

Derved bliver forudsigelsen ud fra Bayes Naive metode, at en ældre kvinde, der bor på Sjælland, med størst sandsynlighed stemmer på rød blok.

## Opgave: Ham or Spam?

I denne opgave skal vi se på, hvordan man kan lave et simpelt spamfilter. Vi har et datasæt med 35000 mails med oplysninger om emailens oprindelse (Danmark, Europa uden Danmark, USA og anden oprindelse), afsenders mailadresse (firma, Google, Hotmail og anden) samt indhold (dating, spil og andet). Målet er, at man gerne ud fra disse oplysninger gerne automatisk vil kunne afgøre, om det er spam og derved at mailen ikke skal vises i mail-boxen, eller om det er Ham (non-spam). For hver af de 35000 mails er det også noteret om, der er tale om spam eller ham.

Træningsdata består af

| | Oprindelse | | | |
|---|---|---|---|---|
| | DK | Europa | USA | Andet |
| Spam | 20% | 30% | 35% | 55% |
| Ham | 80% | 70% | 65% | 45% |
| Antal | 10000 | 12000 | 8000 | 5000 |

og

| | Mail | | | | Indhold | | |
|---|---|---|---|---|---|---|---|
| | Firma | Google | Hotmail | Andet | Dating | Spil | Andet |
| Spam | 10% | 20% | 60% | 80% | 80% | 90% | 12.5% |
| Ham | 90% | 80% | 40% | 20% | 20% | 10% | 87.5% |
| Antal | 17000 | 6450 | 5400 | 6150 | 4325 | 4975 | 25700 |

**Opgave 1**

a) Hvis man blot modtager en mail fra en Hotmail-konto, vil man da tænke, at det er spam eller ham?

b) Hvis man blot modtager en tilfældig mail, vil man da tænke, at det er spam eller ham?

c) Forklar hvorfor det kan være svært at afgøre, om det er spam, hvis man modtager en mail fra Danmark,

som er sendt fra en firma-mail og hvor indhold er relateret til dating.

**Opgave 2**

Indfør selv stokastiske variable $X1$, $X2$, $X3$ og $Y$ og angiv udfaldsrum for hver af dem.

**Opgave 3**

a) Opstil posterior forholdet

$$\frac{P(Y = 0 \mid X_1 = x_1, X_2 = x_2, X_3 = x_3)}{P(Y = 1 \mid X_1 = x_1, X_2 = x_2, X_3 = x_3)}$$

og forklar det smarte, der er sket ved at bruge Bayes formel.

b) Redegør for betydningen af forholdet og forklar hvorfor man ser på, om det er over eller under 1.

**Opgave 4**

Angiv den antagelse (modelforudsætning), som anvendes ved Bayes Naive klassifikation, og forklar hvad det gør for beregningen af sandsynlighederne fra opgave 3.

**Opgave 5**

Bestem prior forholdet

$$\frac{P(Y = 0)}{P(Y = 1)}$$

og beregn på den baggrund vægten

$$w_0 = \ln\left(\frac{P(Y = 0)}{P(Y = 1)}\right)$$

**Opgave 6**

Bestem alle betingede sandsynligheder

$$P(X_i = x_i \mid Y = 0) \quad \text{og} \quad P(X_i = x_i \mid Y = 1)$$

for hver enkelt information, givet at det er henholdsvis spam og ham (i alt 22 sandsynligheder) og forklar idéen bag én af disse udregninger.

**Opgave 7**

For hver af de 11 informationer bestemmes forholdet mellem sandsynlighederne

$$\frac{P(X_i = x_i \mid Y = 0)}{P(X_i = x_i \mid Y = 1)}$$

**Opgave 8**

Forklar hvordan man kommer fra resultaterne i opgave 7 til vægte og udregn vægtene hørende til hver af informationerne (i alt 11).

**Opgave 9**

Forklar hvad der sker ved at benytte logaritmen på udtrykket fra opgave 3 og 4, hvor man ellers ganger faktorer sammen.

**Opgave 10**

Afgør ud fra de vægte, som du har beregnet i opgave 8, hvilke informationer der taler for spam, og hvilke der taler for ham.

**Opgave 11**

Afgør ved at beregne scoren $S$, om man vil tænke at en mail er ham eller spam i følgende to situationer:

- Mailens oprindelse er andet, den er sendt fra en

hotmail-konto og omhandler ikke dating eller spil.
- Mailen er en firma-mail fra Danmark med indhold relateret til dating.