UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	STK4030/9030 — Moderne data analyse.
Eksamensdag:	Fredag 5. desember 2008.
Tid for eksamen:	14.30 – 17.30.
Oppgavesettet er på 3 sider.	
Vedlegg:	Ingen.
Tillatte hjelpemidler:	Godkjent kalkulator.

Kontroller at oppgavesettet er komplett
før du begynner å besvare spørsmålene.

Oppgave 1.

Anta vi har en regresjonssituasjon med input variable $\boldsymbol{x} \in \mathcal{R}^p$ og y en numerisk output. Basert på $\{(\boldsymbol{x}_i, y_i), i = 1, ..., n\}$ vil vi finne en regresjonstilpasning $\hat{f}(\boldsymbol{x})$.

(a) Anta $Y = f(\boldsymbol{x}) + \varepsilon$ der $E(\varepsilon) = 0$, $\text{Var}(\varepsilon) = \sigma_\varepsilon^2$ og støyleddene fra ulike observasjoner er uavhengige.

La $\hat{f}(\boldsymbol{x})$ være en regresjonstilpasning basert på data. Definer

$$\text{Err}(\boldsymbol{x}_0) = E[(Y - \hat{f}(\boldsymbol{x}_0))^2 | \boldsymbol{x} = \boldsymbol{x}_0]$$

og vis at

$$\text{Err}(\boldsymbol{x}_0) = \sigma_\varepsilon^2 + \text{Bias}^2(\hat{f}(\boldsymbol{x}_0)) + \text{Var}(\hat{f}(\boldsymbol{x}_0)).$$

Diskuter konsekvensene av dette resultatet.

Ridge regresjon (blandt mange andre metoder) bruker en lineær regresjons model

$$\hat{f}(\boldsymbol{x}) = \hat{\beta}_0 + \sum_{j=1}^{p} x_j \hat{\beta}_j$$

(Fortsettes side 2.)

der $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p)^T$ er definert gjennom

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

(b) Anta nå at $\sum_{i=1}^{N} x_{ij} = 0$ for alle j og $\sum_{i=1}^{N} x_{ij}x_{ij'} = 0$ for alle $j \neq j'$.

Utled analyttiske uttrykk for $\hat{\beta}_j, j = 0, 1, ..., p$.

Basert på disse analyttiske uttrykkene, diskuter hvordan leddet $\lambda \sum_{j=1}^{p} \beta_j^2$ influerer på disse uttrykkene og relater det til (a).

Oppgave 2.

Betrakt en generell situasjon der du for $i = 1, ..., n$ har observert input $\boldsymbol{x}_i \in \mathcal{R}^p$ og output y_i der y_i enten er numerisk eller kategorisk. Du ønsker å bruk data for å tilpasse en model som predikerer fremtidige Y'er.

(a) Det er vanlig å dele datasettet i et *trenings-sett* og et *test-sett* og noen ganger også et *validerings-sett*. Diskutér rollen disse settene har og fordeler/ulemper ved å gjøre en slik oppdeling av datasettet.

(b) Forklar hva vi mener med kryss-validering. Diskutér dens bruk og hvordan denne metoden relaterer seg til trening-/validering-/test-sett.

Oppgave 3.

Betrakt en klassifikasjonssituasjon med input variable $\boldsymbol{x} \in \mathcal{R}^p$ og output variabel $Y \in \{1, ..., K\}$. Anta en modell

$$\Pr(Y = k | \boldsymbol{x} = \boldsymbol{x}) = p_{m(\boldsymbol{x}),k}$$

der vi antar $\mathcal{R}^p$ er delt opp i M disjunkte regioner $R_m, m = 1, ..., M$ og $m(\boldsymbol{x})$ er regionen som $\boldsymbol{x}$ tilhører.

(a) Anta uavhengige observasjoner $\{(\boldsymbol{x}_i, y_i), i = 1, ..., n\}$ er tilgjengelig. Skriv ned likelihood funksjonen for den gitte modell og vis at maksimering av likelihooden korresponderer med miminering av

$$Q = -\sum_{m=1}^{M} \sum_{k=1}^{K} N_{m,k} \log p_{m,k}$$

der $N_{m,k}$ er antall observsjoner innen region R_m som tilhører klasse k.

(b) Definer $\hat{p}_{m,k} = N_{m,k}/N_m$ der $N_m = \sum_k N_{m,k}$ er antall observasjoner innen region R_m. Hvorfor er $\hat{p}_{m,k}$ et rimelig estimat for $p_{m.k}$?

Hvis vi innsetter $\hat{p}_{m,k}$ for $p_{m,k}$ og $N_m\hat{p}_{m,k}$ for $N_{m,k}$ i Q, hva slags mål innen terminologien for klassifikasjonstrær svarer da Q til?

(Fortsettes side 3.)

(c) For å bruke en slik model, må regionene $\{R_m\}$ også spesifiseres. Diskutér fordelene ved å bruke en trestruktur der en oppdeling (split) bare avhenger av én input variabel for å definere disse regionene.

SLUTT