Masinõpe

Konstantin Tretjakov

11. mai 2009
Milleks on meil arvutid?
Milleks on meil arvutid?
Milleks on meil protsessoriga arvutid?
Milleks on meil protsessoriga arvutid?
Automatiseerimine!

- Protsesside automaatne juhtimine
- Süsteemide kontrollimine
- (Enese)kontrollitavad süsteemid
- Robootika, küberneetika
- Intelligent systems
Kuidas juhtida?

- Olgu meil mingi süsteem (nt. robot).
- Mida on vaja selleks et seda kontrollida?
Kuidas juhtida?

- Olgu meil mingi süsteem (nt. robot).
- Mida on vaja selleks et seda kontrollida?
- **Algoritm:**
  
  \[ f : \text{Sisend} \rightarrow \text{Väljund} \]
Kuidas juhtida?

- Olgu meil mingi süsteem (nt. robot).
- Mida on vaja selleks et seda kontrollida?
- **Algoritm:**
  
  \[ f : \text{Sisend} \rightarrow \text{Väljund} \]

- Kuidas seda kirjeldada? Kust seda võtta?
Kuidas seda kirjeldada

- Formaalne loogika ehk programmeerimiskeel

```java
public static void main(String[] args) {
    if (args[0].equals(args[1])) {
        ... 
    } else {
        for (int i = 0; i < 10; i++) {
            ... 
        }
    }
}
```

Kust seda võtta?

- Leiutada
Kuidas seda kirjeldada

- **Formaalne loogika ehk programmeerimiskeel**
  ```java
  public static void main(String[] args) {
      if (args[0].equals(args[1])) {
          ...
      }
  }
  else {
      for (int i = 0; i < 10; i++) {
          ...
      }
  }
  ```

Kust seda võtta?

- **Leiutada = süsteemist aru saada ja üldistada**
Kuidas seda kirjeldada

- Formaalne loogika ehk programmmeerimispeel

```java
public static void main(String[] args) {
    if (args[0].equals(args[1])) {
      //
    } else {
      for (int i = 0; i < n; i++) {
        //
      }
    }
}
```

Kust veda võtta?

- Leiduda ansaamist aru saada ja üldistada
Mida tähendab aru saada?
Mida tähistab aru saada?
Mida tähendab aru saada?
Kuidas seda kirjeldada

- Muster andmetes ehk mudel

  Model: Logistic Regression
  Data: Input1 Input2 Input2 -> Output
       9.1  -1.2  8.1   0.7
       7.2  -0.4  2.8   0.9
       2.7   1.0  5.5  -0.3
       ....

Kust seda võtta?

- Koguda andmeid ning hinnata mudelit
Tere tulemast, Masinõpe

Kui süsteemist ise aru ei saa, lase arvutil seda teha

• Masinõpe = Andmeanalüüs
• Masinõpe = Andmetest *mustrite* otsimine.
• Masinõpe = Andmetele *mudelite* sobitamine.
• Masinõpe = See, mida “algoritmiga” kirja panna ei saa.
Masinõppe abil saab

1. Õpetada *intelligentset käitumist*
Masinõppe abil saab
Masinõppe abil saab
Masinõppe abil saab

2. Saada andmetest *uusi teadmisi*
Masinõppe abil saab
Masinõppe abil saab
Masinõppe abil saab
Masinõppe abil saab

3. Andmeid automaatselt organiseerida

- Kõkkuvõted & visualisatsioonid.
- Pakkimine & müra vähendamine.
Masinõppe abil saab

1. Õpetada intelligentset käitumist
2. Saada andmetest uusi teadmisi
3. Andmeid automaatselt organiseerida
4. ...
Unsupervised learning
Supervised learning
Erindite detekteerimine
Sagedased mustrid

AATAACGGCCCGATGAGGAAACGAACGGTCGCACT
AAGATGAGACATGTCCCGAAAGGTGCATAAGTTAT
GGACGAAAAACTTTCTTTCGCCCTTTTGATGTCAGCCC
AGCGCGGAGGATGAGGATCAGCCCCCCGCAATATTTCAG
ATATGCGAGCTTTTCGCCTCCTCGGAAAGGGCAATAAA
GCGACGGCCCCGATGAGGGTTACTAGATTGGA
TGGGTGTTTACAGATCTCGGCTTACCCTTTATCA
ACCCCTGCTACAGACTCGTTGAGAATGCTACGGGATC
Huvitavamad mustrid
Kuidas

- Konkreetne algoritm on ülesande spetsiifiline
- Lähenemine on aga alati sama:
  1. Määra huvipakkuva mustrite hulga
  2. Määra mustritele headuse mõõdu
  3. Leia parim muster andmetes
- Sellest siis vajadus statistika ja optimiseerimismeetodite järgi.
1     Supervised Learning

2     Näide: Lineaarne regressioon
Supervised Learning

Olgu $\mathcal{X}$ ja $\mathcal{Y}$ hulgad, ning olgu meil *treeningnäidete hulk*:

$$D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \mid x_i \in \mathcal{X}, y_i \in \mathcal{Y}\}$$

Ülesanne on leida funktsioon $f_D : \mathcal{X} \rightarrow \mathcal{Y}$ mis üldistaks andmetes oleva funktsionaalse seose.

Sellise funktsiooniga saaks teha ennustusi uutes andmepunktides.
\begin{itemize}
  \item $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \mathbb{R}$.
  \item $D = \{(0.50, 0.26), (0.43, 0.08), (0.26, 0.00), \ldots \}$
  \item $f_D(x) = x^2$.
\end{itemize}
• $\mathcal{X} = \mathbb{R}^2$, $\mathcal{Y} = \{\text{blue, red}\}$.
• $D = \{((1.3, 0.8), \text{red}), ((2.5, 2.3), \text{blue}), \ldots \}$
• $f_D(x_1, x_2) = \text{if } x_1 + x_2 > 3 \text{ then blue else red.}$
Näiteid?
Supervised masinõppe lähenemised

Kuidas leida funktsiooni, mis “kirjeldab” andmeid hästi?

Palju võimalusi:

• Empiirilise riski minimiseerimine (min \( f_{\mathcal{E}}(f) \))
• Regulariseerimine (min \( f_{\mathcal{E}}(f) + \lambda C(f) \))
• Suurima tõepära meetod (max \( \theta \mathcal{P}(D|\theta) \))
• Bayesian inference (max \( \theta \mathcal{P}(\theta|D) \))
• Struktureeritud riski minimiseerimine (≈ min \( f_{\mathcal{R}}(f) \))
• ... (puud, boosting, etc)
Kuidas leida funktsiooni, mis “kirjeldab” andmeid hästi?

_Palju võimalusi:_
- Empiirilise riski minimiseerimine ($\min_f \mathcal{E}_D(f)$)
- Regulariseerimine ($\min_f \mathcal{E}_D(f) + \lambda \mathcal{C}(f)$)
- Suurima tõepära meetod ($\max_\theta \mathcal{P}(D | \theta)$)
- Bayesian inference ($\max_\theta \mathcal{P}(\theta | D)$)
- Struktureeritud riski minimiseerimine ($\approx \min_f \mathcal{R}_\mu(f)$)
- … (puud, boosting, etc)
Lineaarne regressioon
Empriirilise riski minimiseerimisega

• Vaatleme kõikide lineaarsete funktsioonide hulga \( \mathbb{R} \rightarrow \mathbb{R} \):

\[
f_w(x) = wx
\]

• Otsime nendest sellise, mis ennustab väikseima veaga:

\[
\mathcal{E}_D(f) = \sum_{(x_i, y_i) \in D} (y_i - f(x_i))^2
\]
Funktsioonide hulk $f_w(x) = wx$. 
Lineaarne regressioon

\[ \mathcal{E}_D(f) = \sum_{(x_i, y_i) \in D} (y_i - f(x_i))^2 \]
Optimeerime $\mathcal{E}_D(f_w)$

\[
\mathcal{E}_D(f_w) = \sum_i (y_i - wx_i)^2 = \sum_i y_i^2 + w^2 \sum_i x_i^2 - 2 \sum_i wx_i y_i
\]

millest

\[
\frac{\partial \mathcal{E}_D(f_w)}{\partial w} = w \sum_i x_i^2 - 2 \sum_i x_i y_i
\]

võrdsustame $\frac{\partial \mathcal{E}_D(f_w)}{\partial w}$ nulliga ning saame

\[
w = \frac{2 \sum_i x_i y_i}{\sum_i x_i^2}
\]
Viimane slaid

- Masinõpe = **mustriotsing andmetes**.
- Masinõpet saab **uurida**, välja mõeldes uusi meetodeid.
- Masinõpet saab **kasutada**, oma andmete analüüsimest olemasolevate meetoditega.
- Masinõpet saab **kasutada**, oma süsteemide loomisel.
- Masinõpest tasub igastahes **aru saada** mõnel määral.
- **Maatriksid, tõenäosusteooria, statistika ja optimiseerimismismeetodid** on teie sõbrad.
Küsi musi