

Tentamen LEREN

22 maart 2004
13.30 – 16.00 uur

Vraag 1

Geef definities van de volgende begrippen:

- pruning
- information gain
- inverse resolutie
- MAP hypothesis

Vraag 2

1. Welke aanname maakt de Naive Bayesian Classifier? Geef een preciese formulering van de aanname, bij voorkeur in de vorm van een formule.
2. Ook als deze aanname niet klopt, blijkt de Naive Bayesian Classifier vaak goed te werken. Waarom?
3. Een andere reden waarom men de Naive Bayesian Classifier vaak gebruikt (ipv. Bayes optimal) hoewel de aanname niet klopt, is dat de Bayes optimal classifier ook grote nadelen heeft. Wat is dat nadeel? Licht kort toe.

Vraag 3

1. We passen (a) 3-nearest-neighbour zonder weging van features en (b) decision-tree learning toe op een dataset. Bij kruisvalidatie blijkt de decision-tree learner het aanzienlijk beter te doen. Welke eigenschap(pen) van het domein kan/kunnen hiervan de oorzaak zijn?
2. Stel dat het omgekeerde blijkt: de decision tree leerder doet het veel slechter. Welke eigenschap(pen) van het domein kan/kunnen hiervan de oorzaak zijn?

Vraag 4

Stel dat we een model van een domein leren uit een steekproef en dit model gebruiken om een voorspelling te doen over de klasse van een bepaald voorbeeld (dat we niet in de trainingset gezien hebben). We herhalen het steekproeftrekken een aantal keren en we bepalen bij elke steekproef het totaal aantal foute voorspellingen van het geleerde model. De voorspellingen voor het ene voorbeeld blijken niet steeds hetzelfde te zijn. Een deel van de voorspellingen is dus fout. Het aantal fouten van de geleerde modellen varieert. We bepalen het gemiddeld aantal fouten en de variatie (bv. de variantie) in het aantal fouten over alle steekproeven.

1. Stel dat de gebruikte leermethode een sterke "learning bias" heeft. Vinden we die terug in onze experimenten? Kunnen we die meten?
2. Stel dat de gebruikte leermethode een verkeerde "learning bias" heeft. Vinden we dat terug in onze experimenten? Kunnen we dat meten?
3. Wat is het effect van de grootte van de steekproef op (a) het gemiddeld aantal voorspellingsfouten over alle steekproeven en (b) de variatie in voorspellingsfouten over experimenten?

Vraag 5

De version space methode wordt meestal gebruikt voor zgn. conjunctieve hypothesen. Als de instances worden beschreven als attribuut-waarde vectoren is een conjunctieve hypothese een conjunctie van attribuut-waarde combinaties.

1. Welke actie onderneemt de version space methode op de "boundary sets" S en G bij een nieuw negatief trainingsvoorbeeld?
2. Is het effect van de version space methode op een reeks training data afhankelijk van de volgorde van die data? Leg uit.
3. Stel dat een instance wordt beschreven door 10 booleaanse (tweewaardige) attributen en dat we alleen conjunctieve hypothesen toelaten. We initialiseren de version space met de meest algemene hypothese (voor G) en een speciale hypothese die alle voorbeelden uitsluit voor S. Wat is het maximale en het minimale aantal hypothesen in de boundary sets G en S na:
 - a. twee positieve trainingsvoorbeelden
 - b. twee negatieve trainingsvoorbeelden
 - c. een positief en een negatief trainingsvoorbeeld
4. Kan een Version Space leerder een **nieuw** (ongezien) voorbeeld ooit met zekerheid aan een klasse toekennen? En kan een beslisbomenleerder dat?