Klassificering:
Förväxlingsmatris och Utvärderingsmått

Josef Wilzén

När vi arbetar med klassificering så behöver kunna utvärdera om en modell för klassificering är bra eller dålig. Det finns en mängd olika mått som vi kan använda. Anta att vi har en modell eller ett test som kan ge minst två utfall.

När man diskuterar utvärderingsmått så nämns ofta "positiva" och "negativa" observationer eller fall. Det innebär att att man har en referensklass som är den "positiva" klassen och alla andra klasser är då "negativa". Om ni har binär klassificering med 0/1 kodning så brukar 1 vara den "positiva" klassen. Historiskt så kommer den terminologin från medicinska tester där ett provsvar kan vara positivt eller negativt.

Vi har följande beteckningar:

TP = true positive, FN = false negative
TN = true negative, FP = false positive

Vi börjar ofta med måtten:

- **Träffsäkerhet** (accuracy): andelen korrekt klassificerade observationer givet alla tillgängliga observationer
- **Felkvot** (error rate): andelen inkorrekt klassificerade observationer givet alla tillgängliga observationer

Vi brukar beräkna måtten separat för våra olika datamängder (träning, validering, test). En grundregel är att vi vill vår klassificeringsmodell ska vara tydligt bättre än att slumpmässigt gissa vilken klass ett ny observation ska ha. Så om vi skattar en modell, och sen får ut en hög felkvot så vill vi försöka förbättra modellen på något sätt eller skatta en annan typ av modell.

Exempel: Vi har två klasser med 300 observationer i varje klass och vår modell har en felkvot på 43 %, vilket inte är mycket bättre än att slumpmässigt tilldela klasser till observationer. Här är det tydligt att modellen inte är bra eller lämplig.

Det finns många situationer där träffsäkerhet eller felkvot inte är tillräckliga mått. För vissa problem så kan typ I-fel (falska positiva) vara allvarligare än

typ II-fel (falska negativa) eller tvärtom. Ibland så har vi obalanserade klasser i data vilket gör det olämligt att bara titta på träffsäkerhet/felkvot. Ofta vill veta vilken typ av fel som modellen gör.

Två andra vanliga mått är:

- **Sensitivitet** = sannolikheten att min modell/test ger rätt klass givet att testobservationen är av en viss klass. Exempel: sannolikheten att min modell ger klass 1 när den sanna klassen för min testobservation är av klass 1. Kallas även Recall och True Positive Rate (TPR).
- **Specificitet** = sannolikheten att min modell/test inte ger en viss klass när testobservationen inte är av den klassen. Exempel: sannolikheten att min modell inte ger klass 1 när testobservation inte kommer från klass 1. Om våra klasser är sjuk/frisk så har vi: sannolikheten att min modell visar frisk när personen ifråga inte är sjuk.

Två klasser

Vid klassificering så brukar vi ställa upp en förväxlingsmatris (confusion matrix) för att tydlig redovisa hur vår modell fungerar. Låt f vara antalet observationer i en given cell. Förväxlingsmatris:

		Predikterad klass		
		klass=1	**klass=0**	**Falska negativa**
Sann klass	**klass=1**	f_{11}	f_{10}	$FN_1 = f_{10}$
	klass=0	f_{01}	f_{00}	$FN_0 = f_{01}$
	Falska positiva	$FP_1 = f_{01}$	$FP_0 = f_{10}$	

f_{11} och f_{00} är korrekt klassificerade observationer. Om vi har klass 1 som vår referensklass så motsvarar f_{01} typ I-fel (falska positiva) och f_{10} motsvarar typ II-fel (falska negativa).

- Träffsäkerhet: $T = \frac{f_{11}+f_{00}}{f_{11}+f_{10}+f_{01}+f_{00}}$
- Felkvot: $E = \frac{f_{10}+f_{01}}{f_{11}+f_{10}+f_{01}+f_{00}}$
- Sensitivitet: $sens = \frac{f_{11}}{f_{11}+f_{10}}$
- Specificitet: $spec = \frac{f_{00}}{f_{00}+f_{01}}$

Notera sensitivitet och specificitet defineras utfrån den "positiva" klassen, vilket påverkar vår tolkning för ett specifikt dataset. Notera att vi kan bestäma vad som är den postiva klassen eller "1" i vår analys, tex kan bestämma att sjuk=1 eller att frisk=1.

Fler än två klasser

Nedan visas exempel med tre klasser, vilket kan generaliseras till godtyckligt antal klasser. Låt f vara antalet observationer i en given cell.

		Predikterad klass			
		klass=1	**klass=2**	**klass=3**	**Falska negativa**
Sann klass	**klass=1**	f_{11}	f_{12}	f_{13}	$FN_1 = f_{12} + f_{13}$
	klass=2	f_{21}	f_{22}	f_{23}	$FN_2 = f_{21} + f_{23}$
	klass=3	f_{31}	f_{32}	f_{33}	$FN_3 = f_{31} + f_{32}$
	Falska positiva	$FP_1 = f_{21} + f_{31}$	$FP_2 = f_{21} + f_{32}$	$FP_3 = f_{13} + f_{23}$	

Vi har följande beteckningar:

TP = true positive, FN = false negative
TN = true negative, FP = false positive
sdsfgs

För en specifik klass i, sensitivitet: $sens_i = TP_i / (TP_i + FN_i)$
För en specifik klass i, specificitet: $spec_i = TN_i / (TN_i + FP_i)$

I formlerna ovan så är i den postivia klassen, så vad som är TP, TP, FN och FP defineras utfrån den.

Exempel:
Sensitivitet för klass 2: $f_{22}/(f_{22} + f_{21} + f_{23}) \rightarrow$ sannolikheten att min modell ger klass 2 när den sanna klassen för min testobservation är av klass 2. Så här har vi att $TP = f_{22}$ och $FN = f_{21} + f_{23}$

Exempel:
Specificitet för klass 2: $(f_{11} + f_{33}) / (f_{11} + f_{33} + f_{12} + f_{32}) \rightarrow$ sannolikheten att min modell inte ger klass 2 när testobservation inte kommer från klass 2. Så här har vi $TN = f_{11} + f_{33}$ och $FP = f_{12} + f_{32}$

Nedan följer formler för hur man beräknar sensitivitet och specificitet i de olika fallen.

	Sensitivitet
klass=1	$sens_1 = \frac{f_{11}}{f_{11}+f_{12}+f_{13}}$
klass=2	$sens_2 = \frac{f_{22}}{f_{22}+f_{21}+f_{23}}$
klass=3	$sens_2 = \frac{f_{33}}{f_{33}+f_{31}+f_{32}}$

	Specificitet
klass=1	$spec_1 = \frac{f_{22}+f_{33}}{f_{22}+f_{33}+f_{21}+f_{31}}$
klass=2	$spec_2 = \frac{f_{11}+f_{33}}{f_{11}+f_{33}+f_{12}+f_{32}}$
klass=3	$spec_3 = \frac{f_{11}+f_{22}}{f_{11}+f_{22}+f_{13}+f_{23}}$