732G12 Data Mining

Föreläsning 4

Johan Alenlöv
IDA, Linköping University, Sweden

Dagens föreläsning

- K-närmaste grannar
- Bayesianska klassificerare
- Ensamblemetoder
  - Bagging
  - Boosting
  - Random forest

K-närmaste grannar

Idé basera predikation på de K datapunkter som är närmast.

Ger en icke-parametrisk metod för klassificering och regression.

Problem: Vad är närmast?

Avståndsmått

Vi behöver något som talar om för oss hur nära två datapunkter är.
Finns många alternativ som man kan välja, som ger olika resultat.

Euklidiskt avstånd

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$$

Manhattan avstånd

$$d(\mathbf{x}, \mathbf{y}) = \sum_{k=1}^{n} |x_k - y_k|$$

K-närmaste grannar

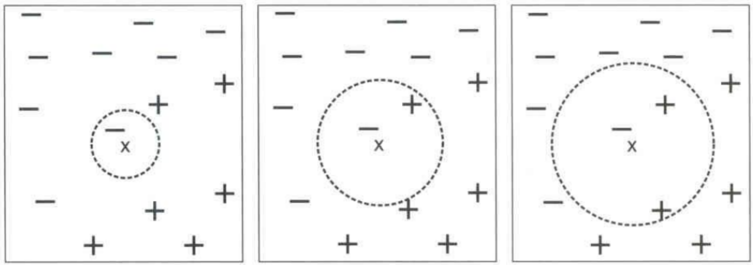
1. Låt k vara ditt valda antal grannar och D din träninigsdata.
2. För varje testdata $z = (\mathbf{x}', y') \in D$:
   2.1 Beräkna $d(\mathbf{x}, \mathbf{x}')$ (avstådet mellan z och all träningsdata)
   2.2 Välj $D_z \subseteq D$, de k närmaste träniningsdatan till z
   2.3 Låt $y' = \arg\max_v \sum_{(\mathbf{x}_i, y_i) \in D_z} \mathbf{I}_{v=y_i}$

2.3 är majoritetsvalet. Kan också vikta detta värde med avståndet:

2.3 $y' = \arg\max_v \sum_{(\mathbf{x}_i, y_i) \in D_z} w_i \mathbf{I}_{v=y_i}$.

För regression används medelvärde alternativt viktat medelvärde.

K-närmaste grannar



(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

K-närmaste grannar

- Målet med modellen är att prediktera nya observationer.
- Påverkas stort av olika skalor.
- Långsam anpassning.
- Känslig mot brus.
- Val av K har stor betydelse!
    - Litet K ger överanpassning.
    - Stort K ger underanpassning.
    - Korsvalidering kan användas för att bestämma K.
- Producerar godtyckligt utformade beslutsgränser.
- Problem i högre dimensioner.

Bayesiansk klassificerare

Att direkt modellera en icke-deterministisk funktion kan vara mycket svårt.

Exempel:

- $(\text{diet}, \text{träning}) \to (\text{hjärtinfarkt})$ är svårt
- $(\text{diet}, \text{träning}) \to \mathbb{P}(\text{hjärtinfarkt})$ lättare

Använd Bayes sats för att hjälpa till i modelleringen

$$\mathbb{P}(Y \mid \mathbf{X}) = \frac{\mathbb{P}(\mathbf{X} \mid Y)}{\mathbb{P}(\mathbf{X})} \cdot \mathbb{P}(Y) \propto \mathbb{P}(\mathbf{X} \mid Y) \cdot \mathbb{P}(Y)$$

$$\text{posterior} = \frac{\text{likelihood}}{\text{evidence}} \cdot \text{prior} \propto \text{likelihood} \cdot \text{prior}$$