AI för naturligt språk

Introduktion till textklassificering

Marco Kuhlmann
Institutionen för datavetenskap

LINKÖPINGS UNIVERSITET

Textklassificering

- **Textklassificering** handlar om att automatiskt sortera textdokument i fördefinierade klasser.
- Begreppet *dokument* används på alla sorters texter, från korta twitterinlägg till fullständiga böcker.
- För att träna och utvärdera textklassificerare krävs dokument som är annoterade med rätt klass.

  Textklassificering kan inte upptäcka egna klasser.

Sentimentanalys

Men den välanvända tropen och välbekanta strukturen till trots är ”Palm Springs” en riktig liten pärla, genomförd med både finess och ett stort känslomässigt gehör. Det är … en berättelse som vibrerar av hjärta och smartness under sin småfåniga exteriör.

Källa

positiv

Tyvärr är ”Bliss”, utöver det ganska vackra fotot, en enda röra. Den överlastade, men svårt undergestaltade, intrigen solkas av kass dialog och ett skådespeleri som förvandlar både Wilson och Hayek till elaka karikatyrer på sig själva.

Källa

negativ

Klassificering av nyhetstexter

Sverige	Världen	Näringsliv	Sport
löntagare sossarna	president kongress	bolag redovisning	stafett skiathlon
ostlänken stambanor	brexit EU	aktier kurser	allsvenskan hemmaplan
ekholmen rattonykterhet	hertig sussex	handelsavtal protektionism	målskytt styrning

Forensisk lingvistik

Antalet mordutredningar som klarats upp på just detta sätt är för all del inte enormt, men växer stadigt. Samma tekniker används även (fast detaljerna är av lätt insedda skäl inte offentliga) för att få span på potentiella terrorister och pedofiler. Är det i själva verket Kent-Åke (68), som utger sig för att vara Lisa (13) på det ungdomliga diskussionsforumet?

Språkpolis på riktigt: ord kan fria eller fälla (SvD 2020-01-07)

AI för naturligt språk

Sentimentanalys

Marco Kuhlmann
Institutionen för datavetenskap

LINKÖPINGS UNIVERSITET

Sentimentanalys

Men den välanvända tropen och välbekanta strukturen till trots är ”Palm Springs” en riktig liten pärla, genomförd med både finess och ett stort känslomässigt gehör. Det är … en berättelse som vibrerar av hjärta och smartness under sin småfåniga exteriör.

Källa

positiv

Tyvärr är ”Bliss”, utöver det ganska vackra fotot, en enda röra. Den överlastade, men svårt undergestaltade, intrigen solkas av kass dialog och ett skådespeleri som förvandlar både Wilson och Hayek till elaka karikatyrer på sig själva.

Källa

negativ

Sentimentanalys

- **Sentimentanalys** i sin enklaste form försöker bestämma polariteten i en åsikt om t.ex. en film eller en produkt.

  positiv/negativ, positiv/negativ/neutral

- Detta kan vara en förhållandevis lätt uppgift, då åsikter ofta kan kopplas till enstaka ord.

  pärla, *finess*, *vibrerar*; *röra*, *överlastade*, *elaka*

- En enkel metod för sentimentanalys är därför att matcha orden i texten mot listor med positiva och negativa ord.

Utmaningar med sentimentanalys

- **Negationer och kontrafaktiska uttryck**

  Det var inte illa.
  Det är inte det värsta som kan hända.
  Denna film kunde ha blivit en riktig höjdare – men så blev det inte.

- **Ironi och sarkasm**

  Vad trevligt, ännu en hjärtvärmande romantisk komedi!
  Skådespelarna är så otroligt roliga, de borde bli komiker.
  Han är ju bra på det här. 😂

Aspektbaserad sentimentanalys

NEGATIV ASPEKT

Jag hatade deras fajitas,

men salladerna var jättegoda!

ASPEKT POSITIV

{fajitas: negativ, salladerna: positiv}

Liknande problem

- **Detektera subjektiva textpassager**

  Att identifiera de delar av en text som uttrycker subjektiva uppfattningar eller spekulationer, snarare än fakta.

  eng. *subjectivity detection*

- **Detektera ställningstaganden**

  Att i en debatt avgöra vilken sida en debattör står för, t.ex. om hen är för eller emot ett lagförslag.

  eng. *stance classification*

AI för naturligt språk

Textklassificering som maskininlärning

Marco Kuhlmann
Institutionen för datavetenskap

LINKÖPINGS UNIVERSITET

Textklassificering som maskininlärning

A
president
kongress

B
bolag
redovisning

C
stafett
skiathlon

A
brexit
EU

B
aktier
kurser

C
allsvenskan
hemmaplan

A
hertig
sussex

B
handelsavtal
protektionism

C
målskytt
styrning

Textklassificering som maskininlärning

träningsmängd
träning

president kongress	bolag redovisning	stafett skiathlon
A	B	C
brexit EU	aktier kurser	allsvenskan hemmaplan
A	B	C

testmängd
utvärdering

hertig sussex	handelsavtal protektionism	målskytt styrning
A	B	C
A	B	A

Varifrån får vi annoterade datamängder?

- manuell annotering

  exempel: taggning av nyhetsartiklar

- använda parallella, icke-språkliga signaler

  exempel: emotikoner i twitterinlägg

- kombinera olika datamängder

  exempel: anföranden i riksdagen + omröstningsresultat

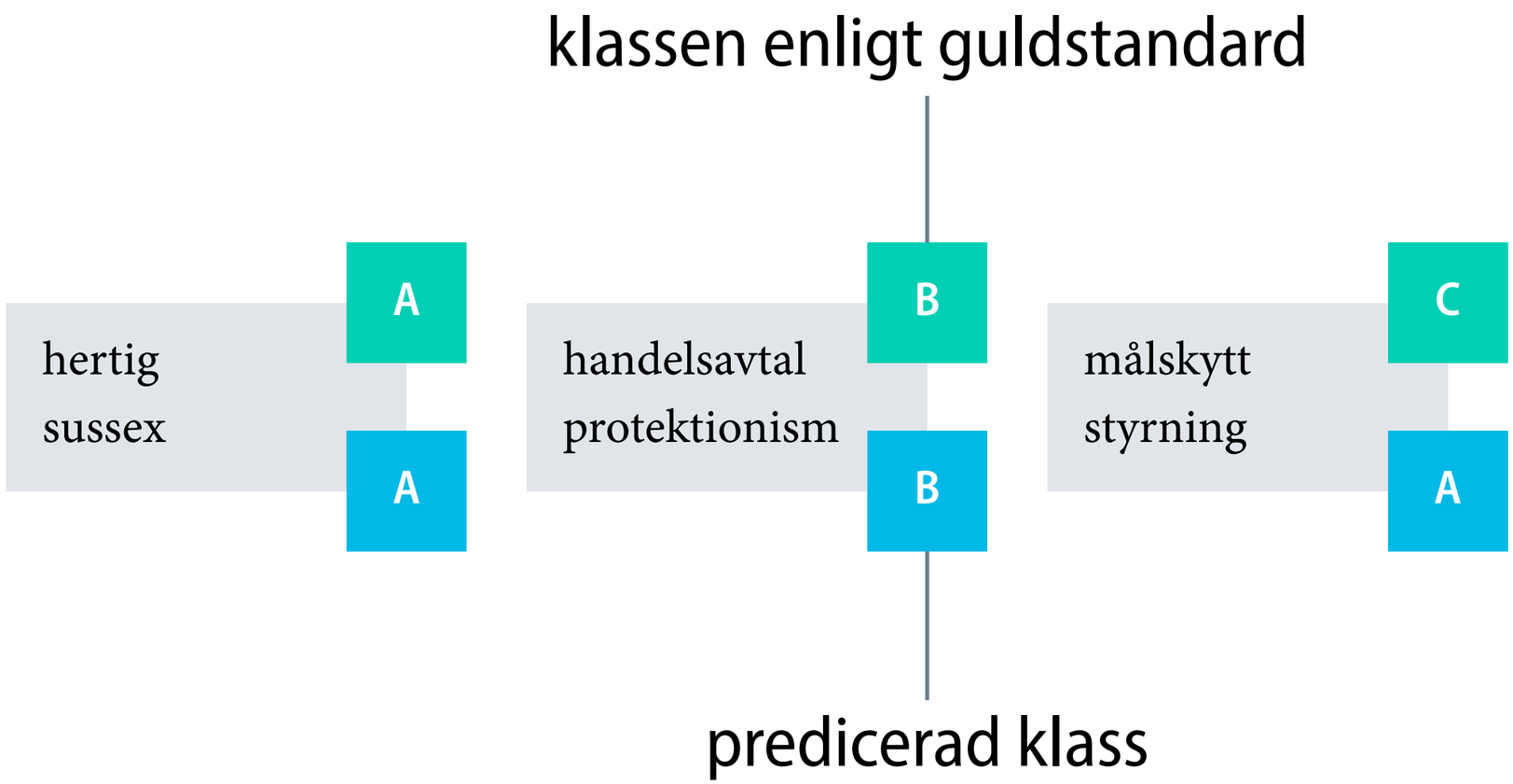
AI för naturligt språk

Korrekthet

Marco Kuhlmann
Institutionen för datavetenskap

LINKÖPINGS UNIVERSITET

Utvärdering av textklassificerare



Korrekthet

Begreppet **korrekthet** (eng. *accuracy*) betecknar andelen av alla dokument i testmängden för vilka systemet predicerar rätt klass:

$$\text{korrekthet} = \frac{\text{antal korrekt klassificerade dokument}}{\text{totalt antal dokument}}$$

Korrekthet

Dokument	Guldstandard	Predicerad klass
stafett skiathlon	Sport	Sport
allsvenskan hemmaplan	Sport	Sport
målskytt styrning	Sport	Sport
kurs avgift	Sport	Näringsliv

korrekthet: 3/4 = 75%

Förväxlingsmatris

	klassificerare ”positiv”	klassificerare ”negativ”
guldstandard ”positiv”	sanna positiva	falska negativa
guldstandard ”negativ”	falska positiva	sanna negativa

Beräkna korrekthet utifrån förväxlingsmatrisen

	klassificerare ”positiv”	klassificerare ”negativ”
guldstandard ”positiv”	sanna positiva	falska negativa
guldstandard ”negativ”	falska positiva	sanna negativa

Korrekthet vid tre klasser

$$\frac{112}{133} \simeq 84\%$$

	A	B	C
A	58	6	1
B	5	11	2
C	0	7	43

AI för naturligt språk

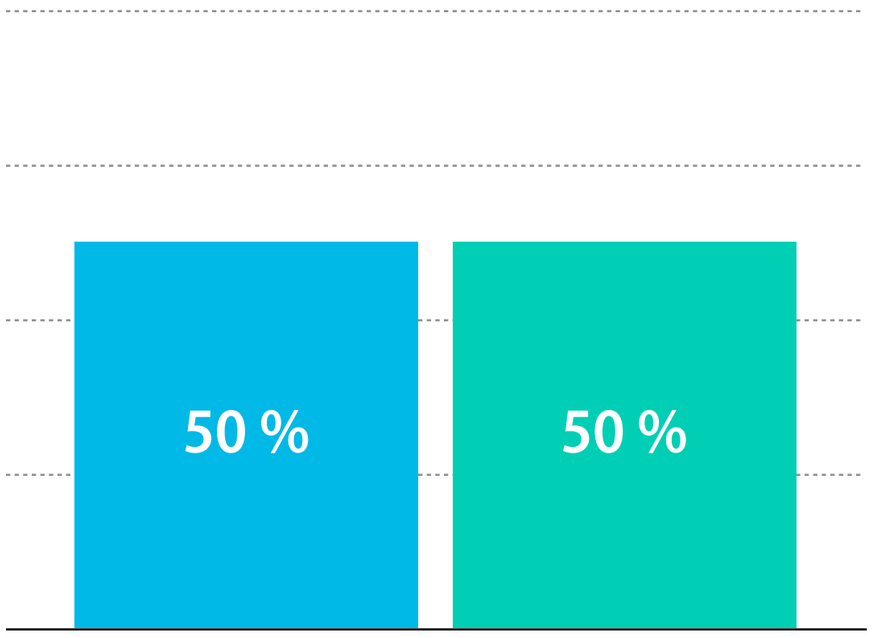
Precision och täckning

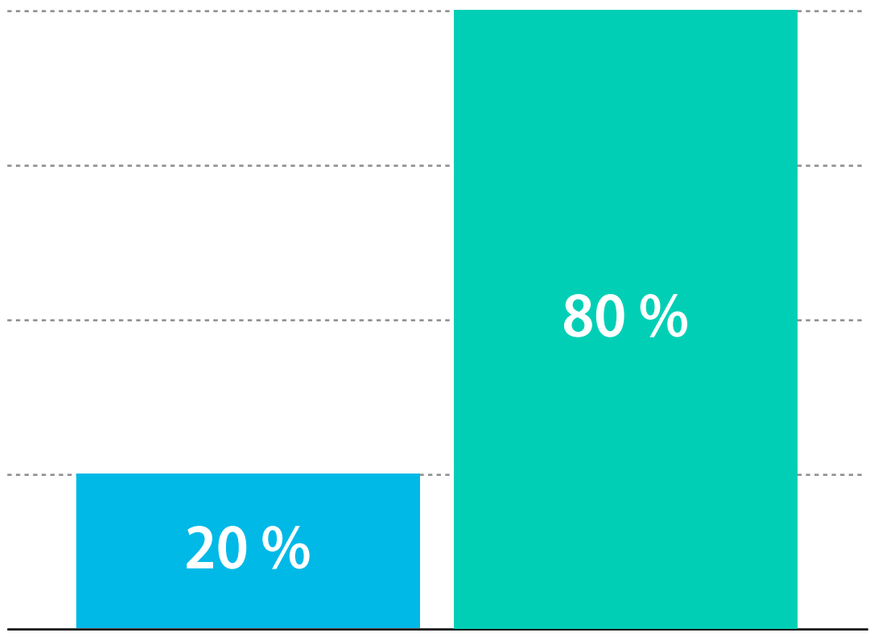
Marco Kuhlmann
Institutionen för datavetenskap

LINKÖPINGS UNIVERSITET

Korrekthet kan vara missvisande

Är 80% korrekthet bra eller dåligt?





Precision och täckning (recall)

- **Precision** och **täckning** (eng. *recall*) ”zoomar in” på hur bra ett system är på att identifiera dokument med en specifik klass *c*.

- **Precision** med avseende på klass *c* svarar på frågan:
  Hur många procent av de dokument för vilka systemet predicerar klass *c* tillhör denna klass enligt guldstandarden?

- **Täckning** med avseende på klass *c* svarar på frågan:
  För hur många procent av de dokument som enligt guldstandarden tillhör klass *c* predicerar systemet att de har denna klass?

Precision med avseende på den positiva klassen

	klassificerare ”positiv”	klassificerare ”negativ”
guldstandard ”positiv”	sanna positiva	falska negativa
guldstandard ”negativ”	falska positiva	sanna negativa

Täckning med avseende på den positiva klassen

	klassificerare ”positiv”	klassificerare ”negativ”
guldstandard ”positiv”	sanna positiva	falska negativa
guldstandard ”negativ”	falska positiva	sanna negativa

Precision med avseende på klass B

$$\frac{11}{24} \simeq 46\%$$

	A	B	C
A	58	6	1
B	5	11	2
C	0	7	43

Täckning med avseende på klass B

$$\frac{11}{18} \simeq 61\%$$

	A	B	C
A	58	6	1
B	5	11	2
C	0	7	43

AI för naturligt språk

Fler utvärderingsmått

Marco Kuhlmann
Institutionen för datavetenskap

LINKÖPINGS UNIVERSITET

Utvärdering av klassificerare i scikit-learn

	precision	recall	f1-score	support
C	0.63	0.04	0.07	671
KD	0.70	0.02	0.03	821
L	0.92	0.02	0.04	560
M	0.36	0.68	0.47	1644
MP	0.36	0.25	0.29	809
S	0.46	0.84	0.59	2773
SD	0.57	0.12	0.20	1060
V	0.59	0.15	0.24	950
accuracy			0.43	9288
macro avg	0.57	0.26	0.24	9288
weighted avg	0.52	0.43	0.34	9288

predicerad = guldstandard

Utvärdering av klassificerare i scikit-learn

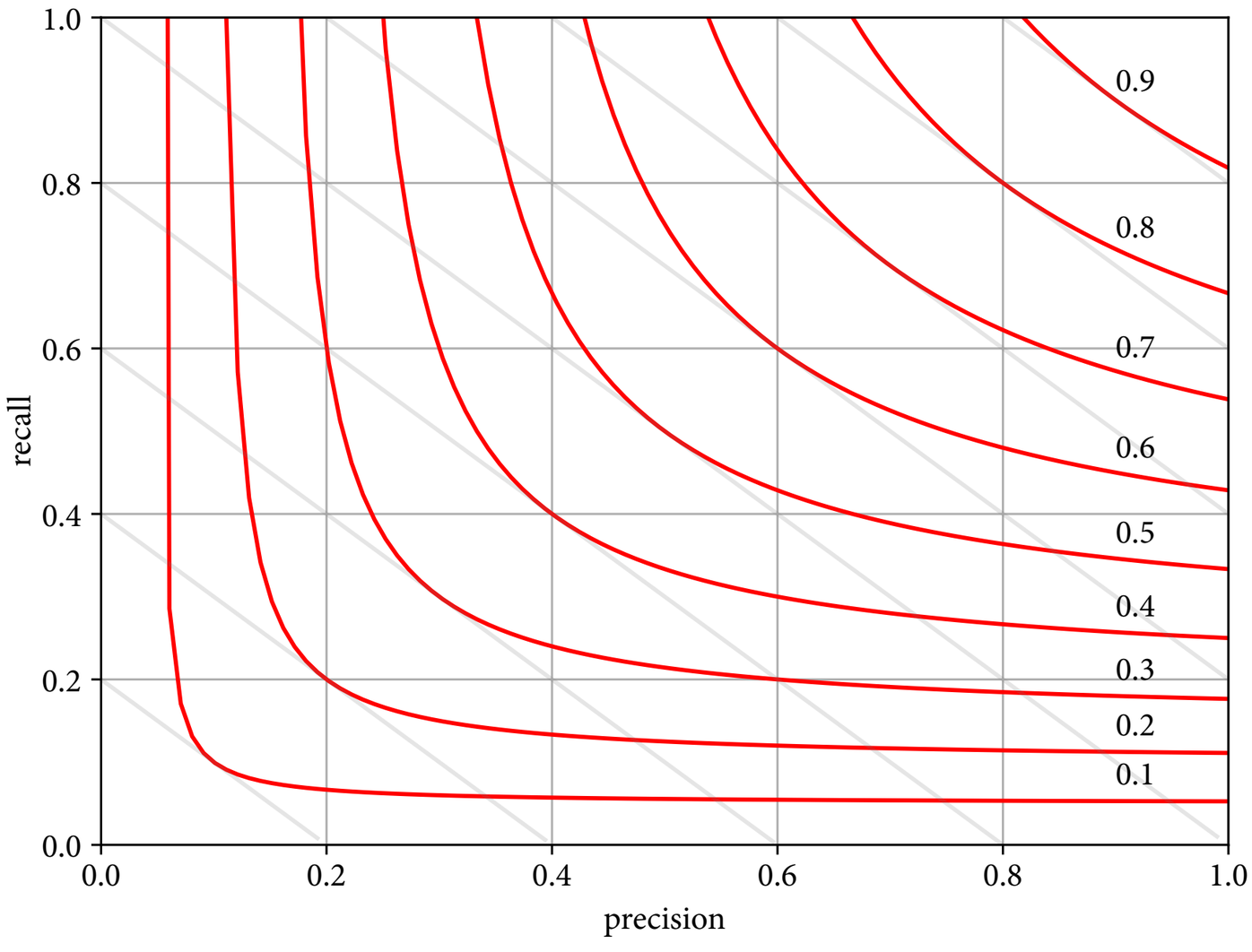
	precision	recall	f1-score	support
C	0.63	0.04	0.07	671
KD	0.70	0.02	0.03	821
L	0.92	0.02	0.04	560
M	0.36	0.68	0.47	1644
MP	0.36	0.25	0.29	809
S	0.46	0.84	0.59	2773
SD	0.57	0.12	0.20	1060
V	0.59	0.15	0.24	950
accuracy			0.43	9288
macro avg	0.57	0.26	0.24	9288
weighted avg	0.52	0.43	0.34	9288

F1-måttet

F1-värdet med avseende på klass *c* är det harmoniska medelvärdet mellan precision och täckning med avseende på denna klass:

$$\mathrm{F1} = \frac{2 \cdot \mathrm{precision} \cdot \mathrm{täckning}}{\mathrm{precision} + \mathrm{täckning}}$$

F1-värdet


1.0
0.8
0.6
0.4
0.2
0.0
recall
0.0
0.2
0.4
0.6
0.8
1.0
precision
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1

Utvärdering av klassificerare i scikit-learn

	precision	recall	f1-score	support
C	0.63	0.04	0.07	671
KD	0.70	0.02	0.03	821
L	0.92	0.02	0.04	560
M	0.36	0.68	0.47	1644
MP	0.36	0.25	0.29	809
S	0.46	0.84	0.59	2773
SD	0.57	0.12	0.20	1060
V	0.59	0.15	0.24	950
accuracy			0.43	9288
macro avg	0.57	0.26	0.24	9288
weighted avg	0.52	0.43	0.34	9288

Utvärdering av klassificerare i scikit-learn

	precision	recall	f1-score	support
C	0.63	0.04	0.07	671
KD	0.70	0.02	0.03	821
L	0.92	0.02	0.04	560
M	0.36	0.68	0.47	1644
MP	0.36	0.25	0.29	809
S	0.46	0.84	0.59	2773
SD	0.57	0.12	0.20	1060
V	0.59	0.15	0.24	950
accuracy			0.43	9288
macro avg	0.57	0.26	0.24	9288
weighted avg	0.52	0.43	0.34	9288

AI för naturligt språk

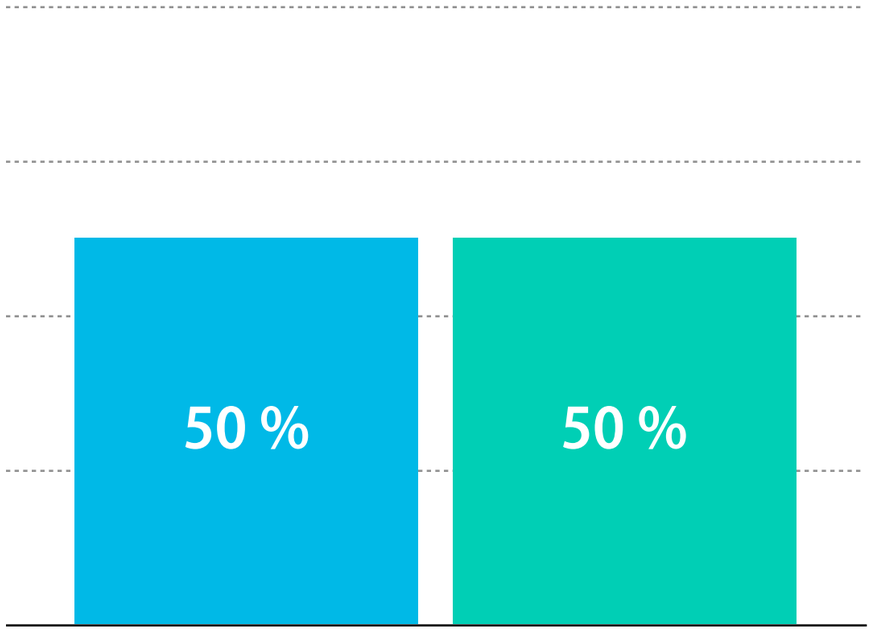
Generellt om utvärderingsmått

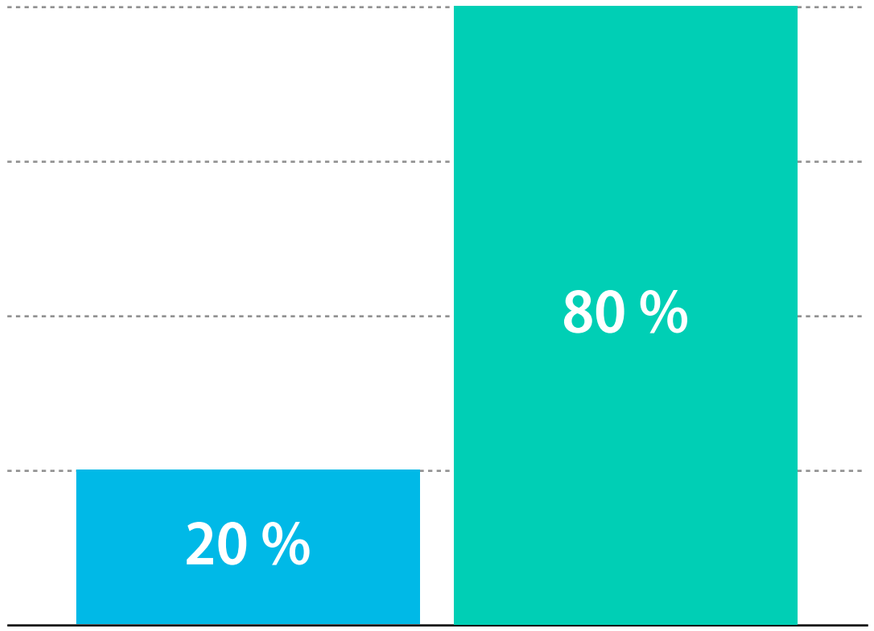
Marco Kuhlmann
Institutionen för datavetenskap

LINKÖPINGS UNIVERSITET

Tolkning av utvärderingsmåtten

Är 80% korrekthet bra eller dåligt?





Utvärderingsmått är relativa mått

- Utvärderingsmått är inga absoluta mått – det meningslöst att säga att 80% korrekthet är ”bra”.
- Utvärderingsmått kan endast användas för att jämföra en klassificerare med ett referensvärde, en **baseline**.

  ”Logistisk regression har 2 poäng högre korrekthet än Naive Bayes.”

- En enkel baseline är att predicera ett dokuments klass genom att slumpa fram den enligt klassernas fördelning i träningsdatan.

Utvärdera = uppskatta generaliseringsförmåga

- Träningsmängd och testmängd ska ses som två av varandra oberoende stickprov från en population av dokument.
- Genom att mäta klassificerarens korrekthet m.fl. på testmängden uppskattar vi dess förmåga att generalisera.

  under antagandet att framtida dokument kommer från samma population

- För att detta ska vara meningsfullt måste träningsmängd och testmängd vara disjunkta och ha samma fördelning av klasserna.

Statisk uppdelning

ursprunglig datamängd

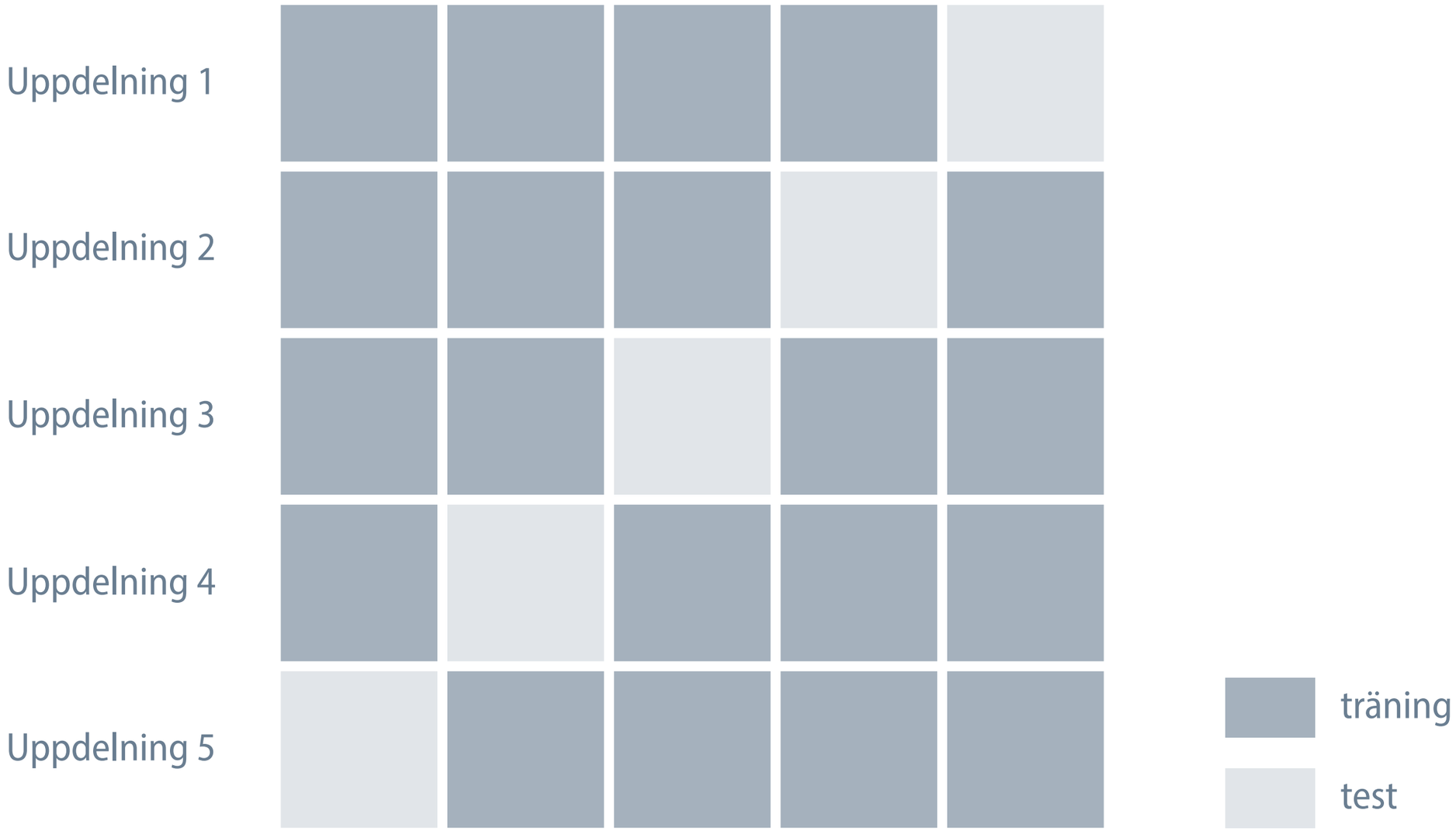
samma fördelning i delmängderna

🙂

olika fördelningar i delmängderna

🙁

Korsvalidering



AI för naturligt språk

Naive Bayes-klassificeraren

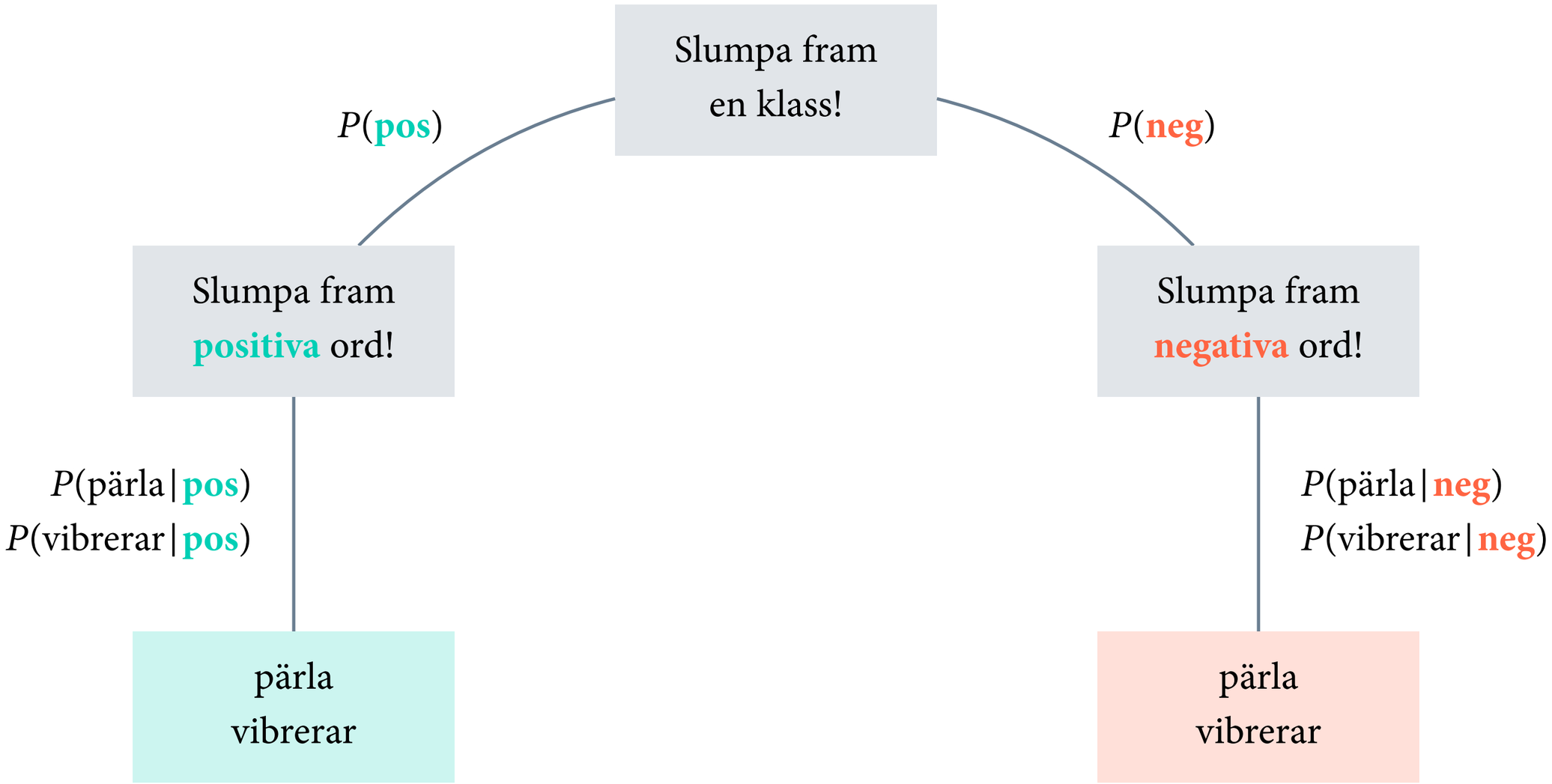
Marco Kuhlmann
Institutionen för datavetenskap

LINKÖPINGS UNIVERSITET

Naive Bayes

- **Naive Bayes-klassificeraren** är en enkel men förvånansvärt användbar klassificeringsalgoritm som bygger på Bayes' regel.
- Den kallas ”naiv” eftersom den gör starka (orealistiska) antaganden om oberoende mellan sannolikheter.
- Naive Bayes-klassificeraren finns i flera olika versioner. Här behandlar vi det som kallas för ”multinomial Naive Bayes”.

Naive Bayes som sannolikhetsmodell



Bayes' rule

- För att klassificera ett dokument skulle vi vilja veta den betingade sannolikheten $P(\text{klass}|\text{ord})$.
- Naive Bayes-modellen innehåller dock endast de omvänt betingade sannolikheterna på formen $P(\text{ord}|\text{klass})$.
- För att konvertera mellan dessa två typer av sannolikheter kan vi använda **Bayes' regel**.

  $P(\text{klass}|\text{ord}) \propto P(\text{klass})\ P(\text{ord}|\text{klass})$

Bayes' regel

Thomas Bayes
(1702–1761)

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Ingredienserna i en Naive Bayes-klassificerare

C	de fördefinierade klasserna, t.ex. {positiv, negativ}
V	de möjliga orden; modellens **vokabulär**
$P(c)$	**klassannolikheter** som svarar på frågan: ”Hur sannolikt är det att ett givet dokument tillhör klass c?”
$P(w\|c)$	**ordsannolikheter** som svarar på frågan: ”Hur sannolikt är det att ett dokument med klass c innehåller ordet w?”

Beslutsregeln i en Naive Bayes-klassificerare

antal ord i det dokument
som vi vill klassificera

$$\hat{c} \;=\; \arg\max_{c \in C} \left[P(c) \cdot \prod_{i=1}^{n} P(w_i \mid c) \right]$$

predicerad klass
för dokumentet

sannolikheten
för ordet w_i, givet klass c

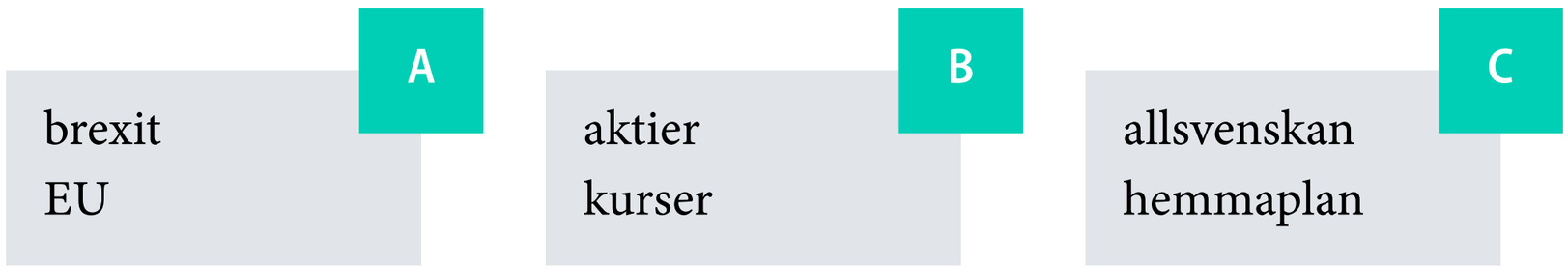
AI för naturligt språk

Inlärningsalgoritmen för Naive Bayes

Marco Kuhlmann
Institutionen för datavetenskap

LINKÖPINGS UNIVERSITET

Att träna upp en Naive Bayes-klassificerare



$P(c)$

klassannolikheter

$P(w|c)$

ordsannolikheter

Maximum Likelihood Estimation (MLE)

- Standardmetoden för att skatta sannolikheter i en modell heter **Maximum Likelihood Estimation (MLE)**.
- När vi följer MLE-principen vill vi hitta de sannolikheter som maximerar sannolikheten för träningsdatamängden.
- För Naive Bayes-klassificeraren är MLE väldigt enkel: det enda vi behöver göra är att räkna dokument och ord.

Maximum Likelihood Estimation (MLE)

	Vanlig slant		Viktad slant	
datamängd	krona	klave	krona	klave
	2	2	3	1
MLE-skattade sannolikheter	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{3}{4}$	$\frac{1}{4}$
datamängdens totala sannolikhet	$\left(\frac{1}{2}\right)^2 \cdot \left(\frac{1}{2}\right)^2 = \frac{16}{256}$		$\left(\frac{3}{4}\right)^3 \cdot \left(\frac{1}{4}\right)^1 = \frac{27}{256}$	

Maximum Likelihood Estimation (MLE)

	Vanlig slant		Viktad slant	
datamängd	**krona**	**klave**	**krona**	**klave**
	2	2	3	1
icke-MLE sannolikheter	$\frac{3}{4}$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{2}$
datamängdens totala sannolikhet	$\left(\frac{3}{4}\right)^2 \cdot \left(\frac{1}{4}\right)^2 = \frac{9}{256}$ ~~16~~		$\left(\frac{1}{2}\right)^3 \cdot \left(\frac{1}{2}\right)^1 = \frac{16}{256}$ ~~27~~	

MLE för Naive Bayes

$\#(c)$ antal dokument med guldstandard-klass c

$\#(w, c)$ antal förekomster av ordet w i dokument med klass c

$$P(c) = \frac{\#(c)}{\sum_{x \in C} \#(x)} \qquad P(w \mid c) = \frac{\#(w, c)}{\sum_{x \in V} \#(x, c)}$$

Sentimentanalys

Men den välanvända tropen och välbekanta strukturen till trots är ”Palm Springs” en riktig liten pärla, genomförd med både finess och ett stort känslomässigt gehör. Det är … en berättelse som vibrerar av hjärta och smartness under sin småfåniga exteriör.

39 token

positiv

Tyvärr är ”Bliss”, utöver det ganska vackra fotot, en enda röra. Den överlastade, men svårt undergestaltade, intrigen solkas av kass dialog och ett skådespeleri som förvandlar både Wilson och Hayek till elaka karikatyrer på sig själva.

36 token

negativ

Skattning av ordsannolikheter

Ord	Förekomster
en	2
karikatyrer	0
pärla	1
röra	0
vibrerar	1
…	

positiv

Ord	Förekomster
en	1
karikatyrer	1
pärla	0
röra	1
vibrerar	0
…	

negativ

Skattning av ordsannolikheter

Sannolikhet	Skattat värde
$P(\text{en} \mid \textbf{pos})$	2/39
$P(\text{karikatyrer} \mid \textbf{pos})$	0/39
$P(\text{pärla} \mid \textbf{pos})$	1/39
$P(\text{röra} \mid \textbf{pos})$	0/39
$P(\text{vibrerar} \mid \textbf{pos})$	1/39
…	

positiv

Sannolikhet	Skattat värde
$P(\text{en} \mid \textbf{neg})$	1/36
$P(\text{karikatyrer} \mid \textbf{neg})$	1/36
$P(\text{pärla} \mid \textbf{neg})$	0/36
$P(\text{röra} \mid \textbf{neg})$	1/36
$P(\text{vibrerar} \mid \textbf{neg})$	0/36
…	

negativ

AI för naturligt språk

Dokumentrepresentationer

Marco Kuhlmann
Institutionen för datavetenskap

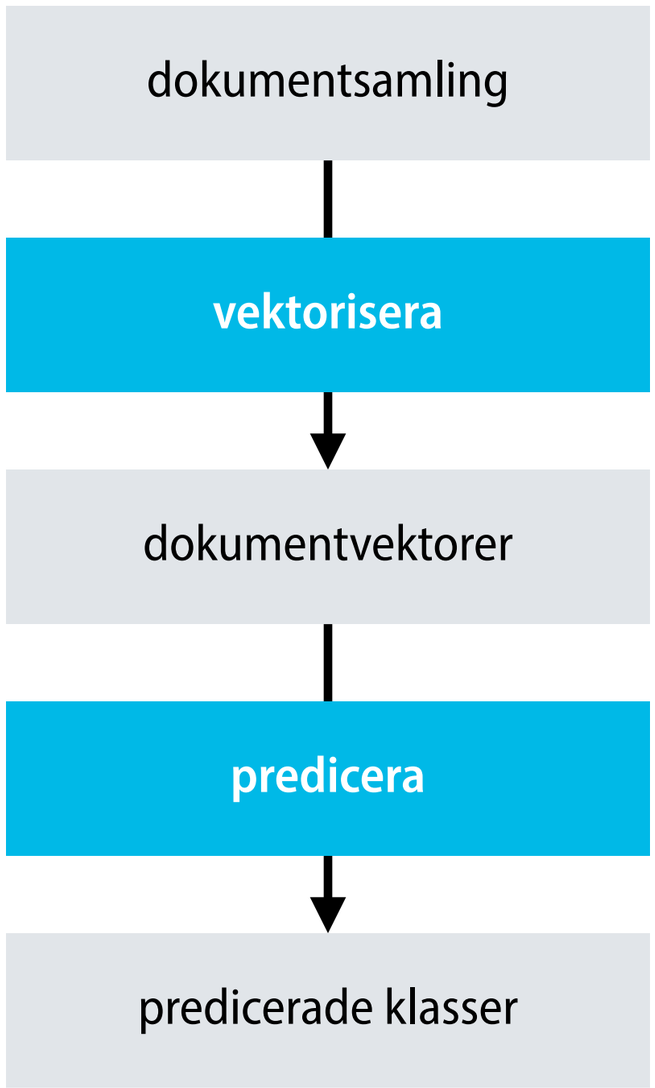
LINKÖPINGS UNIVERSITET

Dokumentrepresentationer

- För att kunna träna och tillämpa en textklassificerare behöver vi representera dokument som tal.
- I Naive Bayes-klassificeraren representerar vi dokument genom att räkna antalet ordförekomster.
- Mera generellt så kommer vi att översätta dokument till listor av tal, som vi tolkar som vektorer.

  nyckeln till artificiella neuronnät

Standardkedjan för textklassificering



Sentimentanalys

Men den välanvända tropen och välbekanta strukturen till trots är ”Palm Springs” en riktig liten pärla, genomförd med både finess och ett stort känslomässigt gehör. Det är … en berättelse som vibrerar av hjärta och smartness under sin småfåniga exteriör.

Källa

positiv

Tyvärr är ”Bliss”, utöver det ganska vackra fotot, en enda röra. Den överlastade, men svårt undergestaltade, intrigen solkas av kass dialog och ett skådespeleri som förvandlar både Wilson och Hayek till elaka karikatyrer på sig själva.

Källa

negativ

Bag-of-words

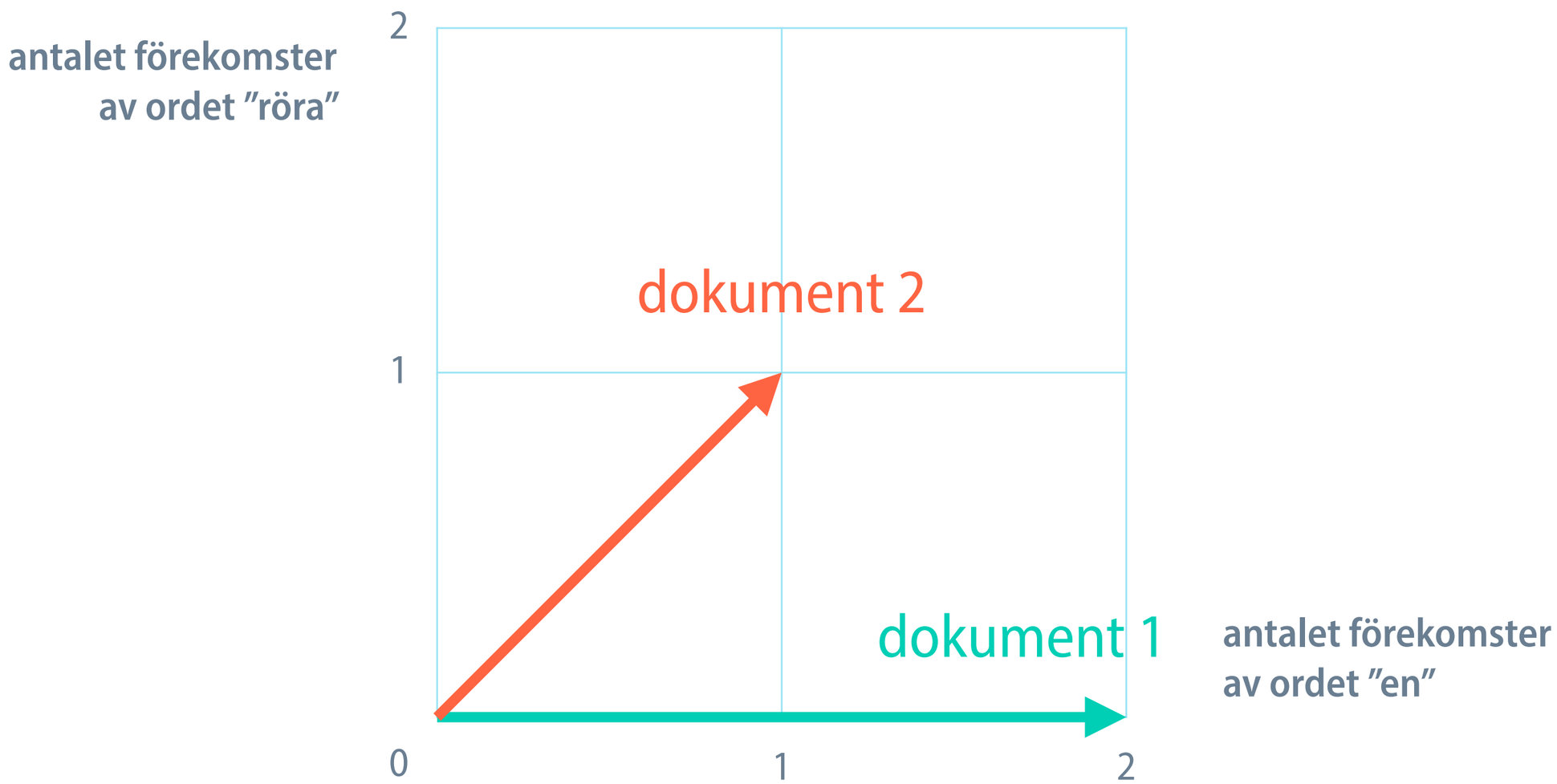
Ord	Förekomster
en	2
karikatyrer	0
pärla	1
röra	0
vibrerar	1
…	

positiv

Ord	Förekomster
en	1
karikatyrer	1
pärla	0
röra	1
vibrerar	0
…	

negativ

Bag-of-words


antalet förekomster
av ordet ”röra”
2
1
dokument 2
dokument 1
0
1
2
antalet förekomster
av ordet ”en”

Tf–idf

Ord	tf–idf
en	0,60
karikatyrer	0,00
pärla	0,40
röra	0,00
vibrerar	0,40
…	

positiv

Ord	tf–idf
en	0,30
karikatyrer	0,40
pärla	0,00
röra	0,40
vibrerar	0,00
…	

negativ

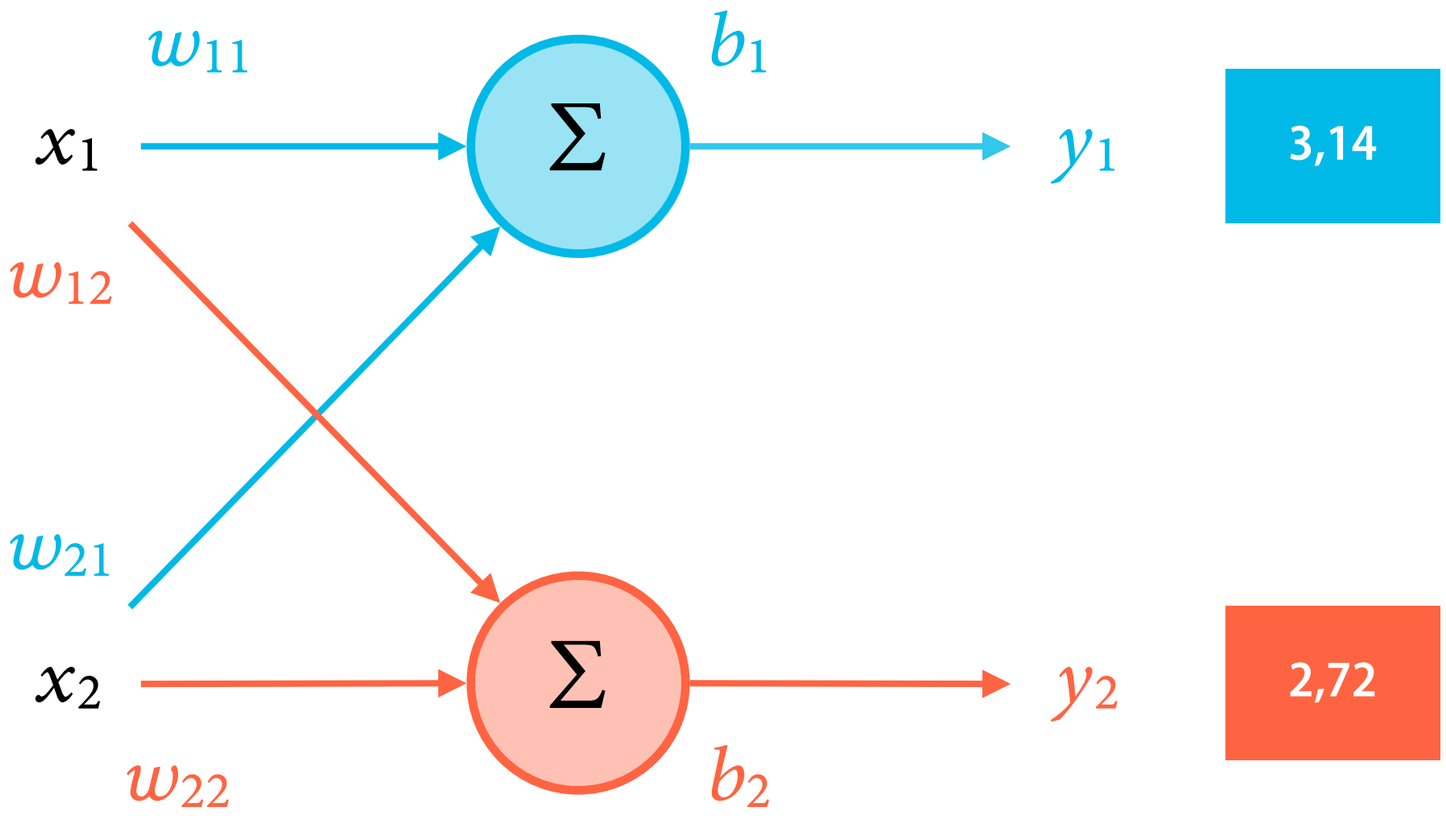
AI för naturligt språk

Logistisk regression

Marco Kuhlmann
Institutionen för datavetenskap

LINKÖPINGS UNIVERSITET

Klassificering med ett linjärt neuronnät



$$y_1 = x_1 w_{11} + x_2 w_{21} + b_1 \qquad y_2 = x_1 w_{12} + x_2 w_{22} + b_2$$

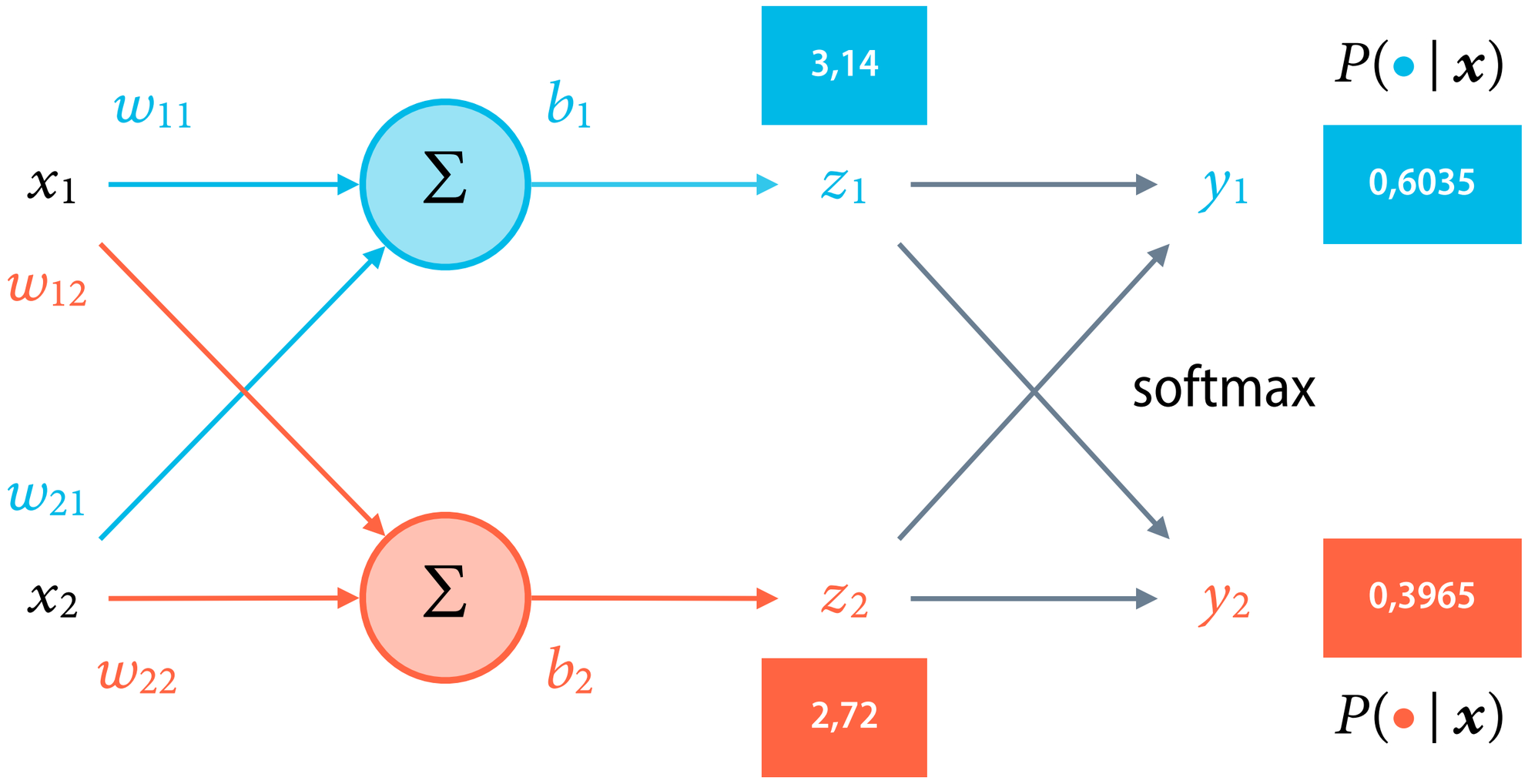
Den logistiska modellen

- Den **logistiska modellen** utökar den generaliserade linjära modellen med en icke-linjär funktion f:

$$\boldsymbol{y} = f(\boldsymbol{z}) \qquad \text{där} \qquad \boldsymbol{z} = \boldsymbol{x}\boldsymbol{W} + \boldsymbol{b}$$

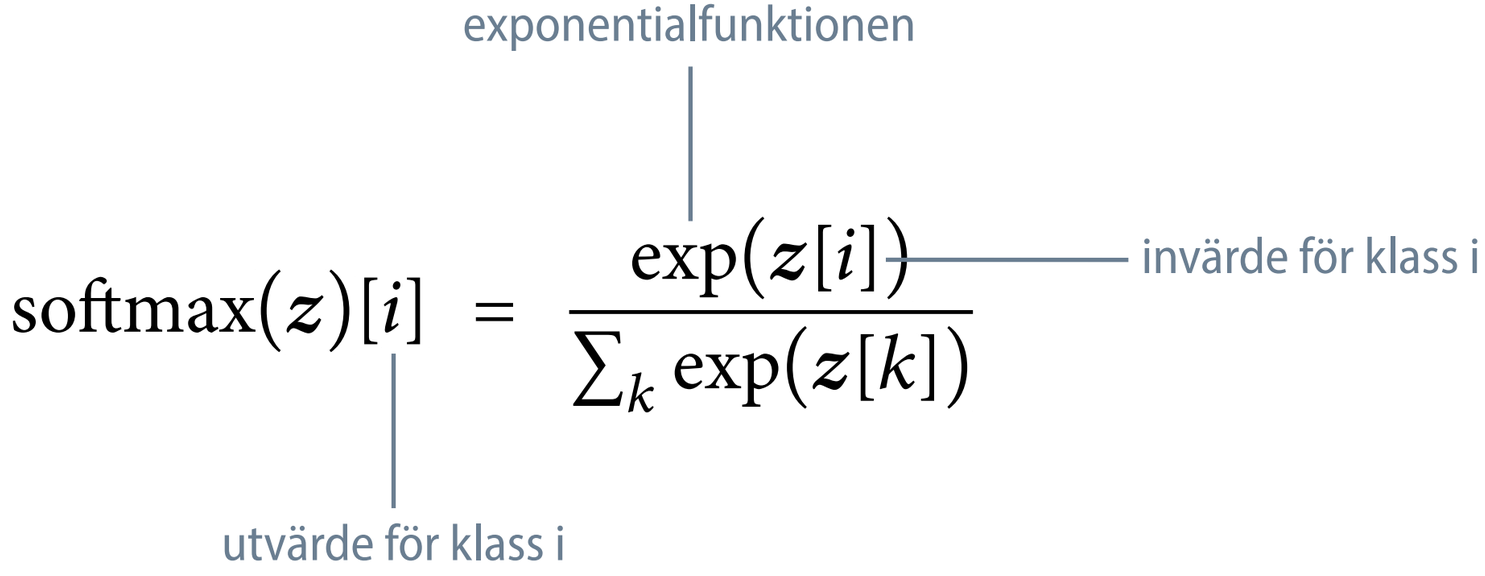
- Denna så kallade **softmax-funktion** transformerar de klasspecifika utvärdena till tal som kan tolkas som sannolikheter.

Logistisk regression

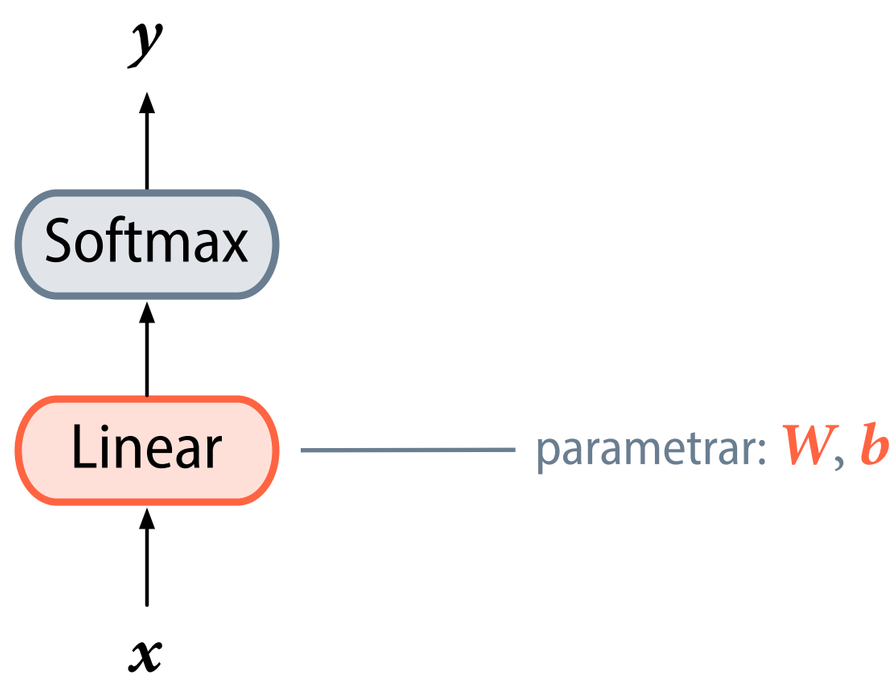


Softmax-funktionen

Softmax-funktionen transformerar alla invärden till icke-negativa tal och normaliserar värdena så att de summerar till 1:



Logistisk regression som neuronnät



Samma parametrar som ett linjärt neuronnät.

AI för naturligt språk

Inlärningsalgoritmen för logistisk regression

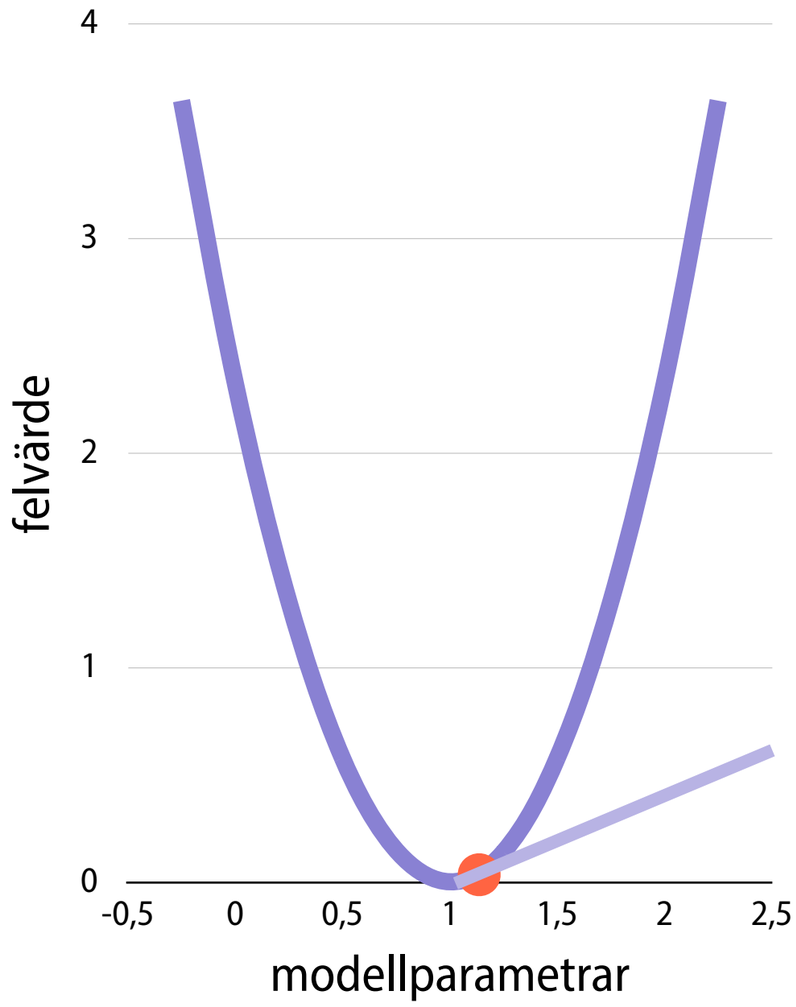
Marco Kuhlmann
Institutionen för datavetenskap

LINKÖPINGS UNIVERSITET

Att träna upp logistisk regression

- Att träna upp en logistisk modell innebär att sätta vikterna och baskonstanterna i den underliggande linjära modellen.
- Eftersom logistisk regression är en probabilistisk modell kan vi använda MLE-principen, precis som för Naive Bayes.
- Den här gången räcker det inte med att räkna – men vi kan använda gradientsökning!

Gradientsökning



modell	felvärde	gradient
2,000	2,33	4,67
1,533	0,66	2,49
1,284	0,19	1,33
1,151	0,05	0,71

steglängdsfaktor = 0,1

subtrahera 0,071

Korsentropi

- Istället för att maximera sannolikheten av träningsmängden minimerar vi modellens **korsentropi** (eng. *cross-entropy*).

  motsvarigheten till kvadratfel i linjär regression

- Korsentropifelet för ett träningsexempel $(\boldsymbol{x}, y)$ är den negativa logaritmen av den sannolikhet som modellen tilldelar y.

  formel: $-\log P(y|\boldsymbol{x})$

Korsentropi

P(A|x)	P(B|x)	guldstandard	felvärde
1,00	0,00	**A**	0,00
0,50	0,50	**A**	0,30
0,25	0,75	**A**	0,60
0,00	1,00	**A**	∞

Korsentropi