Maskininlärning

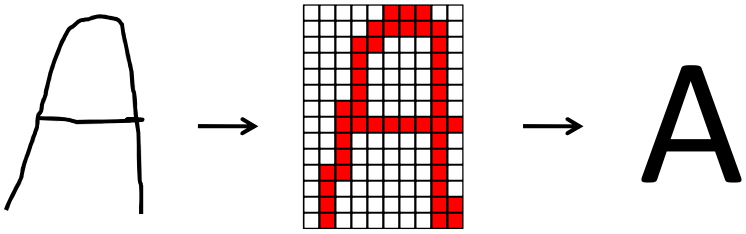
”Field of study that gives computers the ability to learn without being explicitly programmed” (Samuel, 1959)

DD2418 Språkteknologi, Johan Boye

Regler eller ML?

- System som bygger på
  - **handskrivna regler** (labb 1 & 2 & 4)
  - **maskininlärning** (labb 3 & 5)
- Handskrivna regler
  - bygger på språklig intuition, expertkunskaper
- Maskininlärning
  - är **data-driven**
  - **generaliserar** från exempel
  - inkluderar delar från statistik, artificiell intelligens, mm.

Exempel: OCR



Hur hittar vi skattefuskare?

	Kategori	Kategori	Kontinuerlig	Klass
#	Återbäring	Civilstånd	Inkomst	Fuskat?
1	Ja	Singel	600K	Nej
2	Nej	Gift	400K	Nej
3	Nej	Singel	300K	Nej
4	Ja	Gift	420K	Nej
5	Nej	Skild	380K	Ja
6	Nej	Gift	220K	Nej
7	Ja	Skild	800K	Nej
8	Nej	Singel	360K	Ja
9	Nej	Gift	240K	Nej
10	Nej	Singel	340K	Ja

Särdrag (features)

Sven: Ingen återbäring, skild, tjänar 120K?

POS-taggning

Ett annat förslag, som i sig inte spar pengar
DT PN NN HP PP PN AB VB NN

utan bara omfördelar resurser, är att
KN AB VB NN VB IE

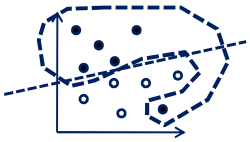
organisera om befälen
VB PL NN

Två typer av inlärning

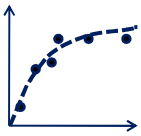
- Supervised (övervakad)
  - vi vet **rätt svar** för ett antal **exempel**
  - vi vill **generalisera** även till osedda exempel
- Unsupervised (oövervakad)
  - en massa exempel, **utan svar**
  - vi vet något mer, t.ex. möjliga värden

Supervised learning

- Klassificering
  - vi försöker förutsäga en **diskret** variabel



- Regression
  - vi försöker förutsäga en **kontinuerlig** variabel



Unsupervised learning

- Data mining
  - leta efter intressanta mönster
  - kluster, vanliga fenomen, ovanliga fenomen, …
- Reinforcement learning
  - en agent som interagerar med sin omgivning **lär sig av sina erfarenheter**

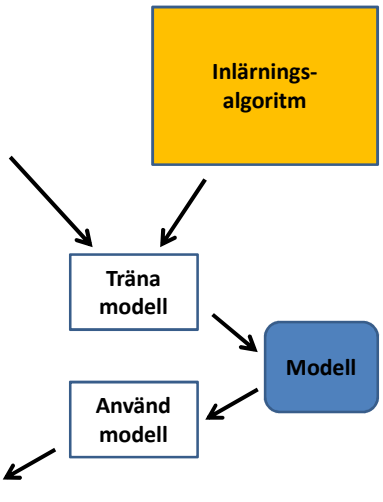
Klassificering

- Givet en **träningsmängd** med ett antal objekt:
  - Varje objekt har ett antal **attribut** *x* + ett ytterligare attribut som är objektets **klass** (*y*)
- Hitta en modell som förutsäger klassen *y* som en funktion av de övriga attributen
- Mål: tidigare osedda objekt skall tilldelas en klass så korrekt som möjligt

Klassificering

#	Återbäring	Civilstånd	Inkomst	Fuskat?
1	Ja	Singel	600K	Nej
2	Nej	Gift	400K	Nej
3	Nej	Singel	300K	Nej
4	Ja	Gift	420K	Nej
5	Nej	Skild	380K	Ja
6	Nej	Gift	220K	Nej
7	Ja	Skild	800K	Nej
8	Nej	Singel	260K	Ja
9	Nej	Gift	240K	Nej
10	Nej	Singel	360K	Ja

11	Nej	Gift	250K	?



Metodik för klassificering

1. Samla en datamängd M med facit
2. Bygg en representation av exemplen/objekten
3. Välj en inlärningsalgoritm och träna den på M
4. Utvärdera på testmängd T
5. Applicera på nya exempel

OBS! M och T är **olika**.

Representation

	Kategori	Kategori	Kontinuerlig	Klass
#	Återbäring	Civilstånd	Inkomst	Fuskat?
1	Ja	Singel	600K	Nej
2	Nej	Gift	400K	Nej
3	Nej	Singel	300K	Nej
4	Ja	Gift	420K	Nej
5	Nej	Skild	380K	Ja
6	Nej	Gift	220K	Nej
7	Ja	Skild	800K	Nej
8	Nej	Singel	360K	Ja
9	Nej	Gift	240K	Nej
10	Nej	Singel	340K	Ja

Särdrag (features)

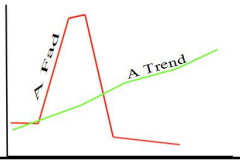
Det man vet om ett exempel kallas **särdrag.**

Särdrag kan vara **kontinuerliga** eller **diskreta** (kategorier).

Oväsentliga särdrag kallas **brus**. Att inse vad som är brus kan vara mycket svårt.

Exempel - betydelsebestämning

- Vad betyder ”fluga”?
  - insekt (”insect”)
  - klädesplagg (”bowtie”)
  - fiskedrag (”fishing”)
  - kortlivad trend (”fad”)



Exempel - betydelsebestämning

- "de dog som flugor", insect
- "jag menar självklart inte att Internet är en övergående fluga", fad
- "vi har exklusiva flugor för direkt leverans", bowtie
- "att knyta fluga är enklare än man tror", bowtie
- "hovmästaren bar en vit fluga", bowtie
- "hovmästaren det är en fluga i min soppa", insect
- "sudoku blev snabbt en fluga i Japan", fad
- "fluga sprutar larver i ögonen", insect
- "flugan är en tidlös herraccessoar", bowtie
- "Lydia Davis fjärde novellsamling heter samarbete med fluga", insect
- "konsten att knyta en fluga", bowtie
- "designade flugor i hundra procent ull", bowtie
- "är facebook en fluga eller en frälsare", fad
- "flugan är på väg att göra comeback", bowtie

Exempel - betydelsebestämning

- "blå flugor av högsta kvalitet", fishing
- "vita flugor av högsta kvalitet", bowtie
- "svarta flugor av högsta kvalitet", bowtie
- "oranga flugor av högsta kvalitet", fishing
- "abborren hugger på fluga", fishing
- "man smaskar till sig själv men flugan flyger och är strax tillbaka", insect
- "pojken är mycket beskedlig och skulle inte göra en fluga förnär", insect
- "en fluga gör ingen sommar", insect
- "när den första flugan kom till är man oense om", fishing
- "app stores en fluga eller framtiden", fad
- "spansk fluga är ingen fluga utan en skalbagge", insect
- "sociala medier är en fluga", fad
- *osv* (*47 exempel*)

Bag/set-of-words-approach

- Låt varje ord vara ett särdrag
- Värdet på ett särdrag är antalet förekomster av motsvarande ord i meningen.
- T.ex. ”*flyg fula fluga flyg*” har flyg=2, fula=1, fluga=1, alla andra ord=0
- Detta kallas för en **bag-of-words**.
- Notera att ordföljd inte är representerat.
- Ofta tar man bort sk ”stoppord”, (dt,kn,sn,pn,pp, vissa verb som ”har”, ”är”, mm.)

Klassificering: Utvärderingsmått

#	Fuskat?	Blev klassificerad som:
1	Nej	Nej
2	Nej	Nej
3	Nej	Ja
4	Nej	Nej
5	Ja	Nej
6	Nej	Nej
7	Nej	Ja
8	Ja	Ja
9	Nej	Ja
10	Ja	Ja
11	Nej	Ja
12	Ja	Nej

← **TN – True Negative**

← **FP – False positive**

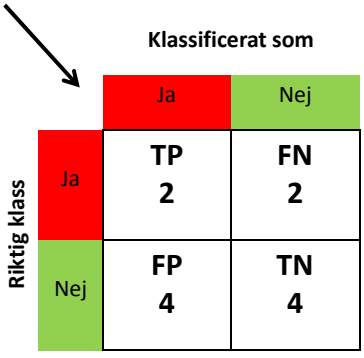
← **FN – False negative**

← **TP – True positive**

Klassificering: Utvärderingsmått

#	Fuskat?	Blev klassificerad som:
1	Nej	Nej
2	Nej	Nej
3	Nej	Ja
4	Nej	Nej
5	Ja	Nej
6	Nej	Nej
7	Nej	Ja
8	Ja	Ja
9	Nej	Ja
10	Ja	Ja
11	Nej	Ja
12	Ja	Nej

Confusion matrix
”Sammanblandningsmatris”



Klassificering: Utvärderingsmått

- Precision = $\frac{TP}{TP + FP}$
- Recall = $\frac{TP}{TP + FN}$
- Korrekthet (accuracy) = $\frac{TP + TN}{TP + TN + FP + FN}$

Fler än två klasser

		Klassificerat som			
		insect	fad	bowtie	fishing
Riktig klass	insect	14	0	0	0
	fad	7	4	0	0
	bowtie	10	0	2	0
	fishing	8	0	0	2

Träningsdata

- ”There is no data like more data”
- Tumregel: **Ju mer data, desto bättre resultat** (dock logaritmiskt)
- Om man har **för lite** data kan **slumpmässiga egenheter** råka läras in

Utvärdering

- Om man har data i överflöd: använd **separat tränings- och testmängd**
- Annars träna på t.ex. 90% av data, testa på 10%
- ***n*-fold cross-validation**
- Vid utveckling: Testa på träningsmängden
  - kan ge missvisande resultat!

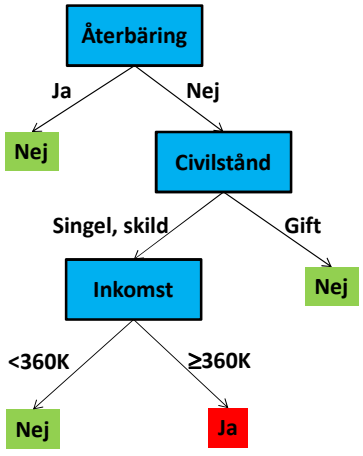
Algoritmer

- Beslutsträd
- Neurala nät
- Markovmodeller
- Regelinduktion (t.ex. TBL)
- Support Vector Machines
- Bayesianska metoder
- Genetiska algoritmer
- … och många fler

Beslutsträd

	Kategori	Kategori	Kontinuerlig	Klass
#	Återbäring	Civilstånd	Inkomst	Fuskat?
1	Ja	Singel	600K	Nej
2	Nej	Gift	400K	Nej
3	Nej	Singel	300K	Nej
4	Ja	Gift	420K	Nej
5	Nej	Skild	380K	Ja
6	Nej	Gift	220K	Nej
7	Ja	Skild	800K	Nej
8	Nej	Singel	360K	Ja
9	Nej	Gift	240K	Nej
10	Nej	Singel	340K	Ja

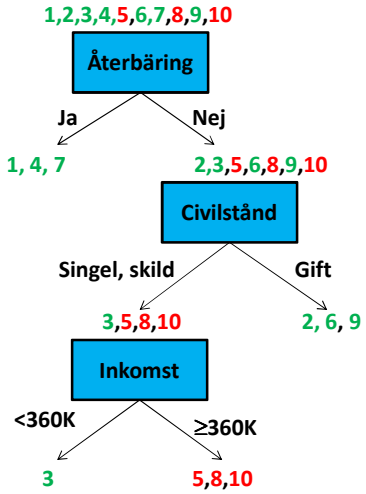
Träningsdata



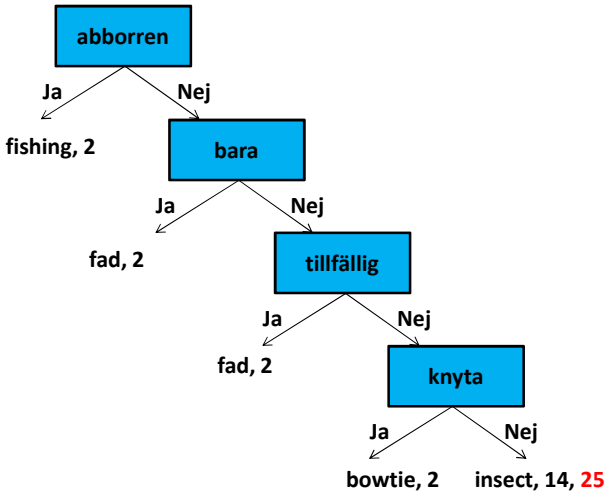
Modell: beslutsträd

Generera beslutsträd

#	Återbäring (Kategori)	Civilstånd (Kategori)	Inkomst (Kontinuerlig)	Fuskat? (Klass)
1	Ja	Singel	600K	Nej
2	Nej	Gift	400K	Nej
3	Nej	Singel	300K	Nej
4	Ja	Gift	420K	Nej
5	Nej	Skild	380K	Ja
6	Nej	Gift	220K	Nej
7	Ja	Skild	800K	Nej
8	Nej	Singel	360K	Ja
9	Nej	Gift	240K	Nej
10	Nej	Singel	340K	Ja



Flug-exemplet



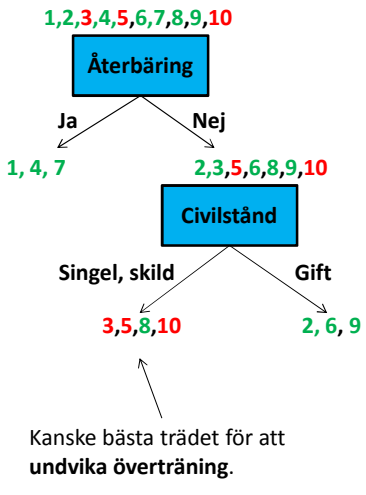
Överträning

- Överträning = **när hypotesen är alltför specialiserad**
- Ger bra resultat på träningsdata, men generaliserar dåligt.
- När händer detta?
  - Träningsdata är inte representativa
  - Fel i träningsdata
- Vad kan man göra?
  - Välj en enklare hypotes, och acceptera några fel för träningsexemplen

Överträning

#	Återbäring	Civilstånd	Inkomst	Fuskat?
	Kategori	Kategori	Kontinuerlig	Klass
1	Ja	Singel	600K	Nej
2	Nej	Gift	400K	Nej
3	Nej	Singel	300K	Ja
4	Ja	Gift	420K	Nej
5	Nej	Skild	380K	Ja
6	Nej	Gift	220K	Nej
7	Ja	Skild	800K	Nej
8	Nej	Singel	360K	Nej
9	Nej	Gift	240K	Nej
10	Nej	Singel	340K	Ja

Byte



Kanske bästa trädet för att **undvika överträning**.

Regelinduktion

- **Transformation-based learning (TBL)**
- Varje exempel ges en initial klassificering
- TBL försöker hitta **transformationsregler** som **förändrar klassificeringen av vissa exempel**, och gör den bättre
- Har applicerats på en mängd språkteknologiska problem, t.ex. POS-taggning

TBL och POS-taggning

Lexikon:

Psykologin	**NN**
hade	**VB**
tidigare	**AB**
mannens	**NN**
utveckling	**NN**
som	**HP**
norm	**NN**

Rules:
tag:ie>sn ← tag:pm@[1]
tag:hp>kn ← tag:nn@[1]

Input:

Psykologin hade tidigare mannens utveckling som norm…

NN VB AB NN NN KN NN ← Resultat

TBL och POS-taggning

- Starta med en taggning direkt från lexikon
- Applicera varje transformationsregel på hela korpuset, en regel i taget
- Senare regler kan ”ogöra” effekten av tidigare regler
- Motto: **”Gissa först, ändra sedan om det blir nödvändigt!”**

Inlärda regler

tag:ie>sn <- tag:pn@[1]	”… jag tror att han…”
tag:ie>sn <- tag:nn@[1]	”… jag tror att bordet …”
tag:hp>kn <- wd:som@[0] & tag:vb@[-1]	”… jag bedömer som viktigt…”
tag:pn>dt <- wd:de@[0] & tag:pp@[-1]	”… på de flesta ställen…”
tag:dt>pn <- tag:vb@[1]	”… ingen tar mig på allvar…”
tag:ie>sn <- tag:pm@[1]	” … jag tror att Nisse… ”
tag:hp>kn <- tag:nn@[1]	”… med mannen som norm…”
tag:pn>dt <- tag:pc@[1]	”… det hotande mörkret…”
tag:dt>rg <- wd:av@[1]	”… bara en av dem… ”
tag:hp>kn <- tag:dt@[1]	”… mannen som en norm….”
tag:ie>sn <- tag:jj@[1]	”… jag tror att rangliga bord…”
tag:pn>dt <- tag:pp@[-1] & tag:jj@[1]	”… på det hela taget…”

Regelmallar

- Bara regler som är instanser av någon **mall** kan läras
- T.ex. reglerna
  - tag:ie>sn <- tag:pn@[1]
  - tag:ie>sn <- tag:nn@[1]

  är instanser av
  - tag:A>B <- tag:C@[1]
  - ”Ändra tagg A till tagg B om efterföljande ord har tagg C”

Andra tillämpningar

- TBL har applicerats på en mängd språkteknologiska problem, t.ex.
  - part-of-speech tagging (Brill 1992)
  - pp-attachment disambiguation (Brill & Resnik 1994)
  - text chunking (Ramshaw & Marcus 1995)
  - spelling correction (Mangu & Brill 1997)
  - dialogue act tagging (Samuel et al. 1998)
  - ellipsis resolution (Hardt 1998)
- Labb 4: Named-entity recognition

Klassificering med hjälp av sannolikheter

- Kan vi beräkna sannolikheten att ett objekt ska tillhöra en viss klass givet dess attribut?
  - till exempel sannolikheten att en person har **fuskat**, givet att han **inte fått återbäring**, är **gift**, och tjänar **250K**.
  - dvs **P(Fuskat=ja | Åter=nej, Civil=gift, Inkomst=250)**
  - Sedan jämför vi denna sannolikhet med **P(Fuskat=nej | Åter=nej, Civil=gift, Inkomst=250)**
  - Om den översta sannolikheten är störst, så tror vi **Fuskat=ja**
  - Annars tror vi **Fuskat=nej**.

Använd Bayes sats

- Använd Bayes sats:

$$P(Fusk \mid ejÅter, gift, 250) = \frac{P(ejÅter, gift, 250 \mid Fusk) \cdot P(Fusk)}{P(ejÅter, gift, 250)}$$

#	Återbäring	Civilstånd	Inkomst	Fuskat?
1	Ja	Singel	600K	Nej
2	Nej	Gift	400K	Nej
3	Nej	Singel	300K	Nej
4	Ja	Gift	420K	Nej
5	Nej	Skild	380K	Ja
6	Nej	Gift	220K	Nej
7	Ja	Skild	800K	Nej
8	Nej	Singel	360K	Ja
9	Nej	Gift	240K	Nej
10	Nej	Singel	340K	Ja

Ett problem

- Använd Bayes sats:

$$P(Fusk \mid ejÅter, gift, 250) = \frac{P(ejÅter, gift, 250 \mid Fusk)\, P(Fusk)}{P(ejÅter, gift, 250)}$$

#	Återbäring	Civilstånd	Inkomst	Fuskat?
1	Ja	Singel	600K	Nej
2	Nej	Gift	400K	Nej
3	Nej	Singel	300K	Nej
4	Ja	Gift	420K	Nej
5	Nej	Skild	380K	Ja
6	Nej	Gift	220K	Nej
7	Ja	Skild	800K	Nej
8	Nej	Singel	360K	Ja
9	Nej	Gift	240K	Nej
10	Nej	Singel	340K	Ja

Problem: Det finns inget sådant exempel i träningsmängden!

Ett ”naivt” antagande

- Lösning: Antag att attributen är **oberoende**.
- I stället för

$$\frac{P(ejÅter, gift, 250 \mid Fusk) \cdot P(Fusk)}{P(ejÅter, gift, 250)}$$

så beräknar vi

$$\frac{P(ejÅter \mid Fusk) \cdot P(gift \mid Fusk) \cdot P(250 \mid Fusk) \cdot P(Fusk)}{P(ejÅter) \cdot P(gift) \cdot P(250)}$$

- Detta antagande är oftast inte korrekt (därav namnet “Naïve Bayes), men visar sig fungera bra i praktiken

Naïve Bayes

- Givet attributvärden E1, E2, E3 vill vi klassificera ett exempel som positivt (+) eller negativt (–).
- Vi jämför sannolikheterna
  1. $P(+ \mid E_1, E_2, E_3) = P(E_1 \mid +) \cdot P(E_2 \mid +) \cdot P(E_3 \mid +) \cdot P(+)$
  2. $P(- \mid E_1, E_2, E_3) = P(E_1 \mid -) \cdot P(E_2 \mid -) \cdot P(E_3 \mid -) \cdot P(-)$
- Om 1 är större än 2, så antar vi ”+”, annars ”–”.
- OBS! Vi kan skippa nämnarna i Bayes sats eftersom de är likadana i 1 och 2.

Uppskattning av sannolikheter

- T.ex. P(*skild*|*fuskat*) kan uppskattas med

$$\frac{antal(skild \cap fuskat)}{antal(fuskat)}$$

  i träningsmaterialet.
- I detta fall: 1/3.
- Hur är det med P(*gift*|*fuskat*)?

#	Återbäring	Civilstånd	Inkomst	Fuskat?
1	Ja	Singel	600K	Nej
2	Nej	Gift	400K	Nej
3	Nej	Singel	300K	Nej
4	Ja	Gift	420K	Nej
5	Nej	Skild	380K	Ja
6	Nej	Gift	220K	Nej
7	Ja	Skild	800K	Nej
8	Nej	Singel	360K	Ja
9	Nej	Gift	240K	Nej
10	Nej	Singel	340K	Ja

Smoothing

- För att inte någon av de uppskattade sannolikheterna ska bli 0:
  - Lägg till 1 i täljaren till varje uppskattad sannolikhet P(*Attr=a*|*Klass=k*)
  - Om *Attr* har *n* möjliga värden, lägg till *n* till nämnaren.
  - T.ex. P(*gift*|*fusk*) = 1/6
    P(*skild*|*fusk*) = 2/6
    P(*singel*|*fusk*) = 3/6

#	Återbäring	Civilstånd	Inkomst	Fuskat?
1	Ja	Singel	600K	Nej
2	Nej	Gift	400K	Nej
3	Nej	Singel	300K	Nej
4	Ja	Gift	420K	Nej
5	Nej	Skild	380K	Ja
6	Nej	Gift	220K	Nej
7	Ja	Skild	800K	Nej
8	Nej	Singel	360K	Ja
9	Nej	Gift	240K	Nej
10	Nej	Singel	340K	Ja

En språkteknologisk applikation

- **Klassificering av dokument** är en viktig applikation
  - identifiering av spam, e-mail-styrning, Net Nanny, nyhetsbevakning, …
- Vi vill klassificera dokument i två klasser
  - FLU – dokument som handlar om influensa
  - ~FLU – resten

Nr	Ord	Klass
1	cough, fever, temperature, flu	FLU
2	cough, pneumonia, flu	~FLU
3	flu, cough, flu	FLU
4	cold, flu	FLU

Uppskattning av sannolikheter

Nr	Ord	Klass
1	cough, fever, temperature, flu	FLU
2	cough, pneumonia, flu	~FLU
3	flu, cough, flu	FLU
4	cold, flu	FLU

- Vi uppskattar P(FLU) till ¾ och P(~FLU) till ¼.
- Hur kan vi uppskatta P(flu|FLU), P(fever|FLU), osv…?
  - Ett sätt: Räkna antal förekomster av ordet i klassen, dividera med totalt antal ord i klassen.
  - P(flu|FLU) = 4/9.
  - P(flu|~FLU) = 1/3.

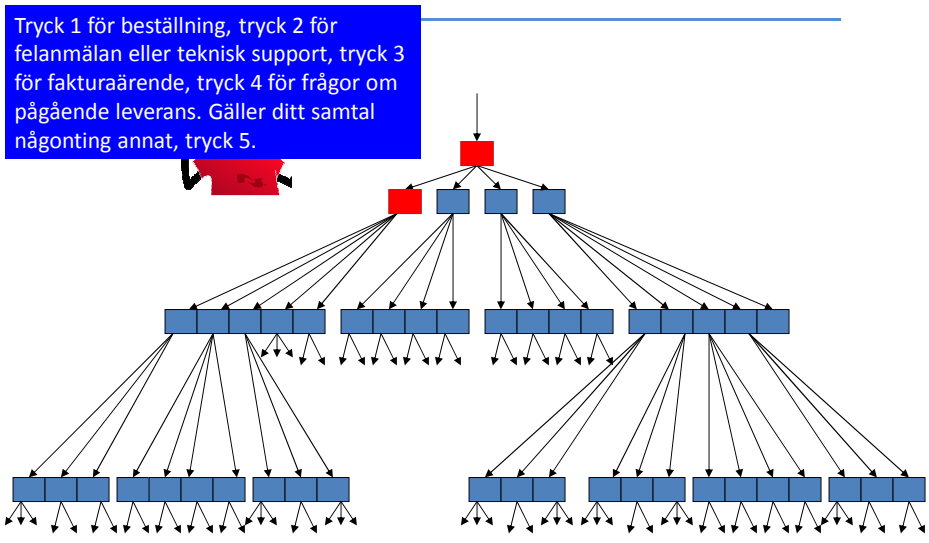
Klassificering

- Hur kan vi klassificera dokumentet ”*flu influenza*”?
- Beräkna sannolikheterna:
  - P(FLU)·P(flu|FLU) ·P(influenza|FLU) = ¾ · 4/9 · 0 = 0
  - P(~FLU) ·P(flu|~FLU) · P(influenza|~FLU) = ¼ · 1/3 · 0 = 0
- Med **smoothing**:
  - Det finns 7 olika ord → lägg till 7 i nämnaren
  - P(FLU)·P(flu|FLU) ·P(influenza|FLU) = ¾ · 5/16 · 1/16
  - P(~FLU) ·P(flu|~FLU) · P(influenza|~FLU) = ¼ · 2/10 · 1/10
  - Predicerad klass: FLU

Exempel: Samtalsstyrning

- Telefonibaserad kundtjänst för ett telekom-företag
  - 14 milj samtal/år
  - Fast telefoni, mobilt, bredband, mobilt bredband, uppringt Internet, IP-telefoni, digital-TV, triple play
  - Knapptryckningsmenyer
- Speciellt telefonnummer för mobilkunder
- Flera direktnummer till olika självbetjäningstjänster

Vad som fanns



Vad man ville ha i stället

Välkommen till XXX. Beskriv kortfattat vad du vill ha hjälp med så kopplar jag fram dig. Vad gäller ditt ärende?

Hej... jag skulle vilja hjälp att tolka några uppgifter på min mobilfaktura.

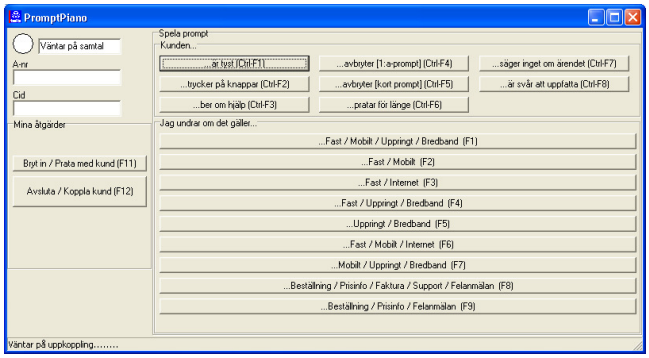
Ett fakturaärende mobiltelefoni, stämmer det?

Målet med projektet

- Ett enda ingångsnummer för alla kunder
- **Öka nöjdheten** hos betjänade kunder
- Öka utnyttjandet av **självbetjäningstjänster**
  - få kunderna att hitta dit
  - skapa nya typer av självbetjäningstjänster
  - skapa fler självbetjäningstjänster
- **Minska** antal **felaktiga samtalstyrningar**
- Få **färre** kunder att **lägga på luren**

Vald teknik

- **Taligenkänning** för att identifiera orden i kundens yttrande
- **Neuralt nätverk** för att identifiera kundens ärende
- Datainsamling med **Wizard-of-Oz**-metodik.



Implementation

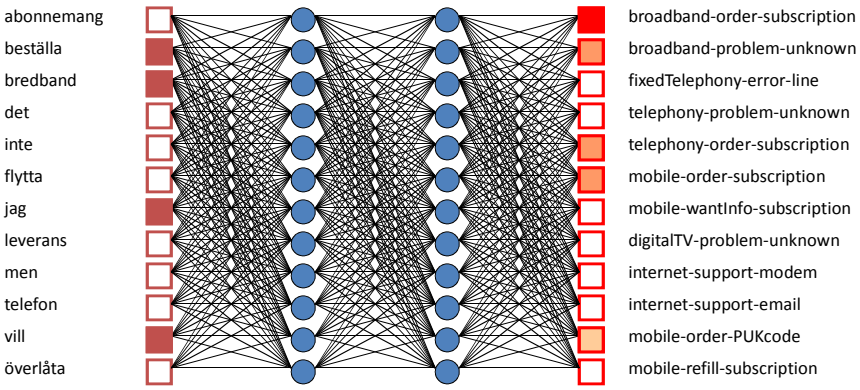
- Samtalsstyrning
  - Design av semantiska kategorier
  - Transkription och uppmärkning
  - Träning av taligenkännare och neuralt nätverk
- Dialogdesign
  - Logiskt dialogflöde
  - Persona och formuleringar
- Implementation
- 12 personer, ~ 12000 persontimmar

Uppmärkning av träningsdata

"Jag har en fråga på min Internet-räkning" → **`internet-wantInfo-billing`**

- Totalt fanns ca 140 sådana klasser.
- En klass motsvarande
  - en kö till handläggare, eller
  - en automatisk självbetjäningstjänst, eller
  - en motfråga från systemet

Klassificering



Hur bra blev det?

- I **riktig drift** klassificerades ca **82%** av yttrandena **helt rätt** (137 kategorier)
  - En version med 1500 kategorier hade ca 80% rätt
  - Klassificerare med **handskrivna regler** hade **67% rätt** (på 137 kategorier)
- Ca 7% av samtalen dirigerades fel
- Ökad kundnöjdhet
- Uppskattad ROI mindre än 1 år