# Repetition DMI, m.m.

I. Terminologi och Grundproblem
II. Linjär algebra
III. Optimering
IV. Sannolikhetslära
V. Parameterestimering



# Några begrepp

- Mönstervektor (egenskapsvektor/indata)
  - lista med numeriska värden som beskriver mönstret. Brukar heta **x**.

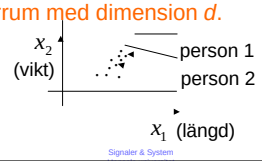
$$\mathbf{x} = \begin{pmatrix} \text{egenskap 1} \\ \text{egenskap 2} \\ \vdots \\ \text{egenskap d} \end{pmatrix}$$

varje egenskap ses som en koordinat i ett *d*-dimensionellt rum



- Börvärde. Ofta vektor (eng. target vector, el. desired output)
  - den vektor som hör ihop med **x**.
  - brukar heta **t**, **y** eller **d**.
- Mönsterrum
  - vi betraktar ett mönster, **x**, som en punkt i ett vektorrum med dimension *d*.

Ex:



## Grundproblem i DMI

- Klassificering (förk. ”klassning”)
  - synonym: mönsterigenkänning
  - nästan synonym: detektion
- Regression
  - nästan synonym: funktionsapproximation
- Kodning
  - effektiv representation av data



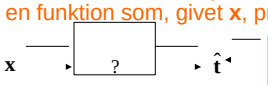
## För detta krävs ...

- Mängd med träningsexempel (träningsmängd)
  - synonymer: designexempel, träningsdata
  - antingen (a) $X = \{\mathbf{x}_1, \mathbf{t}_1, ..., \mathbf{x}_N, \mathbf{t}_N\}$
  - eller (b) $X = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$
- Inlärning
  - synonymer: träning, parameterestimering
  - kan var övervakad (supervised, fall (a)) eller oövervakad (unsupervised, fall (b))



## Övervakad inlärning

- Träningsdata $X = \{\mathbf{x}_1, \mathbf{t}_1, ..., \mathbf{x}_N, \mathbf{t}_N\}$
  - Målet är att, baserat på träningsdata, skapa en funktion som, givet **x**, predikterar **t**.




skattning/estimat/ approximation

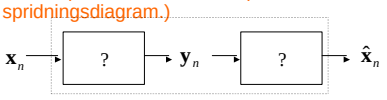
Vanlig vid *klassificering* att låta **t** vara en s.k. *indikatorfunktion*

$$\mathbf{t}_n = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ 0 \\ \vdots \end{pmatrix}$$

1 på pos. *i* om $\mathbf{x}_n \in C_i$
0 annars

## Oövervakad inlärning

- Träningsdata $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$
  - Mål: finna (oftast kompakt) representation av **x**, dvs *koda* eller *komprimera* i syfte att:
    - Förenkla lösningen av problem som kan beskrivas som övervakad inlärning.
    - Visualisera (Koden **y** har typiskt max 3 dimensioner. Kan då plotta mönstren som punkter i spridningsdiagram.)




$\mathbf{y}_n$ har lägre dimension än $\hat{\mathbf{x}}_n$

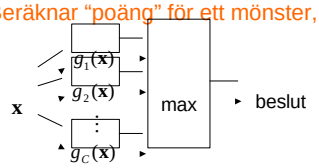


- Exempel på oövervakad inlärning
  - Via principalkomponentanalys (PCA). Koden är kontinuerlig.
  - Olika former av klustring. Koden är diskret (heltal)
    - **Hierarkisk klustring**
    - **K-means clustering (KMC)**
  - Kombinationer av PCA och klustring



## Flera begrepp

- Diskriminantfunktioner.
  - förekommer vid klassificering. En funktion per klass (undantag ev. för 2-klassproblem)
  - Beräknar "poäng" för ett mönster, **x.**

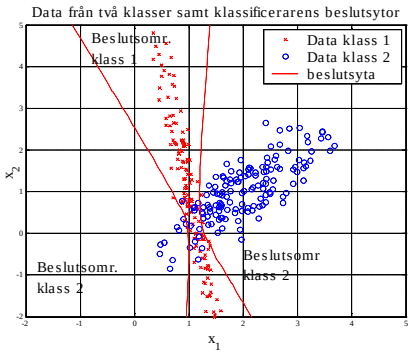


- Beslutsområden
  - ett område som associeras till en viss diskriminantfunktion (klass)
  - beslutsområdet för klass $C_i$ är det område i mönsterrummet där $g_i(\mathbf{x}) > g_j(\mathbf{x})$ för alla $j \neq i$
- Beslutsgräns/beslutsyta
  - gränsen mellan två beslutsområden
  - punkter på beslutsgränsen definieras av ekv:

$$g_i(\mathbf{x}) = g_j(\mathbf{x})$$



Ex:






# II. Linjär algebra

- Hur man "läser" matris/vektor-uttryck lättare

Definitioner: $\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix}$, $\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix}$ och $\mathbf{W} = \begin{pmatrix} w_{11} & \cdots & w_{1d} \\ \vdots & & \vdots \\ w_{m1} & \cdots & w_{md} \end{pmatrix}$

eller $\mathbf{W} = (\mathbf{w}_1 \quad \cdots \quad \mathbf{w}_d)$ där $\mathbf{w}_k = \begin{pmatrix} w_{1k} \\ \vdots \\ w_{mk} \end{pmatrix}$

Dessutom $\mathbf{U} = \begin{pmatrix} \mathbf{u}_1^T \\ \vdots \\ \mathbf{u}_m^T \end{pmatrix}$ radvektorer av dimension *d*

- Hur man "läser" ...

(a) Matris/vektor-produkt sett som en viktad summa av kolumnvektorer

$$\mathbf{y} = \mathbf{W}\mathbf{x} = \sum_{k=1}^{d} \mathbf{w}_k x_k$$

(b) Matris/vektor-produkt sett som en serie skalärprodukter

$$\mathbf{y} = \mathbf{U}\mathbf{x} = \begin{pmatrix} \mathbf{u}_1^T \mathbf{x} \\ \vdots \\ \mathbf{u}_m^T \mathbf{x} \end{pmatrix}$$



- Hur man "läser" ...
  - (c) Matris/matris-produkt sett som en serie matris/vektor-produkter

Bilda $\mathbf{Y} = (\mathbf{y}_1 \quad \cdots \quad \mathbf{y}_N)$, $\mathbf{X} = (\mathbf{x}_1 \quad \cdots \quad \mathbf{x}_N)$

$\Rightarrow \mathbf{Y} = \mathbf{W}\mathbf{X} = (\mathbf{W}\mathbf{x}_1 \quad \cdots \quad \mathbf{W}\mathbf{x}_N)$

- Egenvärde (evä) och egenvektorer (eve)
  - Def.

$\mathbf{A}\mathbf{x} = \lambda \mathbf{x}$ — $\lambda$: evä, $\mathbf{x}$: eve

**A** är kvadratisk matris



- Användbart "verktyg": spåret (trace) av en matris
  - def: $\mathrm{tr}(\mathbf{A}) = \sum_k (\mathbf{A})_{k,k}$, dvs summan av diagonalelementen
  - egenskaper (se utdelat papper)
  - Erbjuder ofta ett kompakt skrivsätt för uttryck som innehåller summor
    - **ex: Kriteriefunktion vid linjär regression (OLS)**

$$J(\mathbf{W}) = \sum_n \|\mathbf{y}_n - \hat{\mathbf{y}}_n\|^2 = \sum_n (\mathbf{y}_n - \mathbf{W}\mathbf{x}_n)^T (\mathbf{y}_n - \mathbf{W}\mathbf{x}_n)$$

kan skrivas $J(\mathbf{W}) = tr\left((\mathbf{Y} - \mathbf{W}\mathbf{X})^T (\mathbf{Y} - \mathbf{W}\mathbf{X})\right)$

med $\mathbf{Y} = (\mathbf{y}_1 \quad \cdots \quad \mathbf{y}_N)$, $\mathbf{X} = (\mathbf{x}_1 \quad \cdots \quad \mathbf{x}_N)$

# III. Optimering

- Problem: minimera $f(\mathbf{x})$ m.a.p **x**, dvs vi söker vektorn $\mathbf{x}^* = \arg\min_{\mathbf{x}} f(\mathbf{x})$
- *Lokalt minimum* definieras av

$$\nabla_{\mathbf{x}} f(\mathbf{x}) = \mathbf{0}, \quad \text{där } \nabla_{\mathbf{x}} f(\mathbf{x}) = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \end{pmatrix} \quad \text{samt att } \mathbf{x}^T \mathbf{H} \mathbf{x} > 0$$

Hessianen



- Metoder för optimering
  - gradientmetoden (steepest descent)
  - Newtons metod: Ide' approximera funktionen (lokalt) med en Taylorutveckling upp till kvadratiska termer. Minimera sedan denna kvadratiska funktion.
    - **Fördel: Mkt snabb konvergens nära ett minimum**
    - **Problem:**
      - Kräver att Hessianen är pos. def.
      - om dimensionen för **x** är stor, stora beräkningar



- Optimering under bivillkor
  - söker $\mathbf{x}^* = \arg\min_{\mathbf{x}} f(\mathbf{x}),$ b.v. $c_i(\mathbf{x}) = 0,$ för $i = 1,\ldots, K$
  - Enligt Lagrange: Om är ett lok. min. till f(.) (och uppfyller bivillkoren) så existerar

$l_1,\ldots, l_K$ sådana att $\mathbf{x}^*$ och $l_1,\ldots, l_K$ uppfyller ekv.

$$\nabla_{\mathbf{x}} L(\mathbf{x}, l_1,\ldots, l_K) = \mathbf{0}$$
$$c_i(\mathbf{x}) = 0, \quad l_i \geq 0 \text{ för alla } i$$

$$L(\mathbf{x}, l_1,\ldots, l_K) = f(\mathbf{x}) - \sum_i l_i c_i(\mathbf{x})$$

# IV. Sannolikhetslära

- *sannolikhet* associeras med utsagor, tex
  - *A*="det regnar i Skövde"
  - I=vår bakgrundsinformation
  - P(A|I) utläses: sannolikhet att utsaga A är sann, givet att I är sann.
  - Egenskap: $0 \le P(A) \le 1$
- Räkneregler för sannolikheter
  - Produktregeln: P(A,B|I)=P(A|B,I)P(B|I)

"givet att"

"och"



- Räkneregler, forts
  - Summaregeln: P(A eller B|I)=P(A|I)+P(B|I)-P(A,B)
- Konsekvens av produktregeln: Bayes sats
  - Gäller att $P(A,B \mid I) = P(B,A \mid I)$

$$\Rightarrow P(A \mid B,I)P(B \mid I) = P(B \mid A,I)P(A \mid I)$$

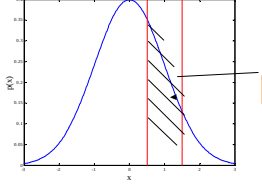
$$\Rightarrow P(B \mid A,I) = \frac{P(A \mid B,I)P(B \mid I)}{P(A \mid I)}$$

- Specialfall av summaregeln:
  - Om A och B är *uteslutande* (kan ej ske samtidigt)
  - P(A eller B|I)=P(A|I)+P(B|I), ty P(A,B|I)=0



- Täthetsfördelning (pdf) för skalär, x
  - Betecknas p(x). Egenskap $p(x) \ge 0$. En funktion sådan att $P(a \le x \le b) = \int_a^b p(x)dx$

Ex: pdf för Normalfördelade x



Arean=P(0.5 < x < 1.5)

- Täthetsfördelning (pdf) för vektor, $\mathbf{x} = (x_1,...,x_d)^T$
  - Betecknas p**(x).** Egenskap: $p(\mathbf{x}) \geq 0.$
  - "betydelse": $p(\mathbf{x}) = p(x_1, x_2,...,x_d).$
  - Produktregeln gäller, dvs

    $p(x_1, x_2,..., x_d) = p(x_1 \mid x_2,..., x_d)\, p(x_2,..., x_d).$ osv.
  - Blandning av diskreta och kontinuerliga variabler möjligt, tex

    $p(\mathbf{x}, \text{klasstillhörighet } C_i) = p(\mathbf{x}, C_i) = p(\mathbf{x} \mid C_i) P(C_i)$

    förkortat skrivsätt



- Tillämpning av shtslära inom MI
  - Anta att **x** "dras" från olika klasser $C_1,..., C_c$ med *a priori-sannolikheter* $P(C_1),...,P(C_c)$
  - Olika pdf:er, $p(\mathbf{x} \mid C_1),..., p(\mathbf{x} \mid C_c)$, för varje klass
  - Minimering av felsannolikhet leder till den optimala beslutsregeln:

    välj $C_i$ om $P(C_i \mid \mathbf{x}) \geq P(C_j \mid \mathbf{x})$ för alla $j \neq i$
  - *a-posteriorisannolikheterna* $P(C_i \mid \mathbf{x})$ fås via Bayes sats

    $$P(C_i \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid C_i) P(C_i)}{p(\mathbf{x})}$$

    måste vanl. skattas



## Några viktiga begrepp:

- Väntevärde

  $$\mathbf{m} = E[\mathbf{x}] = \int_{\mathbf{x}} \mathbf{x} p(\mathbf{x}) d\mathbf{x} = \int_{x_1}\int_{x_2}...\int_{x_d} \mathbf{x} p(\mathbf{x}) dx_d ... dx_2 dx_1$$
  - kan tolkas som en *tyngdpunkt* för en massfördelning, p**(x).**
- Kovariansmatris

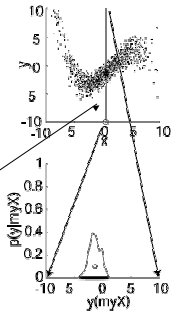
  $$\mathbf{C} = E[(\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})'] = \int_{\mathbf{x}} (\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})' p(\mathbf{x}) d\mathbf{x}$$
  - ger information om fördelningens spridning
  - analyseras m.h.a principalkomponentanalys

- Betingad pdf, $p(\mathbf{y}\,|\,\mathbf{x})$
  - betecknas p(y|x)
  - täthetsfördelningen för y då x är låst.
- Betingat väntevärde

$$m_{y|x} = E[y\,|\,x] = \int_y y\,p(y\,|\,x)\,dy$$

Tyngdpunkten längs denna linje

Ex: y och x är skalärer.
De har simultana pdf:en p(x,y)






- Vanligt ex på vektorvärd pdf, normalfördelningen:

kovariansmatris väntevärde

$$p(\mathbf{x}) = \frac{1}{(2p)^{d/2}\,|\,\mathbf{C}\,|^{1/2}}\,e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m})'\mathbf{C}^{-1}(\mathbf{x}-\mathbf{m})}$$

d=dimension determinant



# V: Parameterskattning

- Problem: Vi har data $X = \{\mathbf{x}_1,\ldots,\mathbf{x}_N\}$ som antas vara "dragna" från pdf p(**x**) som har känd parametrisk form. Parametrarnas värden är dock okända.
  - ex: Vi *antar* att **x** är normalfördelad med väntevärde **m** och kovarianmatris **C**.
- Två (besläktade) priniciper för parameter-skattning:
  - Maximum likelihood (ML)
  - Maximum *a posteriori* (MAP)

Låt $q$ beteckna en vektor som innehåller alla okända parametrar

- ML: $\hat{q}_{ML} = \arg\max p(X|q)$

  s.k. likelihoodfunktion

- MAP: $\hat{q}_{MAP} = \arg\max p(q|X) = \arg\max p(X|q)p(q)$

  s.k. "prior"

  – Obs: om $p(q)$ är relativt konstant för $q$ i ett stort område så sammanfaller ML- och MAP-skattningarna