NTNU
Institutt for datateknikk
og informasjonsvitenskap

Eksamensoppgave i

TDT4173 – Maskinlæring og case-basert resonnering

Mandag 7. desember 2009, kl. 09:00 - 13:00

Oppgaven er utarbeidet av faglærer Agnar Aamodt og kvalitetssikrer Helge Langseth.
Kontaktperson under eksamen er Agnar Aamodt (mobil 92611144)

Språkform: Bokmål
Tillatte hjelpemidler: D
Ingen trykte eller håndskrevne hjelpemidler tillatt.
Bestemt, enkel kalkulator tillatt.

Sensurfrist: Torsdag 7. Januar 2010.

Les oppgaveteksten nøye. Finn ut hva det spørres etter i hver oppgave.

Dersom du mener at opplysninger mangler i en oppgaveformulering gjør kort rede for de antagelser og forutsetninger som du finner det nødvendig å gjøre.

Oppgave 1 – Generelt

a) Definer hva et velformulert læringsproblem (well-posed learning problem) er.

b) Hypotesen for induktiv læring (the inductive learning hypothesis) er en fundamental antakelse for all læring.
Hva uttrykker denne hypotesen?

c) Hva er "induktiv bias"?
Hva er de to hovedformer for induktiv bias?
Gi et eksempel på en metode uten induktiv bias.
Hva kan denne metoden lære?

d) Hva menes med overtilpasning (overfitting) i sammenheng med induktiv læring?
Hva kan forårsake overtilpasning, og hvordan kan man unngå overtilpasning? Du kan eksemplifisere ved å referere til læring av beslutningstrær om du vil.

Oppgave 2 - Bayesiansk læring

a) Noen maskinlæringsalgoritmer utnytter a priori kunnskap, andre gjør ikke det. Gi to eksempler på læremetoder (fra pensum) som benytter a priori kunnskap, og to som ikke gjør det.
For de to metodene som gjør det, beskriv kort hvilken rolle a priori kunnskap spiller i læremetoden hos hver av dem.

b) Hva er en Naive Bayes klassifikator?
Nevn styrker og svakheter denne klassifikatoren har. Forslag til egenskaper du kan vurdere: Kjøretids-kompleksitet av læring og klassifikasjon, plass-kompleksitet for modellen, og antagelsen en Naive Bayes klassifikator gjør om (betingede) uavhengigheter mellom de forskjellige variable.
Sammenling dette med egenskapene til generelle Bayesianske nett.
Begrunn svarene dine.

c) EM algoritmen består av to trinn ("E-trinnet" og "M-trinnet") som gjentas.
Beskriv med egne ord hva som skjer i de to trinnene.
Når er det nødvendig å bruke EM algoritmen?

Oppgave 3 - CBR

a) Nevn den prinsipielle forskjellen mellom k-Nearest Neighbour metoden og case-basert resonnering i mer generell forstand, for hvert av de følgende to trinn i CBR-syklusen:
- Retrieve
- Reuse

b) Gitt CBR-systemet Protos.
Hvordan indekseres casene i Protos?
I hvilke av CBR-syklusens 4 faser benyttes generell domenekunnskap?
Hvordan benyttes den generelle domenekunnskapen i Retrieve fasen?

c) I artikkelen ”Remembering to Forget” foreslås det en kompetanse-modell for casebaser.
Hva uttrykker begrepene

- ”retrieval space”
- ”adaptation space”
- ”coverage set”
- ”reachability set”?

Skisser kort hvordan dette rammeverket brukes til å definere en strategi for å slette case fra en casebase.

Oppgave 4 – Blandete oppgaver

a) Forklar begrepet "kryss-validering".
Hvordan brukes kryss-validering for å beregne godheten av en læringsalgoritme?

b) Et eksempel på en ensemble-algoritme fra pensum er "bagging".
Hva er bagging?
Hva er en "svak klassifikator" (weak classifier)?
Hvordan kan bagging-algoritmen forbedre resultatene en slik svak klassifikator gir?
Hvilke egenskaper bør den svake klassifikatoren ha for at "bagging" skal være virkningsfullt?

c) Support vektor maskiner kalles ofte "stor margin klassifikatorer" (large margin classifiers).
Forklar hvorfor dette er et passende navn for denne klassifikasjonsalgoritmen.

d) Support vektor maskiner omtales også ofte som "kjerne-maskiner" (kernel machines). Forklar hvilken rolle kjerner/kernels har i SVM rammeverket.